USE OF WHOLE GENOME SEQUENCE DATA TO CHARACTERIZE MATING AND RNA SILENCING GENES IN *TILLETIA* SPECIES

By

SEAN WESLEY MCCOTTER

A thesis submitted in partial fulfillment of
the requirements for the degree of

MASTER OF SCIENCE IN PLANT PATHOLOGY

WASHINGTON STATE UNIVERSITY
Department of Plant Pathology

DECEMBER 2014

To the Faculty of Washington State University:

The members of the Committee appointed to examine the thesis of SEAN WESLEY MCCOTTER find it satisfactory and recommend that it be accepted.

Lori M. Carris, Ph.D., Chair

Dorrie Main, Ph.D.

Patricia Okubara, Ph.D.

Lisa A. Castlebury, Ph. D.

ACKNOWLEDGMENTS

The research presented in this thesis could not have been carried out without the expertise and cooperation of others in the scientific community. Significant contributions were made by colleagues here at Washington State University, at the United States Department of Agriculture and at Agriculture and Agri-Food Canada.

I would like to start by thanking my committee members Dr. Lori Carris, Dr. Lisa Castlebury, Dr. Pat Okubara and Dr. Dorrie Main, who provided guidance on procedure, feedback on my research as well as contacts and laboratory resources.

Dr. André Lévesque of AAFC initially alerted me to the prospect of collaboration with other AAFC *Tilletia* researchers and placed me in contact with Dr. Sarah Hambleton, whose lab sequenced four out of five strains of *Tilletia* used in this study (CSSP CRTI 09-462RD). Dr. Prasad Kesanakurti and Jeff Cullis coordinated my access to AAFC's genome and transcriptome data for these species. Separately, Dr. Guus Bakkeren, also of AAFC, provided technical advice on cloning and sequencing smut fungal mating loci.

Dr. Jodi Humann was an invaluable source of technical advice and the key individual responsible for the PacBio+454 hybrid genome assembly. Crucially, Dr. Ping Zheng developed Perl scripts which saved me weeks of work. Mark Wildung and Derek Pouchnik at the WSU Laboratory for Biotechnology and Bioanalysis were responsible for next generation sequencing on both PacBio RS and Roche 454 platforms. Dr. Tobin Peever at WSU graciously allowed me bench space in his laboratory where my lab-mate, Dr. Lydia Tymon, provided a great partner-in-a-pinch for impromptu dance parties and rarely criticized my controversial tastes in music.

USE OF WHOLE GENOME SEQUENCE DATA TO CHARACTERIZE MATING AND RNA-SILENCING GENES IN *TILLETIA* SPECIES

Abstract
by Sean Wesley McCotter, M.S.
Washington State University
December 2014

Chair: Lori M. Carris

Tilletia species (Ustilaginomycotina, Basidiomycota), the bunt fungi, are pathogens of grasses (Poaceae) such as wheat (*Triticum aestivum*) and ryegrass (*Lolium* spp.) and represent a molecularly underexplored branch of the fungal tree of life. Most *Tilletia* species must mate prior to infecting their hosts, highlighting the importance of sex in their life cycles. Mating loci identified in related smut fungi consist of a pheromone precursor, G-protein-coupled pheromone receptor and two divergently transcribed homeodomain transcription factors. The primary objective of this study was to annotate mating and RNA-silencing genes in available genomes of the systemically infecting bunts *T. caries* and *T. contraversa*, and non-systemically infecting bunts *T. indica* and *T. walkeri*. Phylogenetic comparisons of homeodomain proteins in *Tilletia* species reveal four clades with more than two mating-type homeodomain proteins present in *T. caries* and *T. controversa*. Mating genes identified in each species in single copy include putative pheromone precursors and G-protein-coupled-pheromone-receptors. A high level of protein-sequence homology is seen in comparisons of mating-type genes between *T. caries* and *T. contraversa*, as well as between *T. walkeri* and *T. indica*, however lower homology is present in

comparisons between the two groups. Comparisons of RNA-silencing protein copy numbers in *T. caries* 517 with those of other basidiomycetes reveal an expansion of some RNA-silencing-related gene families in *T. caries*. Preliminary genome annotation was carried out using AUGUSTUS. Predicted proteins were clustered by similarity. Putative mating and RNA-silencing-related genes were identified in each species by homology to genes previously identified in Ustilaginomycotina. Transcript evidence for all mating-genes identified was obtained from cDNA. This work is the first to identify mating genes in *T. caries*, *T. contraversa*, *T. indica* and, *T. walkeri*. While it demonstrates conservation of the mating type genes found in *Ustilago* spp., it also shows that mating-type homeodomain proteins in *T. caries* and *T. controversa* are present in multiple copies, rather than just a divergently transcribed pair. This work provides an informative first look into the genomes of these economically and historically important plant pathogens, and highlights unique molecular features of their mating and RNA-silencing mechanisms, which distinguish them from other fungi.

TABLE OF CONTENTS Page

LIST OF TABLES

CHAPTER TWO TABLES

CHAPTER THREE TABLES

SUPPLEMENTARY TABLES

LIST OF FIGURES

CHAPTER TWO FIGURES

CHAPTER THREE FIGURES

Dedication

This work is dedicated to my grandmother, Dorothy McCotter.

CHAPTER ONE

INTRODUCTION TO COMMON BUNT

Common Bunt Biology

Tilletia caries (syn. *T. tritici*) and *Tilletia laevis* (syn. *T. foetida*) (Tilletiales, Exobasidiomycetidae, Ustilaginomycotina, Basidiomycota), also known as common bunt or stinking smut, are two of the world's most devastating pathogens of cultivated wheat (*Triticum aestivum*). Their genomes range from roughly 28 to 42 Mbp and they possess 19 to 22 chromosomes (Russell and Mills 1993). The dikaryon (n+n) infects the wheat seedling pre-emergence by penetrating the coleoptile (Kühn 1874). Once in the plant, the fungus follows the growing tip, eventually reaching the spikelets. There it grows into the developing embryos, in which it proliferates, eventually forming teliospores. The teliospores are formed at the tips of specialized sporogenous hyphae. Immature spores are dikaryotic and karyogamy occurs as the spores mature (Dangeard 1892; Fischer and Holton 1957). Each infected seed is called a sorus and is full of teliospores. When sori are ruptured in large numbers teliospores are released in foul smelling brown clouds (Goates 1996). The odor, caused by the chemical trimethylamine, is reminiscent of dead fish and gives rise to one of the common names for this disease, stinking smut of wheat (Mitra 1935). Goates and Hoffmann (1987) provide a detailed description of teliospore germination and associated nuclear processes. After a period of dormancy on debris, seeds or in the soil, the teliospores undergo meiosis followed by up to two rounds of mitotic divisions. The teliospores then germinate, giving rise to a haploid promycelium (Tulasne and Tulasne 1847). Each nucleus migrates first into the promycelium (basidium) then into one of four to sixteen primary sporidia (analogous to basidiospores). There the nuclei divide mitotically;

one migrates back into the promycelium and from there into any empty primary sporidium. Nuclei remaining in the promycelium following the formation of mature, nucleate sporidia eventually senesce (Goates and Hoffmann 1987). Individual primary sporidia usually undergo rapid conjugation with a sporidium of compatible mating type. This often occurs between adjacent progeny of the same meiotic event. This process is characterized by plasmogamy followed by migration of one nucleus into the adjacent sporidium via a conjugation peg. Alternatively, fusion may occur between unconjugated, haploid mycelia or secondary sporidia (Kollemorgen et al. 1980). Conjugation between sporidia forms dikaryotic "H-bodies", two filiform sporidia arranged in parallel connected by a conjugation peg and in appearance similar to a letter "H". The dikaryotic fungus either forms infection hyphae or divides mitotically to form secondary sporidia. Secondary sporidia may be allantoid or filiform (Goates 1996). The allantoid form is forcibly discharged in a manner analogous to the discharge of basidiospores on a mushroom (Stolze-Rybczynski et al. 2009). Filiform spores are passively dispersed (Carris et al. 2006). Sporidia may germinate to form another sporidium, or infection hyphae that invade a susceptible seedling, thus completing the life cycle (Saari et al. 1996).

Brief History of Common Bunt

Common bunt has been known since ancient times (Fischer and Holton 1957; Gaudet and Menzies 2012; Tillet 1937; Saari et al. 1996). Like wheat cultivation itself, the common bunt fungi are believed to have originated in the Near East and it is there that they persist as major pathogens to this day (Saari et al. 1996). It is believed that the ancient Greeks and Romans referenced smuts of wheat, particularly Theophrastus, Vergel and Pliny the Second (Fischer and Holton 1957). Wheat bunt continued to be a major problem throughout the Middle Ages, in

endemic rather than epidemic proportions. It was accepted that a proportion of the grain crop would be bunted, though certain seasons and growing regions were especially vulnerable (Large 2003). During the Enlightenment of the 17th and 18th Centuries, attempts were made to control the disease, though it was not yet understood as such, for example by brining seed prior to planting. These strategies met with limited success (Gaudet and Menzies 2012). It was during this period that Mathieu Tillet (1714-1791) demonstrated the contagious nature of plant disease using common bunt. In a series of crude yet elegant experiments he debunked various contemporary theories on the genesis of bunt and showed that teliospores acted as an inoculum source (Tillet 1937). The battle with common bunt continued until the mid-20th century, when it was finally controlled through the development of fungicide seed treatments, though not in time to stop more than half a century's worth of major losses in New World wheat-growing regions. Emerging foci for common bunt in the 19th and 20th centuries included the Pacific Northwest and Midwest of the U.S.A., Western Canada, Argentina, Australia and other New World locations (Fischer and Holton 1957). In the late 1800s some fields in Washington State reported losses as great as ninety percent of production, fifty percent in some fields in Kansas and one to four percent of total U.S. production annually by the early 1900s. This cost farmers millions of dollars and bunt was present in nearly every American wheat field (Fischer and Holton 1957). Losses of this scale were also observed in Old World growing areas such as Britain, Russia and Germany, as well as the Near East and Indian subcontinent (Gaudet and Menzies 2012).

In modern times common bunt has been largely eliminated as an economically important pathogen of wheat in developed countries through the combined use of chemical seed treatments, resistant varieties, and preventative cultural practices (Gaudet and Menzies 2012). These

strategies reduce losses from common bunt to less than one percent where properly employed. However, farmers in many developing countries cannot afford, or do not have access to these control measures and losses can be significantly higher. Particularly problematic in Middle Eastern nations such as Syria, Turkey, Iran and Iraq, bunt is also common in Central Asia, for example in Northern India, Pakistan, Kazakhstan and Nepal (Gaudet and Menzies 2012). It is also present in wheat growing parts of Africa. The disease is managed relatively well in East Asia but has the potential to cause major losses if unchecked (Gaudet and Menzies 2012). The areas considered highest risk for common bunt include the Middle East, North African countries, the Indian Subcontinent, West Asia and the republics of the former Soviet Union. New races of common bunt continue to be discovered in these modern disease foci (Saari et al. 1996). Organically grown wheat is at high risk for common bunt infection. This is especially true in Europe where seed treatment prior to planting is not allowed for organic wheat and the seed lot must be organically grown (Matanguihan, Murphy and Jones 2011)

Common Bunt in Organic Wheat

In the United States, organic wheat acreage has increased by over five hundred percent since 1995 (USDA Economic Research Service 2010) but it is in the European Union where organic agriculture has been most aggressively pursued (Matanguihan et al. 2011). Europe is also where common bunt has emerged as a major limiting factor in organic wheat production, with some areas registering yield losses as high as forty percent when untreated seeds are used (Matanguihan et al. 2011). The EU requirement that seeds used in organic agriculture be themselves organically grown and untreated post-harvest is a major factor contributing to the high incidence of common bunt in European states. In the U.S., conventionally produced seed

may be used in organic production. Thus common bunt is relatively rare in the U.S (Matanguihan et al. 2010). Teliospore contamination is common in European organic seedlots, sometimes to the extent that they cannot be used in organic agriculture (Matanguihan et al. 2011). Seed washes from seedlots destined for organic farming are used to determine the feasibility and extent of approved treatment needed (Bänziger et al. 2004).

New methods of control for common bunt are needed for organically grown wheat. While older organic treatments such as hot water baths are still in use, new seed treatments including powdered milk and mustard flour, have been studied for effectiveness in controlling common bunt (Borgen and Kristensen 2003). Much of the research on these alternative control measures has not been peer reviewed, nor has the economic feasibility been assessed (Emmens 2003). New seed treatments, including organic-approved Tillecur, Chitosan, and electron treatments (in which seeds are bombarded with low energy electrons) are being tested (Röder and Tigges 2005. If applied properly, these can limit disease incidence to less than one percent (Vogt-Kaute and Tilcher 2004). The resurgence of common bunt in organic agriculture has, and may continue to stimulate a revival of research on bunt pathogens.

The Phylogenetic Placement of the Bunt Fungi

A large number of phylogenetic studies have been carried out on basidiomycetes in recent decades however only in the past ten years have robust, multi-locus genetic concordance phylogenetic studies been undertaken as part of the Assembling the Fungal Tree of Life (AFTOL) project (Blackwell et al. 2006; Lutzoni et al. 2004; Spatafora 2005). These studies show strong support for the monophyly of Basidiomycota as well as each of its subdivisions,

Agaricomycotina, Pucciniomycotina and Ustilaginomycotina, which are characterized by the mushrooms, rusts and smuts respectively (Lutzoni et al. 2004).

Smuts and bunts were recognized as belonging to a monophyletic group prior to the advent of molecular systematics. Morphological studies starting in the mid-19th century grouped together the orders we recognize today as Ustilaginomycetes and Exobasidiomycetes (Fischer and Holton 1957). Tulasne and Tulasne (1847) distinguished the holobasidiate Tilletiaceae from the phragmobasidiate Ustilaginaceae based upon basidial septation. More recent studies have drawn distinctions between Pucciniomycotina and Agaricomycotina by comparing their septal pore structures as well as the presence and absence of septal caps and swellings (Bauer et al. 1997; Begerow et al. 2006; Lutzoni et al. 2004). However, morphological characters are inadequate as the sole basis for classification in Ustilaginomycotina at the species level. Fischer and Holton, in their seminal *Biology and Control of the Smut Fungi* (1957) wrote:

> "Compared to some other groups of the fungi, the smut fungi have few morphological characters as a basis for classification and the application of an absolute morphological concept would result in consolidation of species beyond good reason."

While Ustilaginomycotina can be successfully distinguished from Agaricomycotina and Pucciniomycotina by the polysaccharide makeup of the cell wall (glucose dominant, xylose absent) (Prillinger et al. 1993), biochemical markers such as these have proven inadequate to distinguish a number of smuts at the species level, particularly within *Tilletia* (Banowitz et al. 1984; Kawchuk et al. 1988).

DNA sequence data has helped to clarify the broad phylogeny of Ustilaginomycotina and led to the inclusion of apparently asexual genera in this lineage such as *Malassezia* (Begerow et al. 2000). The most recent genetic concordance phylogenetic study of Ustilaginomycotina, by Begerow et al. (2006), recognizes three major lineages: the Ustilaginomycetes, the Exobasidiomycetes and the Entorrhizomycetes. The Ustilaginomycetes consist of the Ustilaginales and the Urocystales, both orders of smut fungi. The Exobasidiomycetes consist of a wider range of taxa including the Exobasidiales, Entylomatales, Cercaseorales, Doassansiales, Microstromatales, Malasseziales, Tilletiales and Georgefisheriales. Many of these are plant parasites however the human and animal pathogenic Malasseziales are an interesting exception (Guého et al. 1996). The last lineage of Ustilaginomycotina is the Entorrhizomycetes with just one order, the Entorrhizomycetales, consisting solely of the genus *Entorrhiza*, a homothallic root pathogen of Cyperaceae and Juncaceae (Begerow et al. 2006).

The genus *Tilletia* (bunt fungi) is one of six genera in order Tilletiales and consists of approximately 140 species. *Tilletia* species only infect hosts from the Poaceae (grasses) and like most members of Ustilaginomycotina, are characterized by chlamydospore-like diploid survival structures termed teliospores (Vánky 2002). Teliospores are usually brown in color and may be smooth, or ornamented. Ornamentation may be reticulate, echinulate, pitted, cerebriform, tuberculate or verrucose and the genus produces both pigmented spores and hyaline sterile cells. *Tilletia* species demonstrate both local and systemic modes of infection (Carris et al. 2006). All genera (*Conidiosporomyces*, *Ingoldiomyces*, *Neovossia*, *Tilletia* and, *Oberwinkleria*) in order Tilletiales are parasitic on Pooid hosts (Zogg 1972) with one exception, *Erratomyces*, which parasitizes hosts in the Fabaceae, forming spores in the leaves (Piepenbring and Bauer 1997).

Teliospore germination in *Tilletia* species is characterized by formation of a promycelium culminating in a whorl of primary sporidia, analogous to basidiospores. These may be conjugating or nonconjugating depending upon the species. Locally infecting *Tilletia* species such as *Tilletia indica* and *Tilletia walkeri* produce large numbers of non-conjugating primary sporidia while the systemically infecting bunts such as *T. caries*, *T. laevis* and *T. contraversa* produce 4-16 conjugating primary sporidia (Goates 1996). Other genera within order Tilletiales also produce large numbers of non-conjugating primary sporidia--*Conidiosporomyces* spp., *Neovossia* spp. and *Oberwinkleria* spp. (Vánky and Bauer 1992; Castlebury and Carris 1999; Durán 1987). These non-conjugating species rely on the forcible discharge of secondary haploid allantoid sporidia to reach the infection court. Sporidial discharge is analogous to that of basidiospore discharge in Agaricomycotina as described in Stolze-Rybczynski et al. (2009). For the non-conjugating species, plasmogamy and dikaryon formation is delayed, occurring on or in the host, rather than occurring shortly after basidiospore formation (Carris et al. 2006).

The first molecular phylogenetic study of Tilletiales was carried out by Castlebury et al. (2005) utilizing the nuclear large subunit ribosomal gene. This study included Poaceae-infecting Tilletiales of both conjugating and non-conjugating species from genera *Tilletia*, *Neovossia* and *Ingoldiomyces* as well as Fabaceae infecting *Erratomyces* spp. Those species with germination and conjugation behavior similar to *T. caries* were primarily Pooid infecting but both conjugating and nonconjugating species were spread throughout the host range, infecting Pooid, Panicoid, Chloridoid and Erhartoid hosts. There was also insufficient molecular evidence to classify *Neovossia* separately from *Tilletia*. The phylogeny largely followed that of the host plants (Castlebury et al. 2005). *T. indica* and *T. walkeri* were part of the Pooid infecting clade,

but were not as well supported. Multiple spore ornamentation types were present in this clade including reticulate and verrucose-spored species. The non-Pooid infecting species did not form well-supported groups corresponding to host family, nor were there clades that strictly corresponded to germination pattern. Castlebury et al. (2005) concluded that the non-conjugating germination pattern is likely ancestral, and the conjugating pattern arose multiple times throughout the Tilletiales' evolutionary history.

MATING IN BASIDIOMYCETES

Introduction

Interest in sexual compatibility dates to the work of deBary (1866) and Brefeld (1883). It is driven in part by the economic importance of smuts as agricultural pathogens and the importance of mating in pathogenesis (Fischer and Holton 1957). Development of tests to discern successful mating interactions allowed strains of smut fungi to be divided into compatibility groups based on their ability to mate *in vitro* (Puhalla 1968) or complete their life-cycles on their host (Fischer and Holton 1957). Terminology was developed to describe the different mating interactions observed amongst the progeny of a single meiotic event when paired in all possible combinations, for example the terms "bipolar" and "tetrapolar". When crosses demonstrate that two independently segregating loci control compatibility, the fungi are termed tetrapolar, or multifactorial; when one locus controls compatibility the fungi are termed bipolar, or unifactorial. Thus, the two systems are defined by number mating-type loci and the presence of two or multiple alleles at these loci (Rowell 1955; Rowell and DeVay 1954; Halisky 1965). Most smut fungi with a bipolar mating system have only two mating-type alleles present in their populations but some species, such as *T. contraversa,* demonstrate bipolar mating behavior with more than two alleles (Hoffmann and Kendrick 1969). The phenomenon of multi-allelism is observed by performing crosses between strains from different populations (Halisky 1965). It was not until the late 1980s to mid-1990s that the molecular basis of the genetic phenomena of mating-type was elucidated, primarily through studies utilizing the corn smut fungus, *Ustilago maydis*. Studies with the barley smut fungus, *Ustilago hordei* provided verification of the differences between bipolar and tetrapolar mating systems with regards to

genetic linkage (Bakkeren and Kronstad 1994). The molecular underpinnings of biallelic versus multiallelic mating systems were determined through the study of other smut fungi (Schirawski et al. 2005; Urban et al. 1993; Kellner et al. 2011). These breakthroughs are discussed in the following section.

Molecular Basis of Mating in **Ustilago maydis** *and its Close Relatives*

U. maydis is an excellent model system for the study of fungal mating and pathogenicity due to its fast growth, easily discernable mating reaction, and ability to form sori on vegetative parts of infected maize plants (Feldbrügge et al. 2013; Puhalla 1968; Rowell 1955; Thakur et al. 1989). *Ustilago maydis* has a tetrapolar mating system controlled by a biallelic locus termed "*a*" and a multiallelic locus termed "*b*" (Bakkeren et al. 2008). A compatible mating interaction with different alleles present at both loci is required for infection of the maize host (Rowell and DeVay 1954). The complementary *a*-mating-type (MAT-a) alleles are involved in direction of conjugation tubes prior to plasmogamy (Snetselaar et al. 1996) as well as maintenance of filamentous growth following formation of the dikaryon (Banuett and Herskowitz 1989). Complementary b-mating-type (MAT-b) alleles are required for development of infection hyphae, formation of tumors (sori) on maize and production of teliospores (Rowell 1955; Rowell and DeVay 1954).

Sequencing and molecular characterization of the *a* locus in *U. maydis* revealed that the two alleles have little sequence homology when compared to flanking sequences but occupy the same location on the chromosome, measuring 4.5 kb and 8 kb for MAT-a1 and MAT-a2 alleles respectively. Hence, these are termed "ideomorphs" (Froeliger and Leong 1991). Two genes at

the MAT-a locus, *MFA* (mating factor a) and *PRA* (pheromone receptor a) encode respectively a short pheromone mating factor, and a membrane-bound g-protein-coupled pheromone receptor (GPCR) which recognizes and binds pheromones from compatible mating partners. The receptor includes a number of transmembrane domains and has homology to the STE3 protein of *Saccharomyces cerevisiae*. Each allele encodes alternate versions of these genes (Bölker et al. 1992). The secreted pheromone initiates plasmogamy by binding to the compatible STE3-type GPCR. It is post-translationally modified to form the active pheromone via the farnesylation and the carboxy-methyl esterification of cysteine at its C-terminal Cys-aliphatic-aliphatic-X motif, as well as by the cleavage of the majority of the N-terminal residues prior to this motif and the residues following the terminal cysteine (Spellig et al. 1994).

Molecular characterization of alleles of the *b* locus in *U. maydis* revealed the presence of multiple genes important for mating (Kronstad and Leong 1989). These include genes coding for DNA binding proteins known as homeodomain proteins (HDPs) that recognize a specific DNA sequence termed a "homeobox" (Scott et al. 1989). Homeodomain transcription factors were first characterized in *Drosophila* where they are associated with the antennapedia and bithorax phenotypes (McGinnis et al. 1984). More generally, homeodomain transcription factors are involved in animal limb patterning (Scott et al. 1989). HDPs include a conserved Trp-Phe-x-Asn-x-Arg (WFxNxR) motif important to DNA binding (Scott et al. 1989). The first of these factors at the locus of *U. maydis* was identified via deletion and complementation (Schulz et al. 1990). The open reading frame identified demonstrated strong homology to the HMG-box mating-type protein in *Saccharomyces cerevisiae*. When paired with a MAT-b allele from a compatible mating strain, the encoded HDP was shown to be indispensable for the switch from

yeast-like to filamentous growth in *U. maydis* (Schulz et al. 1990). An additional homeodomain transcription factor at the *b* locus was later identified, which aside from its conserved WFxNxR motif, has minimal homology to its neighboring, divergently transcribed HDP regardless of the allele from which it is sourced (Gillissen et al. 1992). These two proteins are termed bEast (bE) and bWest (bW) followed by a number to indicate the *b* allele of origin (Bakkeren et al. 2008). Only when bE and bW from opposite mating types are paired is filamentous growth and pathogenesis initiated (Gillissen et al. 1992). This led to the hypothesis that these proteins form a heterodimer in their active state. Interestingly, only one complete pair of complementary bE and bW proteins is required for pathogenesis despite the fact that each allele encodes both a bW and bE type protein with the potential to form two distinct heterodimers with mating-type HDPs from a compatible mating partner (Gillissen et al. 1992). Alleles of both bE and bW share areas of high sequence homology at their homeodomain containing C-termini but differ markedly at their N-termini (Schulz et al. 1990; Gillissen et al. 1992). Further work indicated that the sequences determining allelic specificity lie in the hypervariable N-terminal regions of these proteins (Yee and Kronstad 1993). The nature of the interaction between the bW and bE proteins was elucidated when a yeast two-hybrid assay was used to confirm that bE and bW proteins from opposite mating types heterodimerize to become transcriptionally active (Kämper et al. 1995).

Although the pheromone response pathway is required to initiate plasmogamy and upregulate transcription of the b-type homeodomain proteins, an additional nutritional cue provided by cyclic AMP (cAMP) is also required (Gold et al. 1994). In *U. maydis*, extracellular cAMP binds to a membrane bound GPCR, Gpa3 (distinct from the STE3-type GPCR), stimulating adenylyl cyclase Uac1 and increasing intracellular cAMP levels. This activates

protein kinase A (PKA) ADR1 via dissociation of its inhibitory subunit UBC1, leading to elevated levels of pheromone response factor one (PRF1), a transcription factor (*S. cerevisiae* STE12 ortholog) binding to specific DNA sequences near mating genes termed pheromone response elements (PRE) (Müller et al. 2003). The action of PRF1 up-regulates transcription of pheromones, receptors and b-type HDPs (Moore et al. 2011). In low-cAMP conditions, this cascade suppresses filamentous growth. As the name implies, PRF1 also depends on binding by compatible MFA to the STE3-type GPCR PRA previously mentioned. This cascade is under the control of a mitogen-activated protein (MAP) kinase module consisting of three kinases: KPP2, FUZ7 (*S. cerevisiae* STE7 ortholog) and UBC2 (Bölker et al. 2003). For mating to succeed, both nutritional cues and partner-generated compatibility cues must be present in the cell's immediate environment. This cascade is broadly conserved amongst investigated basidiomycetes; however synteny of the genes encoding these peptides is not (James et al. 2013). For example in the bipolar human pathogen *Filobasidiella neoformans*, genes involved in the MAP kinase cascade, which regulates the pheromone response factor, are co-located at the MAT-locus (Loftus et al. 2005; McClelland et al. 2004; Lengeler et al. 2002).

The Organization of Tetrapolar versus Bipolar Mating Loci and the Generation of Non-Parental Mating Types by Recombination between Mating Loci

Despite elucidation of the molecular mechanisms underlying mating behavior in the tetrapolar *U. maydis*, understanding the observed linkage disequilibrium of mating type loci in bipolar-mating smuts required molecular study of a bipolar species. Bakkeren and Kronstad's (1993) molecular characterization of the mating subloci in *U. hordei*, which causes loose smut of barley, delineated the similarities and differences between bipolar and tetrapolar smut fungi.

Though the fungi demonstrate different numbers of compatibility groups in crosses, the genes at the MAT loci of *U. maydis* and *U. hordei* are orthologous. Introduction of *U. hordei*'s *b* genes into *U. maydis* rendered transformants weakly pathogenic on maize, as would occur following successful mating (Bakkeren and Kronstad 1993). The reason for the observed differences in mating behavior in bipolar smuts is physical linkage of the *a* and *b* mating type subloci, which are approximately 150 Kb apart in *U. hordei* (Bakkeren and Kronstad 1994). Subsequent investigation of MAT-a loci across a wide sample of the Ustilaginales revealed broad conservation of synteny at this locus and three orthologous MAT-a ideomorphs conserved across species (Kellner et al. 2011). Recombination between mating subloci in bipolar species has been reported but is rare. One example of recombination in a fungus with a bipolar mating system is found in *Sporidiobolus salmonicolor*, a "red yeast" in Pucciniomycotina, which undergoes occasional recombination between MAT-a and MAT-b subloci (Coelho et al. 2010). In general mating systems in Pucciniomycotina are not well characterized; however studies on the red yeasts have begun to shed light on mating in this economically important subphylum. Mating systems in Pucciniomycotina are discussed in greater detail later in this chapter.

Multiallelic Mating Systems

Multiple ideomorphs at MAT-a have been molecularly characterized in a number of Ustilaginomycetes including *Sporisorium reilianum* (Schirawski et al. 2005), *U. hordei* and *Ustanciosporium gigantosporum* (Kellner et al. 2011). In tetrapolar smuts, it is common to have more than one ideomorph at the MAT-b locus (Bakkeren et al. 2008). Smuts that are multiallelic at MAT-a encode an extra pheromone precursor protein specific to a third pheromone receptor found in the second compatible mating type. This organization of pheromone genes was first

elucidated in *S. reilianum* (Schirawski et al. 2005) and has since been confirmed in a variety of other Ustilaginales (Kellner et al. 2011). Further study of the MAT-a locus in *U. maydis* demonstrated that this fungus likely possessed a second pheromone gene which has been subsequently reduced to a pseudogene, raising the possibility that its bi-allelic *a* locus may have been multiallelic at some point during its evolution (Urban et al. 1996). The presence of multiple pheromone genes in one MAT-a ideomorph, each demonstrating specificity to a different pheromone receptor, is characteristic of multiallelic mating systems. The high frequency of this arrangement in sequenced Ustilaginales indicates that multiallelism is the mode for MAT-a in this family and that the single functional pheromone of *U. maydis* is an exception (Bakkeren et al. 2008; Kellner et al. 2011).

Mating in Agaricomycotina: Molecular Basis and Manifestations in Phenotype

The molecular interactions characterized in *U. maydis* have a great deal in common with those found in other basidiomycetes. In most agaricomycetes, the homeodomain-protein encoding and pheromone/pheromone receptor loci are designated A and B, respectively (Moore et al. 2011), in contrast to the smut fungi where they are referred to as *b* and *a*, respectively (Bakkeren et al. 2008). This inversion of terminology can be confusing, but the letters denoting the respective MAT-loci in agaricomycetes are always capitalized. In Agaricomycotina, features such as mating-type and the general molecular mechanisms of mating (the cleaved, prenylated pheromone, STE3-type GPCR, b-type homeodomain proteins, and presence of multiple ideomorphs) are conserved, but the organization of the genes at mating loci tends to be considerably more complex (Moore et al. 2011).

The mating systems of *Coprinopsis cinerea* and *Schizophyllum commune* are particularly well studied (Moore et al. 2011). The A locus of *C. cinerea* contains three sets of divergently transcribed homeodomain transcription factors located at two subloci (Aα and Aβ; two at one, one at the other) just over 5kb apart. The Bα and Bβ subloci both encode associated pheromones and GPCRs. In a mating interaction, alternative specificity at just one of the A genes is sufficient to initiate A-regulated development. *S. commune*'s second sublocus within its A locus (Aβ) encodes six additional homeodomain transcription factors (Moore et al. 2011; Ohm et al. 2010). The pheromone and pheromone receptor genes found at the B mating subloci in *S. commune* and *C. cinerea* are also present in multiple redundant copies. Six pheromone receptors are present in *S. commune*, some of which are apparently not related to mating functions (Ohm 2010). In the heterokaryon, different specificities at A are required for nuclear pairing, clamp connection formation, conjugate divisions and clamp connection septation while heterokaryons with identical B ideomorphs show inhibited nuclear migration and clamp connections that cannot fuse at their tips (Moore et al. 2011). Isolates with compatible B (pheromone/GPCR) ideomorphs but incompatible A (HDP) ideomorphs fuse but fail to develop into the dikaryon, instead demonstrating the "flat" culture phenotype described by Vaillancourt et al. (1997).

The majority of sequenced members of Agaricomycotina maintain the *Ustilago*-like configuration of two, divergently transcribed homeodomain proteins at MAT-A; however many taxa in orders Agaricales and Russulales demonstrate tandem duplication of the mating-type genes at both their HDP-containing and pheromone and GPCR-containing loci (James et al. 2013). Duplications of genes at each MAT locus (e.g. A and B) are termed subloci (e.g. α or β), similar in concept to the homeodomain and pheromone receptor subloci in bipolar members of

Ustilaginomycotina. In Agaricomycotina, these individual subloci may be present as different ideomorphs, with large numbers present in the population and recombination potentially leading to hundreds of mating types (Moore et al. 2011).

The level of gene duplication at mating loci in Agaricomycotina is generally far greater than that described in smuts; however duplication is not universal in Agaricomycotina. For example, *Pleurotus djamor* has only one pair of divergently transcribed homeodomain transcription factors at the A locus, similar to the organization found in *U. maydis*. Conversely, the B locus encodes three pheromone receptors and at least one pheromone (James et al. 2004). The single pair of homeodomain transcription factors at the A locus is also seen in *Laccaria bicolor*, *Lentinula edodes*, *Pholiota macrospora* and *Postia placenta* (Niculita-Hirzel et al. 2008; Au et al. 2014; Martinez et al. 2009; Yi et al. 2009). Two subloci, each with a pair of HDPs, are seen in *Coprinellus disseminatus* and *Flammulina velutipes* (James et al. 2006; Van Peer et al. 2011). Three pairs of HDPs are found at the A locus in *Coprinopsis cinerea* (Pardo et al. 1996) and eight at both A subloci in *S. commune* (Ohm et al. 2010). Within a single agaricomycete genome, there are often a large number of GPCR orthologs, some located far from the mating loci (James et al. 2013). In twenty species investigated by James et al. (2013) the average number of STE3 orthologs per genome was 6, however often there are only one or two STE3 orthologs located at MAT-B. Many STE3 orthologs located distally to MAT-B are dispensable for mating and their function is not fully understood. Each STE3 receptor is usually paired with one or more pheromone precursors. This also applies to those receptors that are apparently unrelated to mating (James et al. 2013). These examples show that wide arrays of A-and B-locus architecture and MAT-gene content are possible in Agaricomycotina.

Extensive genome sequencing of mushroom-forming fungi shows that, like the smut fungi, species in Agaricomycotina demonstrate both bipolar and tetrapolar mating systems (James et al. 2013; Au et al. 2014). The underlying genetic basis for bipolar versus tetrapolar heterothallism depends on two different mechanisms. One mechanism for bipolar heterothallism is the tight linkage of the mating-type loci as described in Ustilaginales (Bakkeren et al. 2008). Another mechanism for bipolar heterothallism identified by Aimi et al. (2005) and James et al. (2006) and thus far considered unique to Agaricomycotina, involves the loss of recognition function of the pheromone and pheromone-receptor genes at the B-mating locus. In *Pholiota macrospora*, linkage mapping revealed that the pheromone and receptor loci are unlinked as would be expected in a tetrapolar species; however crosses demonstrated only two compatibility groups, and specific pheromone receptor alleles do not correlate with mating type as assayed in crosses (Aimi et al. 2005). A similar situation is encountered in the bipolar basidiomycete *Coprinellus disseminatus* where only the homeodomain transcription factor genes segregate with mating type, although this fungus acts like a bipolar species in crosses (James et al. 2005). The level of recombination suppression between the two MAT loci in bipolar species determines the number of mating types that may exist in the population. Any time recombination between mating type genes or loci occurs there is potential generation of a new mating type with altered heterospecificity. In Agaricomycotina recombination may even occur between adjacent duplicates of the same type of mating gene (e.g. between Aα and Aβ), thus generating even more nuanced mating types with populations of some species containing hundreds of unique "sexes". This has been observed to be the case in *S. commune* and *C. cinerea*, especially with regards to the HDP-encoding A locus (Moore et al. 2011). However, *Pleurotus djamour* does not exhibit the same high level of A locus recombination. Instead recombination occurs more readily at its B

locus which contains a pheromone and three GPCRs (James et al. 2004). For both *S. commune* and *C. cinerea*, frequency of recombination within the A and B loci between their subloci varies depending on the ideomorph and culture conditions (Raper et al. 1960; Day 1960).

Mating in Pucciniomycotina

Mating systems in Pucciniomycotina studied thus far demonstrate interesting variation relative to Agaricomycotina and Ustilaginomycotina. Coelho et al. (2011) analyzed 216 strains of red yeasts representing 32 species of *Rhodotorula*, *Rhodosporidium*, *Sporobolomyces*, and *Sporidiobolus* from order Sporidiobolales exhibiting selfing, outcrossing and asexual phenotypes. All sexual strains of the red yeasts demonstrate bipolar heterothallism in crosses and the organization of the HDP and pheromone/GPCR encoding subloci at the MAT-locus reflected this behavior. An exceptional level of variability was encountered in the homeodomain encoding locus. Nearly every strain analyzed encoded a different ideomorph at this sublocus (Coelho et al. 2011). The high level of interspecies variability at HD makes phylogenetic comparison difficult and suggests that a large amount of diversification occurred in these taxa post-divergence. On the other hand, the PR sublocus demonstrates a far higher degree of similarity indicating that polymorphism likely arose anciently and prior to speciation (Coelho et al. 2011). While these yeast are primarily bipolar, as discussed previously for *S. salmonicolor*, occasional recombination between mating subloci leading to the generation of new mating types is noted (Coelho et al. 2010). Strains of an additional species, *Rhodosporidium babjevae* also showed evidence of occasional recombination between its PR and HD containing subloci providing further evidence for pseudo-bipolarity in Pucciniomycotina (Coelho et al. 2011).

The anther smut fungus, *Microbotryum violaceum*, now recognized as belonging in order Microbotryales in Pucciniomycotina (Bauer et al. 2006), demonstrates bipolar heterothallism and possesses one pheromone receptor gene, one pheromone gene and a pair of divergently transcribed HDPs (Giraud et al. 2008). Like many animals, it has dimorphic sex chromosomes of different sizes that segregate with mating type (Giraud et al. 2008). The MAT-loci in *M. violaceum* demonstrate a high degree of synteny with those of the red yeasts as well as suppression of recombination. Unlike the red yeasts, pheromone receptors in *Microbotryum* are highly differentiated (Petit et al. 2012). In the fern pathogen *Mixia osmundae* (Mixiales), for which only one mating type has been discovered thus far and sexual reproduction is unknown, the homeodomain protein-encoding and pheromone/GPCR-encoding loci are organized in the same fashion as in *U. maydis*, with a pheromone and GPCR gene pair at one locus and two divergently transcribed HDPs at the other (Toome et al. 2013). However, the linkage status of these loci has yet to be elucidated, and because they are found on different scaffolds in the genome assembly, it is not possible to determine without further sequencing whether *M. osmundae* is tetrapolar or bipolar. Additionally, its mating capability will remain in doubt until a specimen of different mating type is collected (Toome et al. 2013).

Cryptic Sex in Basidiomycetes

As discussed previously with regards to *Mixia osmundae* and the red yeasts, there are a number of basidiomycetes which possess MAT-gene orthologs but for which no mating behavior has yet been observed. Sexual reproduction is not known in the human pathogenic yeasts from Ustilaginomycotina, *Malassezia globosa* and *Malassezia restricta*, and like *Mixia osmundae*, only one MAT ideomorph has been found in collected specimens. Orthologs of all mating type

and downstream signaling genes are present suggesting that *Malassezia* spp. are capable of mating (Xu et al. 2007). In their survey of sexuality in the red yeasts, Coelho et al. (2011) noted that asexual strains lacking both GPCR/pheromone and HDP-containing loci were exceedingly uncommon and most asexual strains possessed these genes. It seems unlikely that these species, many with relatively small genomes, would maintain mating genes unless they served some purpose. Population genetic analyses are needed to determine if there is evidence of recombination in these putatively asexual species. It is possible that mating-type associated fitness differentials, as seen in *Filobasidiella neoformans* (McClelland et al. 2004) may have contributed to the predominance of a single mating type in populations of *Malassezia* and red yeasts that have been studied. *F. neoformans* is also known to undergo parasexual recombination in a process termed "fruiting", which involves mating genes but does not require a compatible partner (Lin et al. 2005). Analogous processes may exist in other fungi for which evidence of sexual reproduction is lacking.

Observations on Mating in Tilletia *Species*

Studies of mating in *Tilletia* species, especially among the wheat bunts, have been extensive; however most of this work has dealt with the establishment of compatibility groups rather than the molecular mechanisms of mating. Excellent reviews of this work may be found in Halisky (1965) and Holton et al. (1968). The type species *T. caries* and its close relative, *T. laevis* have been shown to demonstrate simple bipolar heterothallism, including in crosses between the two species (Hanna 1934; Holton 1951; 1953 I; Holton and Kendrick 1957). However, other studies identifying compatibility groups contradict with regards to the number of groups identified (Flor 1932; Becker; 1936). These contradictions may reflect the difficulty in

species identification given intergrading morphological characters, potentially reflective of high levels of hybridization that may occur in nature (Shi et al. 1996). *T. contraversa*'s mating behavior is more complex than the common bunt fungi and raises important questions relevant to the research presented in Chapter Two regarding the numbers of alleles in the dwarf and common bunt population and whether recombinant mating types occur in *Tilletia* species. Hoffmann and Kendrick (1969) showed by pairing secondary sporidia in culture that pairs of sporidia derived from the same teliospore conjugated approximately 50 percent of the time. They also noted that when crossed with lines derived from teliospores of different collections, five compatibility groups were evident. This indicates that multiple ideomorphs of one locus are likely involved in mating reactions (Hoffmann and Kendrick 1969). When crossed with strains of *T. caries*, nearly all matings were successful. A particularly intriguing aspect of Hoffmann and Kendrick's study (1969) was that some sporidial pairs in hybrid crosses fused successfully then failed to develop infection hyphae. This suggests the existence of a second locus or sublocus controlling compatibility, as has been described for *U. maydis* (Rowell 1955). This type of mating reaction may be analogous to that described above in *S. commune*, in which the pheromone/GPCR loci of the two partners are compatible but the HDP loci are not (Vaillancourt et al. 1997). In *T. caries*, *U. hordei* and other bipolar species, sporidial fusion is usually followed by formation of the infectious dikaryon (Fischer and Holton 1957; Bakkeren 2008). In bipolar species, this is ensured because the linkage of the mating type subloci prevents the formation of new mating type alleles via recombination between them (Bakkeren and Kronstad 1994). Although *T. contraversa* behaves as a bipolar strain in crosses among the progeny of a single spore, it appears to have multiple alleles in outcrosses and in hybrid crosses. The failure of some hybrid crosses to progress beyond the fusion stage (Hoffmann and Kendrick 1969) suggests that

they possess compatible pheromone and GPCR alleles, but incompatible HDP alleles. This implies that recombination between the two subloci may have taken place since the divergence of the crossed species. It would also explain the generation of more than the traditional two compatibility groups. Thus, it is possible that some *Tilletia* species demonstrate pseudo-bipolar or even tetrapolar mating systems in which recombination between the MAT-a pheromone/GPCR-encoding locus and the MAT-b HDP-encoding locus occurs.

Basidiospores of other Pooid-infecting *Tilletia* species do not conjugate, as discussed previously. These include *Tilletia elymi* (Holton 1952) and *Tilletia cerebrina* (Siang 1954) which have multinucleate sporidia and are solopathogenic (Holton and Siang 1956). *Tilletia horrida* (Carris et al. 2006), *T. indica* (Durán and Cromarty 1977; Carris et al. 2006) and *T. walkeri* (Castlebury and Carris 1999) are all non-conjugating, heterothallic bunts with monokaryotic primary sporidia. *Tilletia vankyi* produces non-conjugating monokaryotic primary sporidia which are solopathogenic (Carris et al. 2007).

Hybridization in smut fungi has been well documented. Studies by Fischer (1951; 1953) and Hoffmann and Fischer (1963 I; 1963 II) demonstrated hybridization between a variety of *Ustilago* species in axenic culture. A recent study combined molecular analyses of the MAT-a loci in a variety of Ustilaginales with hybridization assays and showed that the ability of Ustilaginales to form hybrid filaments was widespread but that these fusions were not always viable (Kellner et al. 2011). Just as in *Ustilago*, hybridization in *Tilletia* species has been widely documented. Flor (1932) first demonstrated hybridization utilizing *T. caries* and *T. laevis* and later studies confirmed these results as well as hybridization between *T. contraversa* and *T. caries* (Hoffmann and Kendrick 1969; Holton 1942; Holton 1951; Silbernagel 1964; Trail and

Mills 1990). Pimentel et al. (2000), Carris and Gray (1994) Boyd and Carris (1997) and Holton and Siang (1956) extensively documented hybridization between wheat bunt species and various grass infecting *Tilletia* species from the *Tilletia fusca* complex.

The amount of documented hybridization calls into question the species concepts for these fungi. It has been hypothesized that natural introgression could be the basis for the development of new virulence factors in bunt species (Holton 1942; 1953). Pimentel et al. (2000) detected no evidence for recombination in wild isolates of *T. contraversa* and the cheatgrass pathogen *T. bromi*; however Shi et al. (1996) showed that a number of isolates identified morphologically to represent dwarf bunt actually grouped with common bunt based on their RAPD phenotypes. Host specificity likely presents a barrier to hybridization in the field even if pheromone recognition and conjugation is possible in axenic culture. The host bottleneck, however, does not present a barrier to hybridization for the wheat bunts and it is quite likely that the common and dwarf bunt species intergrade in the field (Bao 2010).

CHAPTER ONE REFERENCES

Aimi T, Yohida R, Ishikawa M, Bao D and, Kitamoto Y. 2005. Identification and linkage mapping of the genes for the putative homeodomain protein hox1 and the putative pheromone receptor protein homologue (rcb1) in a bipolar basidiomycete *Pholiota nameko*. Current Genetics 48: 184-194.

Au CH, Wong MC, Bao D, Zhang M, Song C, Song W, Law PTW, Kües U and, Kwan H. 2014. The genetic structure of the *A* mating-type locus of *Lentinula edodes*. Gene 535: 184-190.

Bakkeren G and Kronstad JW. 1993. Conservation of the *b* mating-type gene complex among bipolar and tetrapolar smut fungi. The Plant Cell 5: 123-136.

——— and Kronstad JW. 1994. Linkage of mating-type loci distinguishes bipolar from tetrapolar mating in basidiomycetous smut fungi. Proceedings of the National Academy of Sciences of the United States of America 91: 7085-7089.

———, Kämper J and, Schirawski J. 2008. Sex in smut fungi: structure function and evolution of mating-type complexes. Fungal Genetics and Biology 45: 515-521.

Banowitz GM, Trione EJ and, Krygier BB. 1984. Immunological comparisons of teliospores of the two wheat bunt fungi, *Tilletia* species using monoclonal antibodies and antisera. Mycologia 76(1): 51-62.

Banuett F and Herskowitz I. 1989. Different *a* alleles of *Ustilago maydis* are necessary for maintenance of filamentous growth but not for meiosis. Proceedings of the National Academy of Science of the United States of America 86: 5878-5882.

Bänziger I, Schachermayr G, Zanetti S, Rüegger A, Forrer H and, Vogelgsang S. c2004. Healthy cereal seeds for organic agriculture in Switzerland. In: Lammerts van Bueren E, Ranganathan R and Sorensen N, editors. IFOAM 2004. Proceedings of the first World Conference on Organic Seed: Challenges and opportunities for organic agriculture and the seed industry. 2004 July 5-7; Bonn, Germany. Rome, Italy: UNFAO, IFOAM, ISF; P.84-87

Bao X. 2010. Host specificity and phylogenetic relationships among *Tilletia* species infecting wheat and other cool season grasses [Doctoral dissertation]. Pullman: Washington State University. 143 p.

Bauer R, Begerow D, Sampaio JP, Weiß M and, Oberwinkler F. 2006. The simple-septate basidiomycetes: a taxonomic synopsis. Mycological Progress 5: 41-66.

———, Oberwinkler F and, Vánky K. 1997. Ultrastructural markers and systematics in smut fungi and allied taxa. Canadian Journal of Botany 75: 1273-1314.

Becker T. 1936. Untersuchungen über sexualität bei *Tilletia tritici* Wint im rahmen der imunitätszuchtung. Phytopathologische Zeitung 9: 187-228.

Begerow D, Bauer R and, Oberwinkler F. 2000. Phylogenetic placements of ustilaginomycetous anamorphs as deduced from nuclear LSA rDNA sequences. Mycological Research 105: 53-60.

————, Stoll M and, Bauer R. 2006. A phylogenetic hypothesis of Ustilaginomycotina based on multiple gene analyses and morphological data. Mycologia 98(6): 906-916.

Blackwell M, Hibbett DS, Taylor JW and, Spatafora JW. 2006. Research coordination networks: A phylogeny for kingdom fungi (Deep Hypha). Mycologia 98(6) 829-837.

Bölker M, Urban M and, Kahmann R. 1992. The *a* mating type locus of *U. maydis* specifies cell signaling components. Cell 68: 441-450.

Borgen A and Kristensen L. c2003-2010. Effect of seed treatment with milk powder and mustard flour in control of common bunt (*Tilletia tritici*) in wheat and stem smut (*Urocystis occulata*) in rye [Internet]. In: Biddle AJ, editor. Organic eprints 2013. Proceedings from BCPC Symposium No. 76: "Seed treatment: Challenges and opportunities". 2001, Tjele, Denmark. Farnham United Kingdom: BCPC. Available from: http://orgprints.org/1115/

Boyd ML and Carris LM. 1997. Molecular relationships among varieties of the *Tilletia fusca* (*T. bromi*) complex and related species. Mycological Research 101(3): 269-277.

Brefeld O. 1883. Botanische Untershuchungen über hefepilze Untersuchungen aus dem Gesammtgebiete der Mykologie V. Die Brandpilze I. vi. Liepzig, Germany.

Castlebury LA and Carris LM. 1999. *Tilletia walkeri*, a new species on *Lolium multiflorum* and *L. perenne*. Mycologia 91: 121-131.

———, Carris LM and, Vánky K. 2005. Phylogenetic analysis of Tilletia and allied genera in order Tilletiales (Ustilaginales; Exobasdiomycetidae) based on nuclear LSU rDNA sequences. Mycologia 97: 888-900.

Carris LM, Castlebury LA and, Goats BJ. 2006. Nonsystemic bunt fungi – *Tilletia indica* and *T. horrida*: A review of history, systematics and biology. Annual Review of Phytopathology 44: 113-133.

———, Castlebury LA, Huang G, Alderman SC, Luo J and, Bao X. 2007. *Tilletia vankyi*, a new species of reticulate-spored bunt fungus with non-conjugating basidiospores infecting species of *Festuca* and *Lolium*.

——— and Gray PM. 1994. The ability of *Tilletia fusca* to hybridize with the wheat bunt species under axenic conditions. Mycologia 86(1): 157-163.

Coelho MA, Gonçalves P and, Sampaio JP. 2011. Evidence for maintenance of sex determinants but not of sexual stages in red yeasts, a group of early diverged basidiomycetes. BMC Evolutionary Biology 11: 249-263.

———, Sampaio JP and, Gonçalves. 2010. A deviation from the bipolar-tetrapolar mating paradigm in an early diverged basidiomycete. PLOS Genetics 6(8): e1001052.

Dangeard PA. 1892. Recherches sur la reproduction sexuelle des champignons. Le Botaniste Ser 3 3: 221.

Day PR. 1960. The structure of the A mating-type locus in *Coprinus lagopus*. Genetics 45: 641-651.

De Bary. 1866. Morphologie und Physiologie der Pilze, Flechten und Myxomyceten. W. Hofmeister: Handbuch der Physiologischen Botanik 2 xxi.

Durán R. 1987. Ustilaginales of Mexico. Pullman, WA: Washington State University Press. 331 p.

——— and Cromarty R. 1977. *Tilletia indica*: a heterothallic bunt fungus with multiple alleles controlling incompatibility. Phytopathology 67: 812-815.

Feldbrügge M, Kellner R and, Schipper K. 2013. The biotechnological use and potential of plant pathogenic smut fungi. Applied Microbiology and Biotechnology 97: 3253-3265.

Fischer GW. 1951. Induced hybridization in graminicolous smut fungi I. *Ustilago hordei* x *Ustilago bullata*. Phytopathology 41: 839-853.

———. 1953. Induced hybridization in graminicolous smut fungi II. *Ustilago striiformis* x *Ustilago bullata*. Phytopathology 43: 547-550.

——— and Holton CS. 1957. Biology and Control of the Smut Fungi. Ronald Press, New York, NY.

Flor HH. 1932. Heterothallism and hybridization in *Tilletia tritici* and *T. levis*. Journal of Agricultural Research. 44: 49-58.

Froeliger EH and Leong SA. 1991. The *a* alleles of *Ustilago maydis* are ideomorphs. Gene 100: 113-122.

Gillissen B, Bergemann J, Sandmann C, Shroeer B, Bölker M and, Kahmann R. 1992. A two-component regulatory system for self/non-self recognition in *Ustilago maydis*. Cell 68: 647-657.

Giraud T, Yockteng R, López-Villavicencio M, Refrégier G and, Hood ME. 2008. Mating system of the anther smut fungus *Microbotryum violaceum*: selfing under heterothallism. Eukaryotic Cell 7(5): 765-775.

Goates BJ. 1996. Common bunt and dwarf bunt. In: Wilcoxson RD, Saari EE, Hettel GP and Mcnab A, Eds. Bunt and Smut Diseases of Wheat: Concepts and Methods of Disease Management. Texcoco, Mexico: CIMMYT. 88 p.

Gold S, Gillian D, Barrett K and, Kronstad J. 1994. cAMP regulates morphogenesis in the fungal pathogen *Ustilago maydis*. Genes and Development 8: 2805-2816.

Guého E, Midgley G and, Guillot J. 1996. The genus *Malassezia* with description of four new species. Antonie van Leeuwenhoek 69: 337-355.

Hanna WT. 1934. The physiology of fungi causing bunt of wheat. Fifth Pacific Science Congress Proceedings: 3195-3204.

Halisky PM. 1965. Physiologic specialization and genetics of the smut fungi. III. The Botanical Review 31(1): 114-150.

Hoffmann JA and Fischer GW. 1963 (I). Induced hybridization in graminicolous smut fungi III. Further studies of *Ustilago hordei* x *Ustilago bullata*. Mycologia 55: 549-562.

——— and Fischer GW. 1963 (II). Induced hybridization in graminicolous smut fungi IV. *Ustilago bullata* x *Ustilago trebouxii*. Mycologia 55: 706-712.

——— and Kendrick EL. 1965. Compatibility relationships in *Tilletia contraversa*. Phytopathology 55: 1061.

——— and Kendrick EL. 1969. Genetic control of compatibility in *Tilletia contraversa*. Phytopathology 59(1): 79-83.

Holton CS. 1942. Extent of pathogenicity of hybrids of *Tilletia tritici* and *T. laevis*. Journal of Agricultural Research 65: 555-563.

———. 1951. Methods and results of studies on heterothallism and hybridization in *Tilletia caries* and *Tilletia foetida*. Phytopathology 41: 511-521.

———. 1952. *Tilletia elymi*, an apparently homothallic species. Phytopathology 42: 635-636.

———. 1953. (I) Fusion between secondary sporidia in culture as an index of sex compatibility in *Tilletia* species. Phytopathology 43: 322-323.

———. 1953. (II) Physiologic specialization and genetics of the smut fungi. II. Botanical Review 19(4): 187-208.

———, Hoffmann JA and, Durán R. 1968. Variation in the smut fungi. Annual Review of Phytopathology. 6: 213-242.

——— and Kendrick EL. 1957. Fusion between secondary sporidia in culture as a valid index of sex compatibility in *Tilletia caries*. Phytopathology 47: 688-689.

——— and Siang WN. 1956. Fusion between primary sporidia derived from different species of *Tilletia* occurring on wheat and certain grass hosts. Res. Stud. State Coll. Wash. 24: 326-330.

James TY, Liou SR and, Vilgalys R. 2004. The genetic structure and diversity of the *A* and *B* mating-type genes from the tropical oyster mushroom, *Pleurotus djamor*. Fungal Genetics and Biology 41: 813-825.

———, Sheng S, Li W, Heitman J, Kuo H, Lee Y, Asiegbu FO, Olson Å and, others. 2013. Polyporales genomes reveal the genetic architecture underlying tetrapolar and bipolar mating systems. Mycologia 105(6): 1374-1390.

———, Srivilai P, Kües U and, Vilgalys R. 2006. Evolution of the bipolar mating system of the mushroom *Coprinellus disseminatus* from its tetrapolar ancestors involves loss of mating-type-specific pheromone receptor function. Genetics 172: 1877-1891.

Kawchuk LM, Kim WK and, Nielsen J. 1988. A comparison of polypeptides from the wheat bunt fungi *Tilletia laevis*, *T. tritici*, and *T. contraversa*. Canadian Journal of Botany 66: 2367-2376.

Kellner R, Vollmeister E, Feldbrügge M and, Begerow D. 2011. Interspecific sex in grass smuts and the genetic diversity of their pheromone-receptor system. PLoS Genetics 7(12): e1002436.

Kollmorgen JF, Owczarzak A and, Trione EJ. 1980. Morphology and timing of secondary sporidial mating in a wheat-bunt fungus *Tilletia caries*. Transactions of the British Mycological Society 75(3): 461-471.

Kronstad JW and Leong SA. 1989. Isolation of two alleles of the *b* locus of *Ustilago maydis*. Proceedings of the National Academy of Sciences of the United States of America 86: 978-982.

Kühn J. 1874. Tilletia controversa. In Rabenhorst, Fungi Europaei. Hedwigia 13: 188.

Lengeler KB, Fox DS, Fraser JA, Allen A, Forrester K, Dietrich FS and, Heitman J. 2002. Mating-type locus of *Cryptococcus neoformans*: A step in the evolution of sex chromosomes. Eukaryotic cell 1(5): 704-718.

Lin X, Hull CM and, Heitman J. 2005. Sexual reproduction between partners of the same mating type in *Cryptococcus neoformans*. Nature 414: 1017-1021.

Loftus BJ, Fung E, Roncaglia P, Rowly D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA and, others. 2005. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science: 307: 1321-1324.

Lutzoni F, Kauff F, Cox CJ, McLaughlin D, Celio G, Dentinger B, Padamsee M, Hibbett D, James TY, Baloch E and, others. 2004. Assembling the fungal tree of life: Progress classification and evolution of subcellular traits. American Journal of Botany 91(10): 1446-1480. doi: 10.3732/ajb.91.10.1446

Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P and, others. 2009. Genome, transcriptome and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulose conversion. Proceedings of the National Academy of Sciences of the United States of America 106(6): 1954-1959.

Matanguihan JB, Murphy KM and, Jones SS. 2011. Control of common bunt in organic wheat. Plant Disease 95(2): 92-103. doi:10.1094/PDIS-09-10-0620.

McClelland CM, Chang YC, Varma A and, Kwon-Chung KJ. 2004. Uniqueness of the mating system in *Cryptococcus neoformans*. Trends in Microbiology 12(5): 208-212.

McGinnis W, Levine MS, Hafen E, Kuroiwa A and, Gehring WJ. 1984 A conserved DNA sequence in homoeotic genes of the *Drosophila* antennapedia and bithorax complexes. Nature 308: 428-433.

Moore D, Robson GD and, Trinci APJ. 2011. 21st Century Guidebook to Fungi. New York, NY: Cambridge University Press. 627 p.

Müller P, Weinzierl G, Brachmann A, Feldbrügge M and, Kahmann R. 2003. Mating and pathogenic development of the smut fungus *Ustilago maydis* are regulated by one mitogen-activated protein kinase cascade. Eukaryotic Cell 2(6): 1187-1199.

Niculita-Hirzel H, Labbé J, Kohler A, le Tacon F, Martin R, Sanders IR and, Kües U. 2008. Gene organization of the mating type regions in the ectomycorrhizal fungus *Laccaria bicolor* reveals distinct evolution between the two mating type loci. New Phytologist 180(2): 329-342.

Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, de Vries RP, Record E, de Vries RP, Record E and, others. 2010. Genome sequence of the model mushroom *Schizophyllum commune*. Nature Biotechnology 28: 957-963.

Olson Å, Aerts A, Asiegbu F, Belbahri L, Bouzid O, Broberg A, Canbäck B, Coutinho PM, Cullen D, Dalman K and, others. 2012. Insight into trade-off between wood decay and parasitism from the genome of a fungal forest pathogen. New Phytologist 194(4): 1001-1013.

Organic Production: USDA Economic Research Service [Internet]. Washington (DC): United States Department of Agriculture: c2012-2013 [cited 2014 Sep 21]. Available from: http://www.ers.usda.gov/data-products/organic-production.aspx.

Pardo EH, O'Shea SF and, Casselton LA. 1996. Multiple versions of the *A* mating type locus of *Coprinus cinereus* are generated by three paralogous pairs of multiallelic homeobox genes. Genetics 144: 87-94.

Petit E, Giraud T, de Vienne DM, Coelho MA, Aguileta G, Amselem J, Kreplak J, Poulain J, Gavori F, Wincker P and, others. 2012. Linkage to the mating-type locus across the genus *Microbotryum*: Insights into nonrecombining chromosomes. Evolution 66(11): 3519-3533.

Piepenbring M and Bauer R. 1997. *Erratomyces*, a new genus of Tilletiales with species on Leguminosae. Mycologia 89: 924-936.

Pimentel G, Peever TL and, Carris LM. 2000. Genetic variation among natural populations of *Tilletia contraversa* and *T. bromi*. Phytopathology 90(4): 376-383.

Prillinger H, Oberwinkler F, Umile C, Tlachac K, Bauer R, Dörfer C and, Traufratzhofer E. 1993. Analysis of cell wall carbohydrates (neutral sugars) from ascomycetous and

basidiomycetous yeasts with and without derivatization. Journal of General and Applied Microbiology 39: 1-34.

Puhalla JE. 1968. Compatibility reactions on solid medium and interstrain inhibition in *Ustilago maydis*. Genetics 60: 461-474.

Raper JR, Baxter MG and, Ellingboe AH. 1960. The genetic structure of the incompatibility factors of *Schizophyllum commune*: The A factor. Proceedings of the National Academy of Sciences of the Unite States of America 46: 833-842.

Rowell JB. 1955. Functional role of compatibility factors and an in vitro test for sexual compatibility with haploid lines of *Ustilago zeae*. Phytopathology 45: 370-374.

Rowell JB and DeVay JE. 1954. Genetics of *Ustilago zeae* in relation to basic problems of its pathogenicity. Phytopathology 44: 356-362.

Russell BW and Mills D. 1993. Electrophoretic karyotypes of *Tilletia caries*, *T. contraversa*, and their F1 progeny: Further evidence of conspecific status. Molecular Plant Microbe Interactions 6(1):66-74.

Saari EE, Mamluk OF and, Burnett, PA. 1996. In: Wilcoxson RD, Saari EE, Hettel GP and Mcnab A, Eds. Bunt and Smut Diseases of Wheat: Concepts and Methods of Disease Management. Texcoco, Mexico: CIMMYT. 88 p.

Schirawski J, Heinze B, Wagenknecht M and, Kahmann R. 2005. Mating type loci of *Sporisorium reilianum*: Novel pattern with three *a* and multiple *b* specificities. Eukaryotic Cell 4(6): 1317-1327.

Schulz B, Banuett F, Dahl M, Schlesinger R, Schäfer W, Martin T, Herskowitz I and, Kahmann R. 1990. The *b* alleles of *U. maydis*, whose combinations program pathogenic development, code for polypeptides containing a homeodomain-related motif. Cell 60: 295-306.

Scott MP, Tamkun JW and, Hartzell GW, III. 1989. The structure and function of the homeodomain. Biochimica et Biophysica Acta 989: 25-48.

Silbernagel MJ. 1964. Compatibility between *Tilletia caries* and *T. contraversa*. Phytopathology 54, 1117-1120.

Shi YL, Loomis P, Christian D, Carris LM and, Leung, H. 1996. Analysis of the genetic relationships among the wheat bunt fungi using RAPD and ribosomal DNA markers. Phytopathology 86: 311-318.

Spatafora JW. 2005. Assembling the fungal tree of life (AFTOL). Mycological Research 109: 755-756.

Spellig T, Bölker M, Lottspeich F, Frank RW and, Kahmann R. 1994. Pheromones trigger filamentous growth in *Ustilago maydis*. The EMBO Journal 13(7): 1620-1627.

Stolze-Rybczynski JL, Cui Y, Stevens MHH, Davis DJ, Fischer MWF and Money NP. 2009. Adaptation of the spore discharge mechanism in the Basidiomycota. PLoS ONE 4(1): e4163. doi:10.1371/journal.pone.0004163

Toome M, Ohm RA, Riley RW, James TW, Lazarus KL, Henrissat B, Albu S, Boyd A, Chow J, Clum A and, others. 2013. Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. New Phytologist 202: 554-564.

Trail F and Mills D. 1990. Growth of haploid *Tilletia* strains *in planta* and genetic analysis of a cross of *Tilletia caries* x *T. contraversa*. Phytopathology 80: 367-370.

Tulasne L and Tulasne C. 1847. Mémiore sur les Ustilaginées compares Uredinées. Annales des Sciences Naturelles-Botanique et Biologie Vegetale 3: 12-127.

Urban M, Kahmann R and, Bölker M. 1996. The biallelic *a* mating type locus of *Ustilago maydis*: Remnants of an additional pheromone gene indicate evolution from a multiallelic ancestor. Molecular and general Genetics 250: 414-420.

Van Peer AF, Park S, Shin P, Jang K, Yoo Y, Park Y, Lee B, Sung G, James TY and, Kong W. 2011. Comparative genomics of the mating-type loci of the mushroom *Flammulina velutipes* reveals widespread synteny and recent inversions. PLoS One 6, e22249.

Vánky K and Bauer R. 1992. *Conidiosporomyces*, a new genus of Ustilaginales. Mycotaxon 43: 427-436.

Vaillancourt LJ, Raudaskoski M, Spech CA and, Raper CA. 1997. Multiple genes encoding pheromones and a pheromone receptor define the Bβ1 mating-type specificity in *Schizophyllum commune*. Genetics 146(2): 541-551.

Vogt-Kaute W and Tilcher R. 2004. Control of common bunt in wheat (*Tilletia caries*) by alternative seed treatment. In: Martinez A, ed. Challenges and Opportunities for Organic Agriculture and the Seed Industry. Turrialba, Costa Rica: Bib. Orton IICA/CATIE. 188 p.

Xu J, Saunders CW, Hu P, Grant RA, Boekhout T, Kuramae EE, Kronstad JW, Deangelis YM, Reeder NL, Johnstone KR and, others. 2007. Dandruff-associated *Malassezia* genomes reveal convergent and divergent virulence traits shared with plant and human fungal pathogens. Proceedings of the National Academy of Sciences of the United States of America: 104(47): 18730-18735.

Yee AR and Kronstad JW. 1993. Construction of chimeric alleles with altered specificity at the *b* incompatibility locus of *Ustilago maydis*. Proceedings of the National Academy of Science of the United States of America 90: 664-668.

Yi R, Tachikawa M, Mukaiyama H, Bao D and, Aimi T. 2009. Genomic structure of the *A* mating-type locus in a bipolar basidiomycete *Pholiota nameko*. Mycological Research 113: 240-248.

Zogg H. 1972. Die *Tilletia*-streifenbrandkrankheiten der grässer. Phytopathologische Zeitschrift 74: 218-229.

CHAPTER TWO

INTRODUCTION

Sexual reproduction is one of the primary drivers of evolution (Heitman 2006) and has been shown experimentally to improve adaptation by fungi to new environments (Goddard et al. 2005). Sexual reproduction in basidiomycetes leads to dikaryon formation. In plant pathogenic smut fungi from subclasses Ustilginomycetidae and Exobasidiomycetidae, sexual reproduction is a prerequisite for infection of the host plant (Bakkeren et al. 2008; Kahmann and Kämper 2004; Fischer and Holton 1957). Among the smut fungi, sexual reproduction has been best studied in *Ustilago maydis*, the corn smut fungus (Feldbrügge et al. 2004).

In *U. maydis*, two loci determine mating compatibility (Puhalla 1968; Rowell and DeVay 1954). One locus encodes a G-protein-coupled pheromone receptor (GPCR) and a pheromone precursor peptide (MFA) (Bölker et al. 1992; Spellig et al. 1994; Snetselaar et al. 1996) while the other encodes a pair of divergently transcribed homeotic, or homeodomain proteins (HDPs) (Kronstad and Leong 1989; Schulz et al. 1990; Gillissen et al. 1992). These loci will henceforth be referred to as MAT-a and MAT-b respectively. Compatibility relies on the presence of different alleles, also termed ideomorphs (Froeliger and Leong 1991), at each of these loci (Rowell and DeVay 1954). Two ideomorphs are present at MAT-a and multiple ideomorphs are present at MAT-b (Puhalla 1968, Rowell and DeVay 1954). Only when strains of different specificities at these loci mate can the resulting dikaryon infect its maize host (Rowell and DeVay 1954).

The membrane-bound GPCR at MAT-a includes a number of transmembrane domains and has homology to the STE3 protein of *Saccharomyces cerevisiae*. Each ideomorph encodes

alternative versions of these genes (Bölker et al. 1992). The pheromone precursor, MFA, also located at MAT-a, initiates the processes leading to plasmogamy by binding to the GPCR of a compatible mating partner. It is post-translationally modified to form the active pheromone via the farnesylation and the carboxy-methyl esterification of cysteine at its C-terminal Cys-aliphatic-aliphatic-X motif, as well as by the cleavage of the majority of the N-terminal residues prior to this motif and the residues following the terminal cysteine (Spellig et al. 1994).

The HDP genes at MAT-b encode a conserved Trp-Phe-x-Asn-x-Arg (WFxNxR) motif important for homeobox-sequence binding (Scott et al. 1989). These HDPs act as transcription factors. The first of these factors at the locus of *U. maydis* demonstrates strong homology to the HMG-box mating-type protein in *Saccharomyces cerevisiae*. When paired with proteins encoded by a MAT-b ideomorph from a compatible mating strain, the HDP is indispensable for the switch from yeast-like to filamentous growth in *U. maydis* (Schulz et al. 1990). The additional homeodomain transcription factor present at MAT-b has minimal homology to its neighboring, divergently transcribed HDP gene aside from its conserved WFxNxR motif regardless of the allele from which it is sourced (Gillissen et al. 1992). These two proteins are termed bEast (bE) and bWest (bW) followed by a number to indicate the b allele of origin (Bakkeren et al. 2008). When bE and bW from compatible mating types are paired, filamentous growth and pathogenesis are initiated (Gillissen et al. 1992). The heterodimer of bE and bW from compatible mating strains forms an active transcription factor (Kämper et al. 1995). A single heterodimer of complementary bE and bW proteins is required for pathogenesis even though each ideomorph encodes both a bW and bE type protein, meaning it can form two distinct heterodimers with mating-type HDPs from a compatible mating partner (Gillissen et al. 1992). Orthologs of both

bE and bW share areas of sequence homology at their homeodomain containing C-termini but differ markedly at their N-termini (Schulz et al. 1990; Gillissen et al. 1992). The sequences determining allelic specificity correspond to the hypervariable N-terminal regions of these proteins (Yee and Kronstad 1993).

Successful mating also depends on a nutritional cue supplied by cyclic AMP (Gold et al. 1994). In *U. maydis*, extracellular cAMP binds to a membrane-bound GPCR Gpa3, stimulating adenylyl cyclase Uac1 and increasing intracellular cAMP levels. This activates protein kinase A (PKA) ADR1 via dissociation of its inhibitory subunit UBC1, leading to elevated levels of pheromone response factor one (PRF1), a transcription factor (*S. cerevisiae* STE12 ortholog) binding to specific DNA sequences near mating genes termed pheromone response elements (PRE) (Müller et al. 2003). The action of PRF1 up-regulates transcription of pheromones, receptors and b-type HDPs (Moore et al. 2011). PRF1 is not activated unless cAMP is readily available and a pheromone from a compatible partner binds to the STE3-type GPCR (Bölker et al. 2013).

U. maydis is a tetrapolar fungus and contains two unlinked loci determining mating compatibility. Fungi with a single locus determining compatibility are termed bipolar. Tetrapolar versus bipolar compatibility systems determine the number of matings that will be successful between the progeny of a given meiotic event (Halisky 1965). Because fewer matings will be successful for tetrapolar species, outcrossing is more likely than for bipolar species (Halisky 1965). By comparing the tetrapolar *U. maydis* with the bipolar *Ustilago hordei*, Bakkeren and Kronstad (1993) showed that bipolar and tetrapolar smut fungi encode orthologous mating-type

genes at their MAT-loci and the linkage status of the MAT-a and MAT-b genes is what distinguishes these two systems.

While the molecular basis of mating in *Ustilago* species has been studied in great detail, the same has not been done for the bunt fungi (*Tilletia* species). A large body of work on mating systems in Agaricomycotina (reviewed in Raudaskoski and Kothe 2010; James et al. 2013) and Pucciniomycotina (Coelho et al. 2011; Giraud et al. 2008; Petit et al. 2012; Toome et al. 2013) has shown conservation of form and function of mating type proteins with only limited conservation of copy number of MAT-genes and synteny at mating loci. Studies utilizing crosses indicate bipolar heterothallism with multiple alleles for a number of *Tilletia* species including *T. indica* (Durán and Cromarty 1977) and *T. contraversa* (Hoffmann and Kendrick 1965; 1969) and simple bipolar heterothallism for the common bunts *T. caries* and *T. laevis* (Flor 1932; Holton and Kendrick 1957).

Because of the similarities in mating behavior identified by these authors, it is expected that similar conservation of mating gene form and function will be present in *Tilletia* species. This study takes a comparative genomics approach, mining available *Tilletia* genome sequences for the mating genes and loci. By including two strains of *T. caries* as well as three other *Tilletia* species, molecular variation at *Tilletia* mating loci can be investigated for the first time.

METHODS

Strains and Culture Conditions

Tilletia caries strain 517 (WSP72095) was collected at a Washington State University wheat research plot in Whitman Co., WA, USA in July of 2006 by Xianming Chen (USDA-

ARS, Pullman, WA) (Table S1). A haploid monosporidial line was established in culture using the methods described in Boyd and Carris (1997). Secondary sporidia and mycelial fragments scraped from cultures grown on M19, a semi-defined medium (Trione 1969), were suspended in 200 µL sterile distilled de-ionized water and streak-plated on to 100 x 15 mm Petri dishes of M19 amended with 40 U/mL penicillin and 40 µg/mL streptomycin (Life Technologies, Grand Island, NY) overlain with sterile SpectraPor dialysis membrane (VWR Scientific, Radnor, PA). Plates were incubated in the dark at 15C for 2 – 4 weeks in unsealed plastic culture boxes that allowed air exchange.

Nucleic Acids Extraction

Mycelium for DNA extraction was removed from dialysis membranes with a sterile metal spatula, placed in a 1.5 mL microfuge tube, frozen at -80 C for 24 hours, and lyophilized overnight. Lyophilized mycelium was ground using a sterile mortar and pestle and stored at -20 C prior to use. DNA was extracted from approximately 50 mg of ground, lyophilized mycelium using a modified phenol-chloroform extraction procedure (Lee and Taylor 1990). Mycelium for RNA extraction was removed from membranes as previously described, frozen in liquid nitrogen, and ground using a sterile mortar and pestle. RNA was extracted using TRIzol reagent (Thermo Fisher Scientific, Waltham, MA) per the manufacturer's instructions.

Whole Genome Sequencing

Tilletia caries monokaryotic strain 517, isolated as described above, was sequenced in the Washington State University Laboratory for Biotechnology and Bioanalysis on PacBio RS and Roche 454 sequencing platforms. A hybrid assembly of both datasets was carried out by Dr. Jodi

Humann in the WSU Horticulture Dept. utilizing Allora (SMRT analysis: http://www.pacificbiosciences.com/products/software/algorithms/) to create corrected continuous long reads. Genome assembly statistics are available in Table S2. One strain each of *Tilletia caries, T. contraversa, T. indica and T. walkeri*, was selected for whole genome and RNA sequencing on an Illumina HiSeq by Agriculture and Agri-Food Canada scientist Sarah Hambleton [Eastern Cereal and Oil Seed Research Center, Ottawa, ON; Canadian Safety and Security Program (CSSP), project number CRTI 09-462RD)] and assembled with the Tuxedo suite (Trapnell et al. 2009; 2010; 2013; Kim et al. 2011; Roberts et al. 2011; 2011). The sources of all strains included in this study are listed in Table S1.

cDNA Preparation

RNA was treated with RQ1 RNase-Free DNase (Promega Corporation, Madison, WI) per the manufacturer's instructions. cDNA was prepared from DNase-treated RNA using cloned AMV Reverse Transcriptase (Thermo Fisher Scientific, Waltham, MA) and OligoDT_{20} primers (Thermo Fisher Scientific, Waltham, MA) per the manufacturer's instructions.

Polymerase Chain Reaction and Sanger Sequencing

PCR amplification was used to verify the sequence and transcription of the mating genes in *T. caries* 517 and bridge adjacent contigs at the mating subloci. Primers were designed manually or with the aid of Primer3 v0.4.0 (Rozen and Skaletsky 2000) and are listed in Table S4. Reactions utilized KOD Xtreme Hot-Start DNA polymerase (Merck KGaA, Darmstadt, Germany) per the manufacturer's instructions. Thermal cycling was carried out on a MyCycler thermal cycler (BioRad, Hercules, CA) using the following parameters: one cycle each of 75 C

for fifteen minutes and 94 C for two minutes followed by forty cycles of 98 C for ten seconds, the annealing temperature (selected as described below) for thirty seconds, 68 C for 2+ minutes (see below) and one cycle at a final extension temperature of 68 C for ten minutes. Samples were held at 4 C following completion of this protocol. Temperature gradients were used to identify the optimum annealing temperature for each target amplicon. Extension time was modified depending on the size of the desired amplicon, with an additional minute added to extension per kb over 1kb. Amplification was verified via gel-electrophoresis in 1% agarose gels. DNA concentration and amplicon-length were visually assayed in the same gel on a UV transilluminator using GeneRuler DNA Ladder Mix (ThermoFisher Scientific, Waltham, MA). PCR products were purified with ExoSAP-IT (USB, Cleveland, OH) according to the manufacturer's instructions, and forward and reverse strands were sequenced by Elim Biopharmaceuticals (Hayward, CA) and assembled in a pairwise manner in Geneious 7.1.7 (Biomatters Ltd.) using the de novo assemble option.

Genome Annotation and Protein Family Clustering

Genome annotation was carried out for *T. caries*, *T. caries* 517, *T. contraversa*, *T. indica* and *T. walkeri* utilizing WebAUGUSTUS with each organism's transcript and genome assemblies as input sequences for parameter training (Hoff and Stanke 2013). Output is included in AUGUSTUS_SUPP.zip and summarized in Table S3. Conceptual translations of annotations were pooled and "all against all" pairwise alignments (using the annotations as both query sequences and the subject database) were carried out with the BLASTp algorithm (Altschul et al. 1990; Altschul et al. 1997) using an Expectation-value (E-value) cutoff of 1. BLAST results were clustered by E-value into protein families utilizing MCL Edge (Enright et al. 2002; van

Dongen et al. 2000) with the inflation parameter set to 6.0. Gene clusters are listed in MCL_SUPP.txt. Short ORFs in the genome assembly (90-200 nt) were identified with Emboss GetORF (http://emboss.open-bio.org/rel/dev/apps/getorf.html) and conceptually translated into proteins in Geneious 7.1.7 (Biomatters Ltd.). These were clustered with MCL Edge as described previously. A variety of Perl scripts were used to process these annotations and gene clusters. These are included in the Perl Scripts Supplement along with author and source information.

Annotation of the Mating Loci

Sequences associated with mating loci in *U. maydis* (GenBank accessions XP_758530, XP_756724 and, XP_756724) were downloaded from GenBank. tBLASTn was used to perform pairwise alignments with all available databases of genomic contigs, transcript assemblies and AUGUSTUS gene predictions with an E-value cutoff of $1e^{-6}$. MCL clusters containing AUGUSTUS prediction matches were investigated for the presence of additional mating orthologs. Proteins were only designated as b-mating-type if they demonstrated E-values of less than $1e^{-6}$ against homeodomain-containing mating-type proteins found in other basidiomycetes in the NCBI RefSeq Basidiomycota database. AUGUSTUS-predicted mating gene coding sequences were mapped to their respective contigs or scaffolds, which were then re-annotated with FGENESH (Solovyev et al. 2006) under the "*Ustilago*" parameter settings. Putative mating-type-determining HDP-encoding gene predictions were not considered credible unless supported by transcript evidence. Transcripts at putative mating loci were identified using BLASTn (E=0). All coding sequence annotations identified in this manner were mapped to their respective genomic contigs with the Geneious 7.1.7 (Biomatters Ltd.) map to reference command. FGENESH or AUGUSTUS annotations were compared to the NCBI Conserved Domain

Database (Marchler-Bauer et al. 2011) and BLASTx was run against NCBI RefSeq Fungi to assign putative function. Genes exhibiting homology to basidiomycete mating type proteins were aligned with BLASTn hits from their respective transcriptomes using Clustal Omega (Sievers et al. 2011) to verify the accuracy of gene predictions and were manually modified in Geneious 7.1.7 (Biomatters Ltd.) to reflect their transcripts in cases where transcript evidence did not match gene predictions. All annotations are provided in FASTA format as a supplement.

Phylogenetic Analyses

All sets of protein orthologs of interest were aligned using Clustal Omega (Sievers et al. 2011; McWilliam et al. 2013). ProtTest 3.2 (Darriba et al. 2011; Guindon and Gascuel 2003) was run on the HDP alignment to select the best models of protein substitution for maximum likelihood and Bayesian Markov Chain Monte Carlo analyses. The JTT+G+F model of amino acid substitution was selected using ProtTest 3.4 (Darriba et al. 2011; Guindon and Gascuel 2003) and employed in the maximum likelihood phylogenetic analysis. JTT+G was employed in MrBayes 3.2.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). A maximum likelihood analysis with 100 bootstrap replicates was carried out in PhyML 3.0 (Guindon and Gascuel 2003) and Bayesian analyses were carried out in MrBayes 3.2.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003).

RESULTS

Homeodomain Proteins

Putative HDP encoding genes identified in the *Tilletia* genomes are listed in Table 2-1. All except Tco_bE_2 encode a traditional WFxNxR homeodomain motif (Scott et al. 1989).

Only two potential homeotic genes out of twenty seven identified could not be resolved from the sequence data (Table 2-1). Transcript contig names, genomic contig names and corresponding AUGUSTUS predictions are listed in Table 2-1. In addition to bE and bW clusters identified, two additional groups of HDPs were identified. These demonstrated a higher degree of conservation than the putative mating-type-determining HDPs. No putative functions could be assigned to these genes via similarity searches in GenBank, however they appear to be conserved across the fungal kingdom (Table 2-1). The phylogenetic relationships between the bW type, bE type and non-b-type HDPs are shown in Figure 2-1. Fig. 2-1A depicts the Bayesian Markov Chain Monte Carlo analysis and Figure 2-1B the maximum likelihood analysis. The same major clades are present in both analyses, and the only difference between the two trees is branching order of the bE cluster (Fig. 2-1A, B). The non-mating-type HDPs form two clades corresponding to their MCL clusters which may be found in MCL_SUPP.txt.

Table 2-1. Homeodomain-containing-proteins encoded in the genomes of *T. caries* (Tca), *T. caries* 517 (Tc517), *T contraversa* (Tco), *T. indica* (Tin) and *T. walkeri* (Twa). Orange and blue labels represent putatively non-mating-type determining groups of orthologs. Red and teal labels represent putatively mating-type determining homeodomain proteins. Motifs highlighted in red represent departures from the classic WFxNxR motif (Scott et al. 1989).

Homeodomain protein	Motif	Length (CDS/AA)	Transcript(s)	Gene (Augusutus)	Contig	BLASTp Hit Description NCBI RefSeq Ustilaginomycotina	Accession No.
Tc517_g1873	WFINAR	1941/646	none	Tc517_g1873	73415_contig	hypothetical protein *Pseudozyma flocculosa*	XP_007881851.1
Tca_g10368	WFINAR	1941/646	Tca_DAOM_238032_R01V2.3693.2; 3694.1; 3693.1	Tca_g10368	Tca_DAOM_238032_R01_4013	hypothetical protein *P. flocculosa*	XP_007881851.1
Tco_g8961	WFINAR	1941/647	Tc_DAOM_236426_R01V2.4220.1; 4219.1	Tco_g8961	Tc_DAOM_236426_R01_472	hypothetical protein *P. flocculosa*	XP_007881851.2
Twa_g197	WFINAR	1947/648	Tw_DAOM_236422_R01V2.2124.1	Twa_g197	Tw_DAOM_236422_R01_1433	hypothetical protein *Cyphellophora europaea*	XP_008718427.1
Tin_g1181	WFINAR	1938/649	Ti_DAOM_236416_R01V2.3483.1	Tin_g1181	Ti_DAOM_236416_R01_3724	hypothetical protein *P. flocculosa*	XP_007881851.1
Tc517_g394	WFQNRR	1563/518	none	Tc517_g394	32530_contig	hypothetical protein *Ustilago maydis*	XP_762430.1
Tca_g5325	WFQNRR	1563/519	Tca_DAOM_238032_R01_99	Tca_g5325	Tca_DAOM_238032_R01_99	hypothetical protein *U. maydis*	XP_762430.1
Tco_g4688	WFQNRR	1563/520	Tc_DAOM_236426_R01V2.1083.1	Tco_g4688	Tc_DAOM_236426_R01_1821	hypothetical protein *U. maydis*	XP_762430.2
Twa_g3288	WFQNRR	1704/567	Tw_DAOM_236422_R01V2.1802.1	Twa_g3288	Tw_DAOM_236422_R01_1394	hypothetical protein *U. maydis*	XP_762430.3
Tin_g6649	WFQNRR	1704/567	Ti_DAOM_236416_R01V2.7207.1	Tin_g6649	Ti_DAOM_236416_R01_631	hypothetical protein *U. maydis*	XP_762430.4
Tc517_bE	WFINMR	1986/661	Sanger trace evidence from cDNA; see cDNA_SUPP.fasta	Tc517_g2987	73754_contig	mating-type locus allele B1 protein *U. maydis*	XP_756724.1
Tco_bE_1	WFINMR	1875/624	Tc_DAOM_236426_R01V2.1433.1	Tco_g8544	Tc_DAOM_236426_R01_21	mating-type locus allele B1 protein *U. maydis*	XP_756724.2
Tco_bE_2	WTNILR	1890/629	Tc_DAOM_236426_R01V2.2095.1	Tco_g4281	Tc_DAOM_236426_R01_2669	mating-type locus allele B1 protein *U. maydis*	XP_756724.3
Tca_bE_2	WFGNMR	1206/401	Tca_DAOM_238032_R01V2.6610.2; 6610.1 (may be incomplete)	Tca_g4472	Tca_DAOM_238032_R01_665	mating-type locus allele B1 protein *U. maydis*	XP_756724.4
Tca_bE_1	WFINMR	1989/662	Tca_DAOM_238032_R01V2.366.1	Tca_g4235	Tca_DAOM_238032_R01_1227	mating-type locus allele B1 protein *U. maydis*	XP_756724.5
Tin_bE	WFINMR	2,112/703	Ti_DAOM_236416_R01V2.3128.1	Tin_g5612	Ti_DAOM_236416_R02_3367	mating-type locus allele B1 protein *U. maydis*	XP_756724.6
Twa_bE	WFINMR	2,100/699	Tw_DAOM_236422_R01V2.6028.1	Twa_g988	Tw_DAOM_236422_R01_699	mating-type locus allele B1 protein *U. maydis*	XP_756724.2
Tc517_bW_1.1	WFQNRR	1677/558	Sanger trace evidence from cDNA; see cDNA_SUPP.fasta	Tc517_g2986	73754_contig	putative homeodomain transcription factor bw2 *P. flocculosa*	XP_007878973.1
Tc517_bW_1.2	WFCNRR	1110/369	Sanger trace evidence from cDNA; see cDNA_SUPP.fasta	Tc517_g2988	73754_contig	similar to B42094 bw1 *U. maydis*	XP_756725.1
Tca_bW1	WFCNRR	1002/333	Tca_DAOM_238032_R01V2.7861.1; Tca_ CUFF.10731	Tca_g8104	Tca_DAOM_238032_R01_7477	putative homeodomain transcription factor bw2 *P. flocculosa*	XP_007878973.1
Tca_bW2	WFQNRR	1782/593	Tca_CUFF.10610	Tca_g4234	Tca_DAOM_238032_R01_1227	similar to B42094 bw1 *U. maydis*	XP_756725.1
Tco_bW_2	WFCNRR	1212/445	Tc_DAOM_236426_R01V2.3898.1	Tco_g9842	Tc_DAOM_236426_R01_4416	putative homeodomain transcription factor bw2 *P. flocculosa*	XP_007878973.1
Tco_bW_1	WFQNRR	1782/593	Tc_DAOM_236426_R01V2.8690.1	Tco_g7249	Tc_DAOM_236426_R01_887	similar to B42094 bw1 *U. maydis*	XP_756725.1
Tin_bW	WFQNRR	1440/479	Ti_DAOM_236416_R01V2.3086.1	Tin_g3384	Ti_DAOM_236416_R01_345	similar to B42094 bw1 *U. maydis*	XP_756725.2
Twa_bW	WFQNRR	1440/479	Tw_DAOM_236422_R01V2.6024.1; 6026.1	Twa_g989	Tw_DAOM_236422_R01_699	similar to B42094 bw1 *U. maydis*	XP_756725.3
Tc517_g4063	WFYRTR	unable to resolve	Runs off end of contig; no transcripts available	Tc517_g4063	74166_contig	N/A	N/A
Tco_g5544	WFYRTR	unable to resolve	ORF is abnormally short and lacks transcript support	Tco_g5544	Tc_DAOM_236426_R01_7140	N/A	N/A

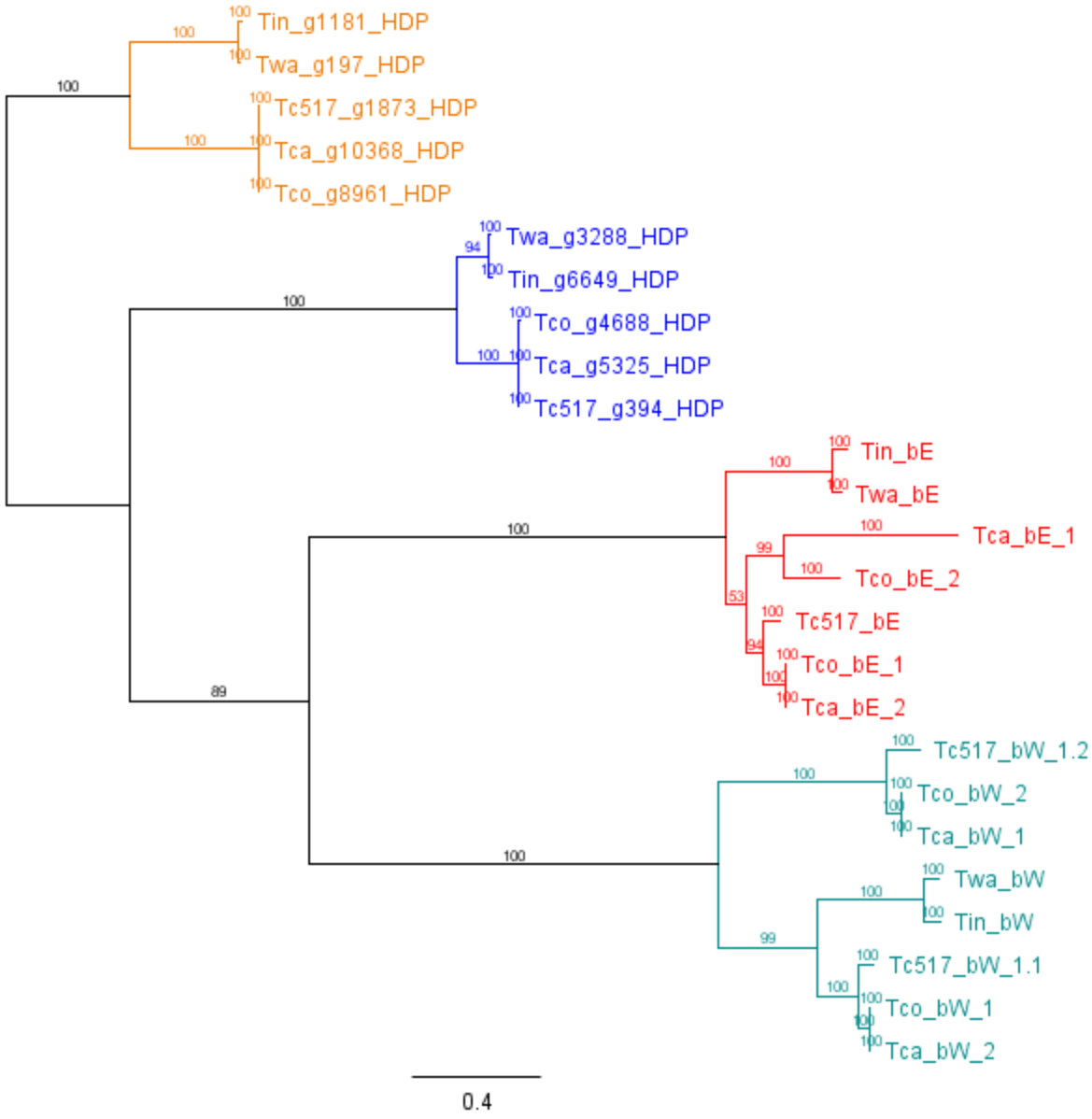

Figure 2-1 A. A Bayesian Markov Chain Monte Carlo phylogenetic analysis carried out in MrBayes 3.2.2 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Numbers at nodes represent posterior probabilities; distances are in substitutions per site. Clades represent putative bE and bW MAT-b homeodomain proteins as well as two groups of highly conserved, putative non-MAT HDPs. Colors of boxes around clades correspond to those in Table 2-1.

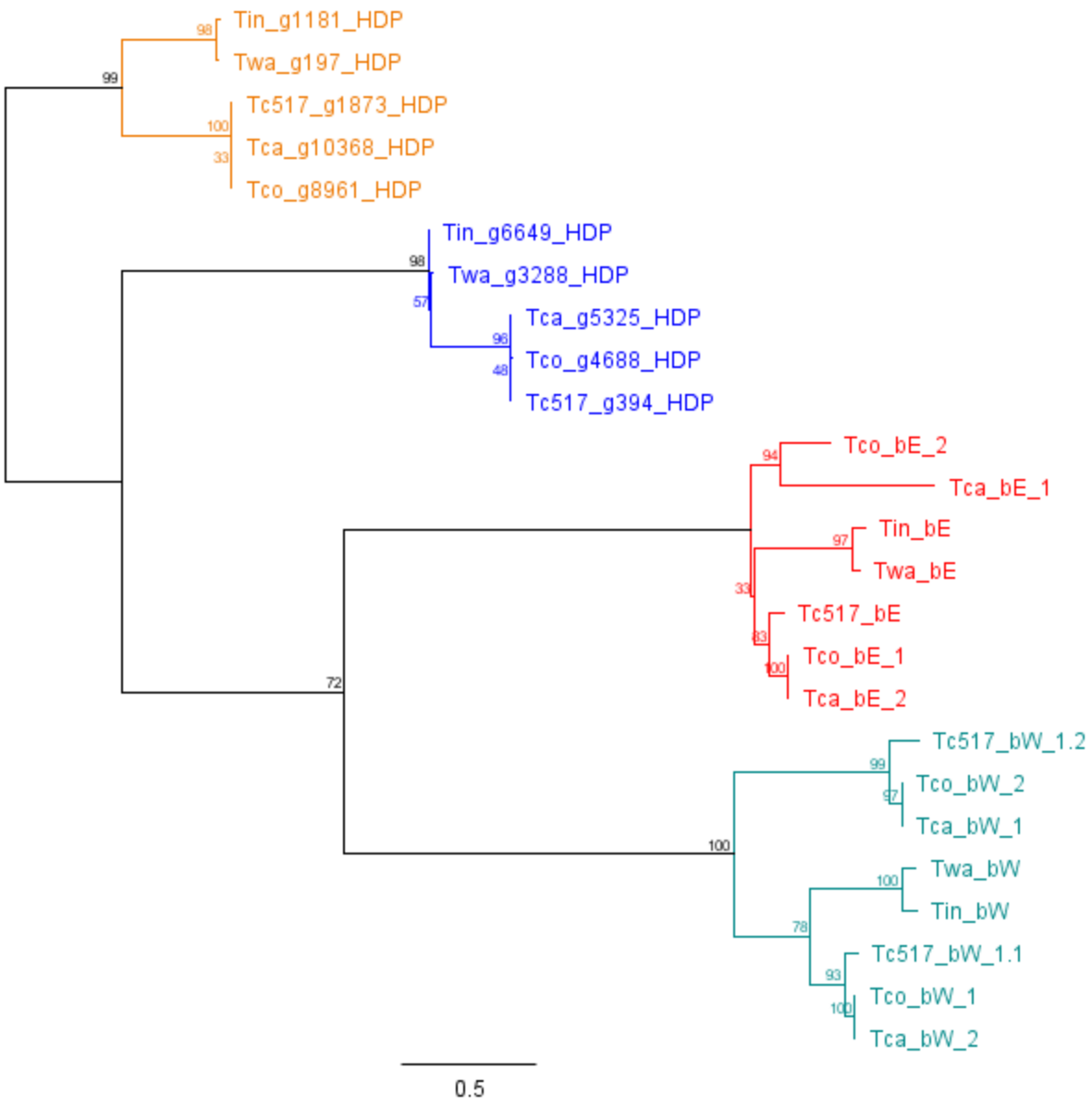

Figure 2-1 B. A 100 bootstrap maximum likelihood phylogenetic analysis. Numbers at nodes represent bootstrap support values; distances are in substitutions per site. Clades represent putative bE and bW MAT-b homeodomain proteins as well as two groups of highly conserved, putative non-MAT HDPs. Colors of labels and clades correspond to those in Table 2-1.

Putative b-East Mating Type HDPs

An alignment of all bE orthologs is depicted in Figure 2-2 and pairwise similarity scores from the alignment are listed in Table 2-2. Two bE orthologs were identified for both *T. caries* (Tca_bE_1; Tca_bE_2) and *T. contraversa* (Tco_bE_1; Tco_bE_2), but not for *T. caries* 517, *T. indica* \and *T. walkeri*. The HDPs demonstrate the lowest degree of conservation in the N-termini, which is also where the homeodomain-motifs are found, as indicated in the red box in Figure 2-3. Tca_bE_2 is 40 residues shorter than the other peptides. Colocation at a MAT-b HDP-encoding locus, characterized by divergently transcribed adjacent HDPs (Bakkeren et al. 2008) could only be verified for Tc517_bE, Tca_bE_1, Tin_bE and, Twa_bE due to incompleteness of the genome assemblies. Other HDPs were assigned as mating-type or non-mating-type HDPs by similarity ($E<1e^{-3}$) to known fungal mating-type HDPs identified using BLASTp against the NCBI RefSeq Fungi database.

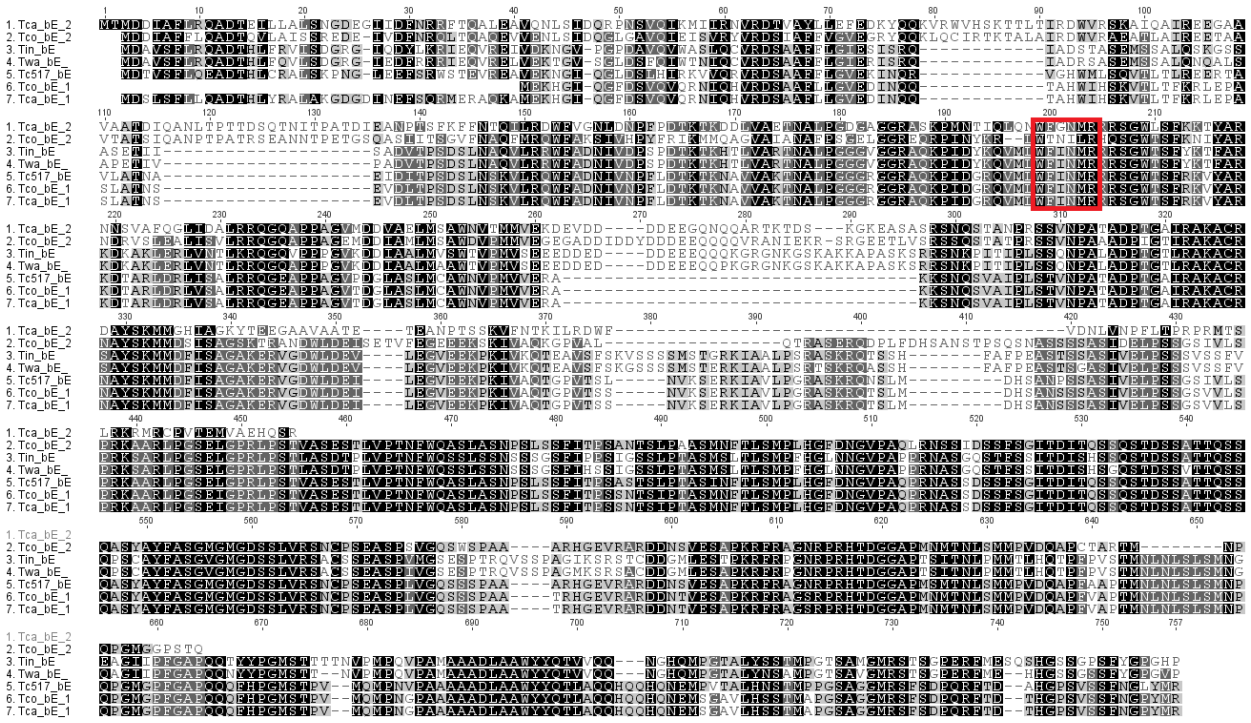

Figure 2-2.Alignment of bE orthologs. Alignment is shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. The homeodomain motif is boxed in red.

Table 2-2. A pairwise similarity matrix depicting percent identity between the bE type polypeptides aligned in Fig. 2-2.

	Tca_bE_2	**Tco_bE_2**	**Tin_bE**	**Twa_bE**	**Tc517_bE**	**Tco_bE_1**	**Tca_bE_1**
Tca_bE_2		43.349	33.79	34.703	35.28	34.872	35.748
Tco_bE_2	43.349		46.596	46.899	58.218	58.143	56.748
Tin_bE	33.79	46.596		91.75	58.924	59.342	58.557
Twa_bE_	34.703	46.899	91.75		60.597	60.06	59.375
Tc517_bE	35.28	58.218	58.924	60.597		91.346	88.973
Tco_bE_1	34.872	58.143	59.342	60.06	91.346		100
Tca_bE_1	35.748	56.748	58.557	59.375	88.973	100	

Based on this analysis, Tca_bE_1 and Tco_bE_1 are identical except for the first 40 residues, as shown in Figure 2-2 and Table 2-2, and may represent the same bE ideomorph. Tc517_bE is 91% identical to Tco_bE_1 and approximately 89% identical to Tca_bE_1, and may also represent the same ideomorph. The Twa_bE and Tin_bE HDPs have greater than 90% identity, suggesting that they could represent an orthologous ideomorph. Tca_bE_2 and Tco_bE_2 have < 43% and 58% identity, respectively, with any other protein depicted.

Putative b-West Mating-Type HDPs

An alignment of all bW orthologs is depicted in Figure 2-3 and pairwise similarity scores are listed in Table 2-3. Two bW-type peptides are identified for *T. caries* 517, *T. caries* and *T. contraversa*. No two orthologs from the same genome share >32% identity along their shared lengths. Subalignments of peptides demonstrating >78% identity along their shared lengths are depicted in Figures 2-4, 2-5 and, 2-6. The pairs of Tco_bW_1 and Tca_bW_2, as well as Tco_bW2 and Tca_bW_1, have 100% identity. Differences between these peptides and the *T. caries* 517 peptides may represent artefacts of assembly or different ideomorphs within the population. The different homeodomain motifs may indicate that the two respective groups shown in figures 2-4 and 2-5 could be ideomorphs, though colocation at a MAT-b HDP-encoding locus (characterized by divergently transcribed adjacent HDPs) could only be verified for Tc517_bW_1.1, Tc517_bW_1.2, Tca_bW_2, Tin_bW and Twa_bW due to incompleteness of the genome assemblies. Other genes were assigned as mating-type HDPs by similarity determined using BLASTp against the NCBI RefSeq Ustilaginomycotina database. Most variability is located towards the C-termini of the peptide group containing Tc517_bW_1.2, Tco_bW_2 and, Tca_bW_1 depicted in Figure 2-3. An indel is present from residues 111-132

and is lacking in Tc517_bW_1.2 (Figure 4). Conversely, the group of Tco_bW_1, Tca_bW_2 and Tc517_bW_1.2 have greater variability at their N-termini and length polymorphisms are not evident (Figure 2-4). The two subgroups of HDPs shown in Figures 2-4 and 2-5 are differentiated by homeodomain motifs as well as within-group homology. The subalignment of putative bW-type peptides (Figure 2-4) includes a WFCNRR homeodomain motif, while subalignment (Figure 2-5) includes a WFQNRR homeodomain motif. The subalignment of *T. indica* and *T. walkeri* peptides Tin_bW and Twa_bW (Figure 2-6), demonstrate > 91% identity. The HDPs for *T. indica* and *T. walkeri* have greater similarity to each other than to *T. caries* and *T. contraversa*. For both bE and bW type HDPs, there is 23-60% identity between the *T. indica* and *T. walkeri*-group proteins and the *T. caries* and *T. contraversa*-group proteins, and approximately 91% identity to each other (Table 2-3).

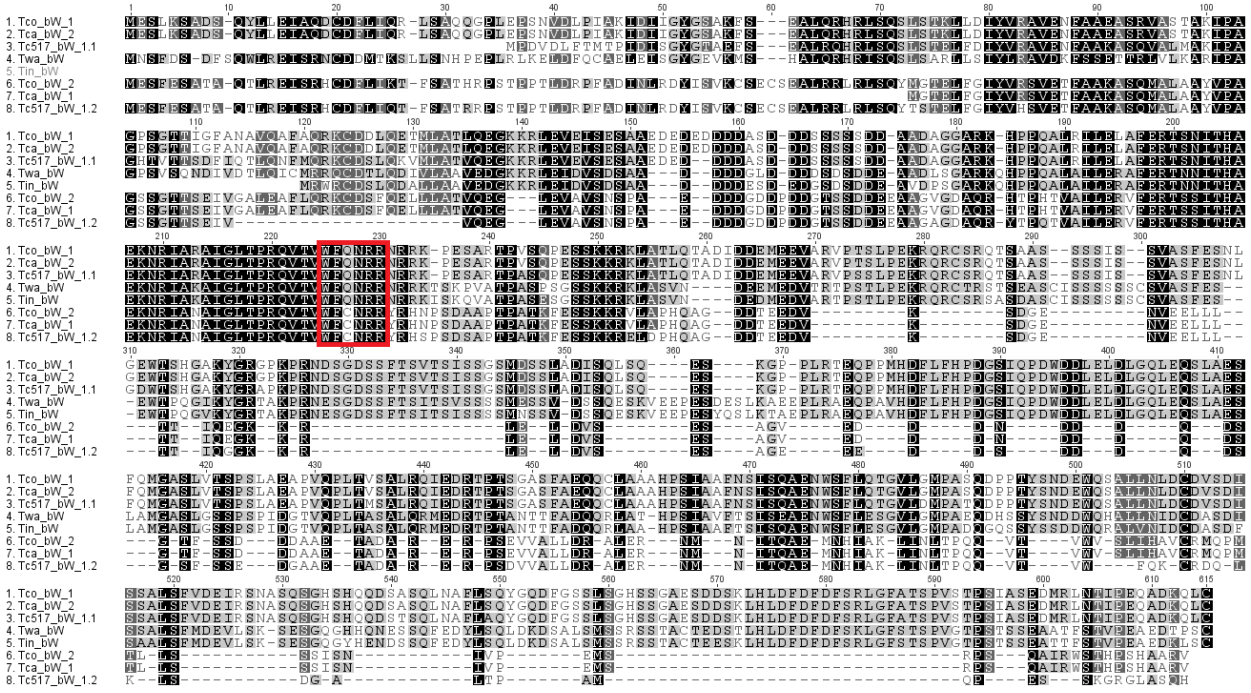

Figure 2-3. Alignment of all bW orthologs. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. Homeodomain motif is boxed in red.

Table 2-3. A pairwise similarity matrix depicting percent identity between the bW type polypeptides aligned in Fig. 2-3.

	Tc517_bW_1.2	Tco_bW_2	Tca_bW_1	Tco_bW_1	Tca_bW_2	Tc517_bW_1.1	Twa_bW	Tin_bW
Tc517_bW_1.2		81.795	78.438	30.932	30.932	28.929	27.176	22.946
Tco_bW_2	81.795		100	32.99	32.99	31.195	29.132	25.41
Tca_bW_1	78.438	100		30.361	30.361	31.084	27.119	25.41
Tco_bW_1	30.932	32.99	30.361		100	91.25	57.781	60.327
Tca_bW_2	30.932	32.99	30.361	100		91.25	57.781	60.327
Tc517_bW_1.1	28.929	31.195	31.084	91.25	91.25		58.421	60.532
Twa_bW	27.176	29.132	27.119	57.781	57.781	58.421		90.397
Tin_bW	22.946	25.41	25.41	60.327	60.327	60.532	90.397	

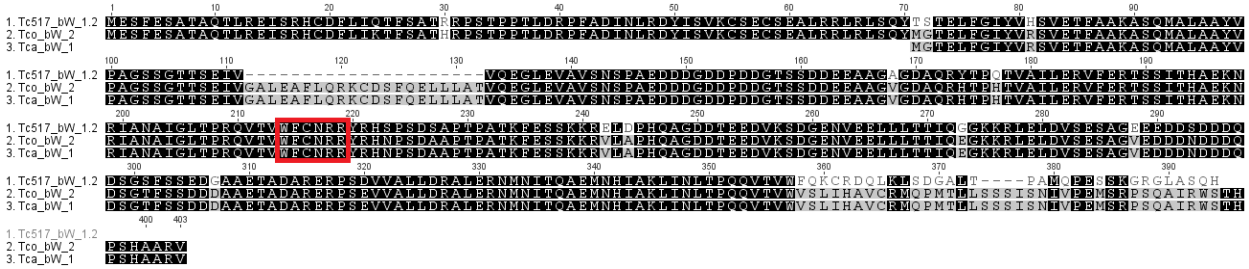

Figure 2-4. Alignment of Tc517_bW_1.2, Tco_bW_2 and, Tca_bW1, a subset of bW polypeptides. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. Homeodomain motif is boxed in red.

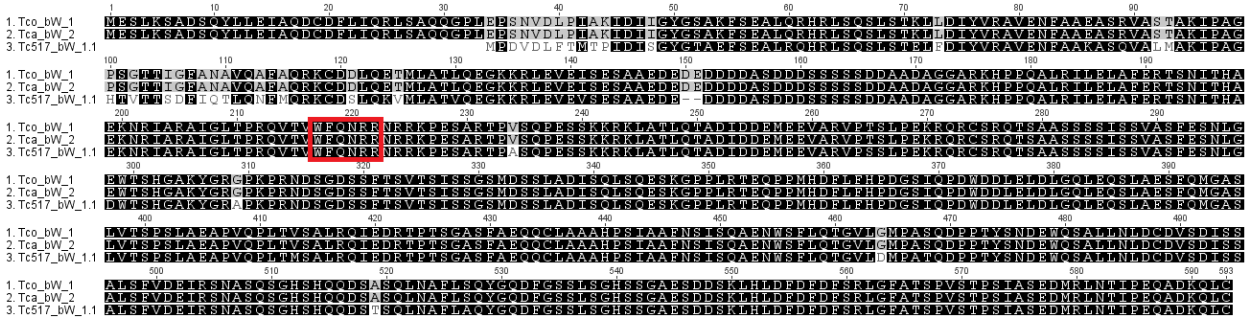

Figure 2-5. Alignment of Tco_bW_1, Tca_bW_2 and, Tc517_bW_1.1, a subset of bW polypeptides. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. Homeodomain motif is boxed in red.

Figure 2-6. Alignment of Tin_bW and Twa_bW, a subset of bW polypeptides. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. Homeodomain motif is boxed in red.

Other HDPs

Two additional clusters of orthologous HDPs were identified in the genomes analyzed and were grouped together based on their BLAST scores during the MCL-clustering analysis (MCL_SUPP.txt). Each cluster of proteins shares the same homeodomain motifs (WFQNRR and WFINAR, respectively). Alignments and similarity matrices for each cluster are shown in Figures 2-7 and 2-8 and Tables 2-4 and 2-5 below. A greater amount of conservation was present within each group relative to the putative b-mating-type HDPs. None of these HDP-encoding genes were identified as divergently transcribed and are unlikely to be involved in mating as it is currently understood. Sequence conservation is far greater within the groups comprising the systemically infecting bunts, *T. caries* and *T. contraversa* (Figure 2-7 Tco_g4688_HDP, Tca_g5325_HDP, Tc517_g394_HDP; Figure 8. Tc517_g1873_HDP, Tca_10368_HDP, Tco_g8961_HDP) and the non-systemically infecting bunts, *T. indica* and *T. walkeri* (Figure 2-7 Tin_g1181_HDP, Twa_g197_HDP; Figure 8. Twa_g3288_HDP, Tin_g6649_HDP) than between the two groups. The species group in Fig. 2-7 is better conserved (71%) than the species group shown in Fig. 2-8 (41%). The HDPs are similar only to hypothetical proteins in NCBI RefSeq Fungi database.

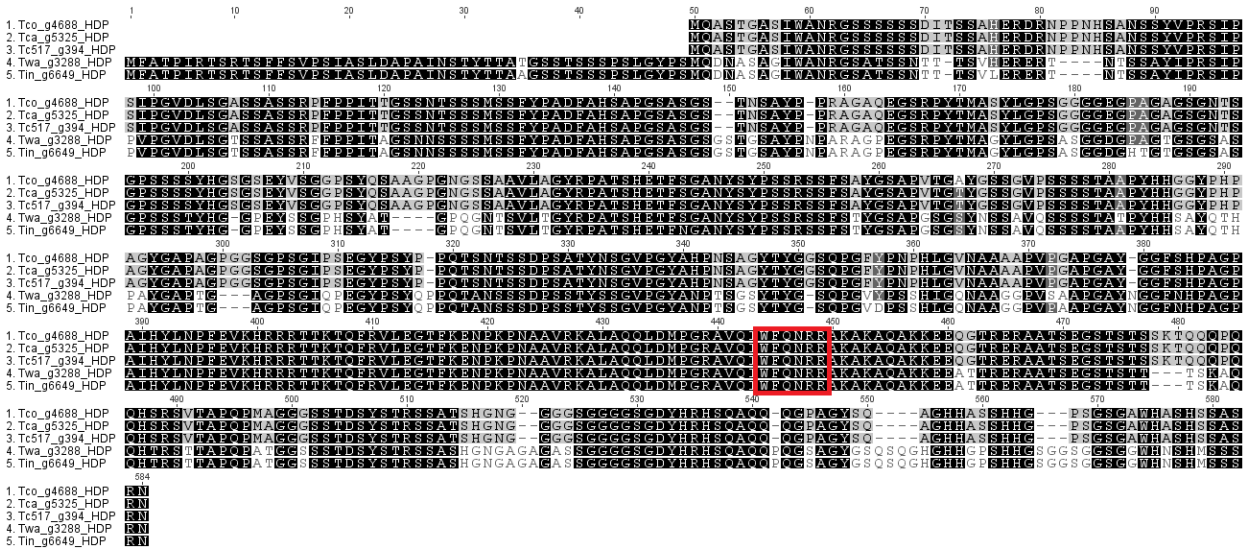

Figure 2-7. Alignment of a cluster of orthologous homeodomain proteins from sequenced *Tilletia* genomes. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. Homeodomain motif is boxed in red.

Table 2-4. A pairwise similarity matrix depicting percent identity between the HDP polypeptides aligned in Fig. 2-7.

	Tco_g4688_HDP	**Tca_g5325_HDP**	**Tc517_g394_HDP**	**Twa_g3288_HDP**	**Tin_g6649_HDP**
Tco_g4688_HDP		99.808	99.808	71.028	70.654
Tca_g5325_HDP	99.808		100	71.028	70.654
Tc517_g394_HDP	99.808	100		71.028	70.654
Twa_g3288_HDP	71.028	71.028	71.028		98.413
Tin_g6649_HDP	70.654	70.654	70.654	98.413	

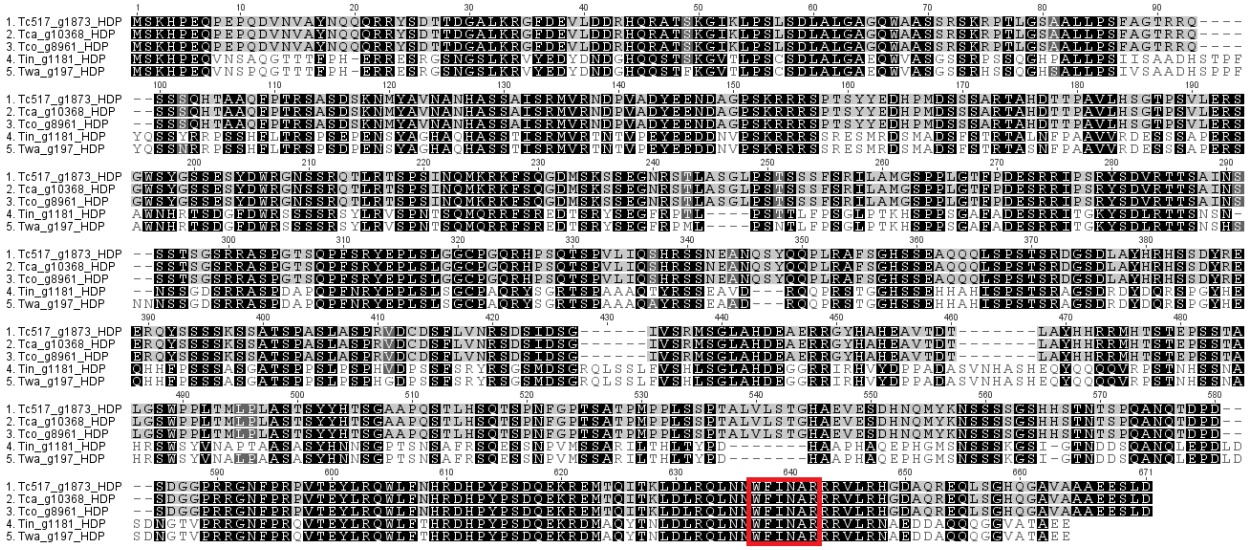

Figure 2-8. Alignment of a cluster of orthologous homeodomain proteins from sequenced *Tilletia* genomes. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. Homeodomain motif is boxed in red.

Table 2-5. A pairwise similarity matrix depicting percent identity between the HDP type polypeptides aligned in Fig. 2-8.

	Tc517_g1873_HDP	**Tca_g10368_HDP**	**Tco_g8961_HDP**	**Tin_g1181_HDP**	**Twa_g197_HDP**
Tc517_g1873_HDP		100	100	47.583	47.44
Tca_g10368_HDP	100		100	47.583	47.44
Tco_g8961_HDP	100	100		47.583	47.44
Tin_g1181_HDP	47.583	47.583	47.583		96.914
Twa_g197_HDP	47.44	47.44	47.44	96.914	

*Putative GPCR (***STE3** *Orthologs) and Pheromone Precursors*

At MAT-a, a putative pheromone receptor (*S. cerevisiae* STE3 ortholog) and pheromone precursor (MFA) were identified in each genome investigated. Summaries of STE3 genes are shown in Table 2-6. and MFA genes in Table 2-7.

Table 2-6. STE3 orthologs identified in the five genomes investigated.

STE3 (GPCR Ortholog)	Length (CDS/AA)	Transcript(s)	Exons	Gene (Augustus)	Contig
Tca_STE3	1092/363	Tca_DAOM_238032_R01V2.2711.1	5	Tca_g9885	Tca_DAOM_238032_R01_3120
Tc517_STE3	1086/361	Sanger trace evidence from cDNA; see cDNA_SUPP.zip	5	Tc517_g3435	73883_contig
Tco_STE3	1086/361	Tc_DAOM_236426_R01V2.7688.1; 7688.2	5	Tco_g7596	Tc_DAOM_236426_R01_818
Twa_STE3	1086/361	Tw_DAOM_236422_R01V2.1035.1	4	Twa_g7605	Tw_DAOM_236422_R01_1251
Tin_STE3	1086/361	Ti_DAOM_236416_R01V2.2788.1	4	Tin_g7227	Ti_DAOM_236416_R01_3250

Table 2-7. MFA orthologs identified in the five genomes investigated.

MFA Ortholog	Length CDS/AA	Transcript	Transc. Length	Contig
Tc517_mfa	132/44	Sanger trace evidence from cDNA; see cDNA_SUPP.zip	partial	73883_contig
Tca_mfa	132/44	Tca_DAOM_238032_R01V2.2838.1	593	Tca_DAOM_238032_R01_3120
Tco_mfa	135/45	Tc_DAOM_236426_R01V2.7690.1	614	Tc_DAOM_236426_R01_818
Twa_mfa	111/37	none	N/A	Tw_DAOM_236422_R01_1251
Tin_mfa	114/38	Ti_DAOM_236416_R01V2.2790.1	616	Ti_DAOM_236416_R01_3250

STE-3 GPCRs had greater homology than the MAT-b HDPs (Table 2-8 versus Table 2-3 and Table 2-2). In STE3 orthologs, the transmembrane domain prediction software HMMTOP (Tusnády and Simon 1998; Tusnády and Simon 2001) identified all seven conserved transmembrane motifs characteristic of these proteins (see HMMTOP_SUPP.txt) (Schirawski et al. 2005), shown in Figure 2-9 (alignment of predicted STE3 peptides) and Table 2-8. (similarity scores). Short open reading frames of 111 to 132 nt with C-terminal farnesylation motifs characteristic of fungal mating pheromone precursors were associated with each pheromone receptor. Transcripts were identified for all pheromone precursor ORFs with the exception of *T. walkeri* (Twa_MFA). Amino acid alignments and similarity matrices for pheromones and

GPCRs are shown in Figure 2-9, Table 2-10, Figure 2-10 and, Table 2-11 respectively. GPCR STE3 and MFA orthologs have the highest level of conservation at their N-termini, as has been previously described for *U. maydis* (Bölker et al. 1992; Spellig et al. 1994). Each STE3 ortholog has a distinct sequence (Fig. 2-9), but as with the HDPs, two groups can be identified based on higher within-group percent identities (Table 2-8). These subgroups correspond to the systemically infecting species (dwarf and common bunts) and the locally infecting species (*T. indica T. walkeri*). For MFA, conservation is greatest towards the N-termini of the peptides (Fig. 2-10), a region hypothesized to play a role in post-translational processing of the pheromone precursor into the significantly shorter, mature signal peptide, which is farnesylated and carboxy-methyl-esterified at Cys and contains only a portion of the C-terminal residues (Spellig et al. 1994). Duplication of either the STE3 GPCR or MFA, as occurs in Agaricomycotina (Moore et al. 2011), was not apparent from BLAST searches of each genome for additional STE3 orthologs nor MCL similarity clustering of short (90-200 nt) ORFs whose translations demonstrate C-terminal farnesylation motifs characteristic of MFA-type pheromone precursors.

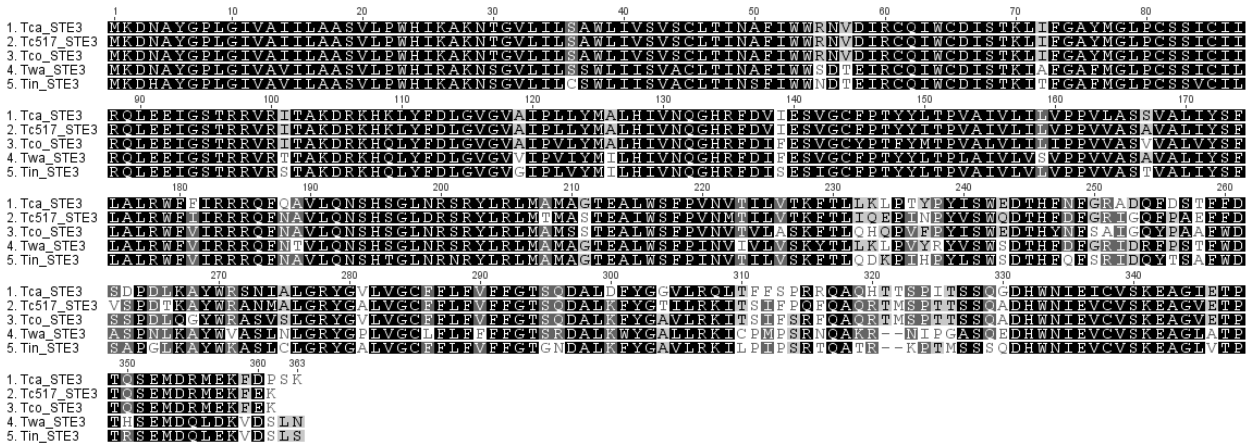

Figure 2-9. Alignment of STE3 orthologs found at MAT-A in *T. caries*, *T. caries 517*, *T. contraversa*, *T. indica* and *T. walkeri*. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels.

Table 2-8. A pairwise similarity matrix depicting percent identity between the STE3 polypeptides aligned in Fig. 3-8.

	Tca_STE3	**Tc517_STE3**	**Tco_STE3**	**Twa_STE3**	**Tin_STE3**
Tca_STE3		86.981	83.934	75.482	76.033
Tc517_STE3	86.981		87.258	75.346	76.454
Tco_STE3	83.934	87.258		75.9	77.839
Twa_STE3	75.482	75.346	75.9		82.825
Tin_STE3	76.033	76.454	77.839	82.825	

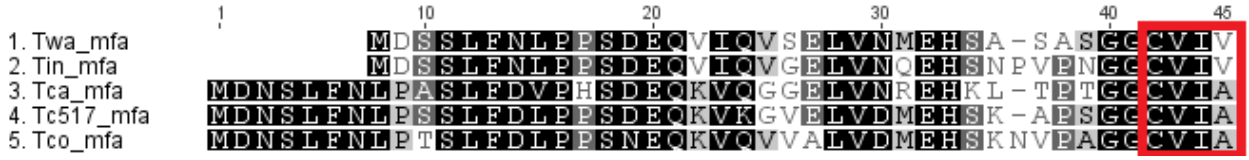

Figure 2-10. Alignment of MFA orthologs found at MAT-A in *T. caries*, *T. caries 517*, *T. contraversa*, *T. indica* and *T. walkeri*. Alignment shaded to indicate similarity with black corresponding to blocks of identical residues and lighter shades of grey corresponding to conservative substitutions or indels. C-terminal farnesylation motif (Cys-aliphatic-aliphatic-x) is boxed in red.

Table 2-9. A pairwise similarity matrix depicting percent identity between the MFA precursor polypeptides aligned in Fig. 3-10.

	Twa_MFA	**Tin_MFA**	**Tca_MFA**	**Tc517_MFA**	**Tco_MFA**
Twa_MFA		81.579	54.054	64.865	57.895
Tin_MFA	81.579		57.895	60.526	60.526
Tca_MFA	54.054	57.895		75	68.889
Tc517_MFA	64.865	60.526	75		82.222
Tco_MFA	57.895	60.526	68.889	82.222	

Synteny Analyses

MAT-a synteny maps are shown in Figure 2-11 and MAT-b synteny maps in Figure 2-12. The contigs containing the MAT-b subloci were only large enough for comparison among *T. caries* 517, *T. indica* and *T. walkeri*. Contig lengths varied for The MAT-a subloci and all are included below.

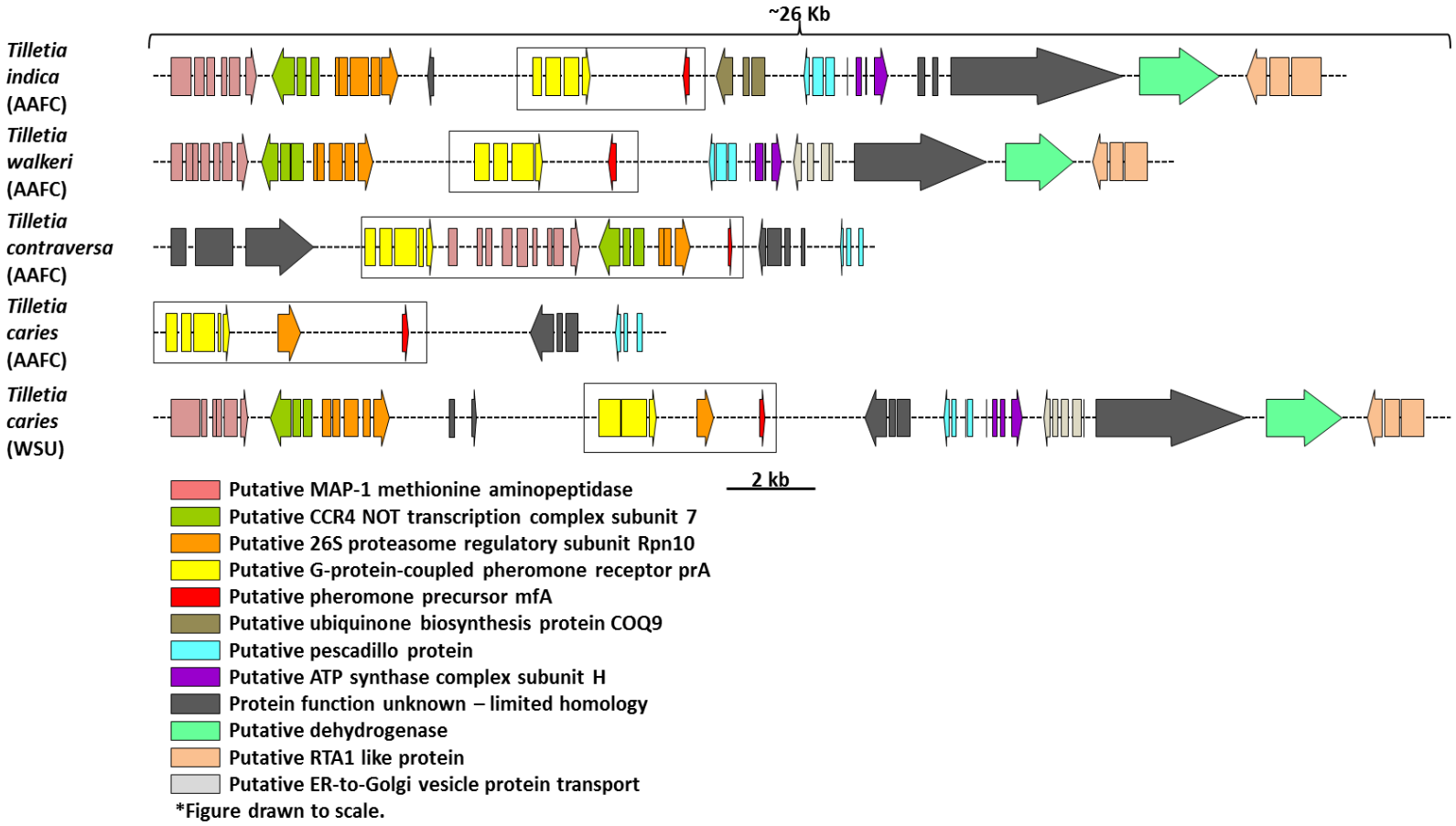

Figure 2-11. Synteny maps of the MAT-a sublocus for *T. caries*, *T. caries 517*, *T. contraversa*, *T. indica* and, *T. walkeri*. Contigs containing the MAT-a genes were identified as described above.

Synteny around the STE3-GPCR and MFA pheromone precursor at MAT-a is generally well conserved with few minor differences. *T. contraversa* represents the largest departure from synteny. Two genes found upstream of the STE3 GPCR in other species, specifically the putative MAP-1 methionine aminopeptidase and putative CCR4 NOT transcription complex subunit 7, are instead located downstream of the STE3 GPCR and between it and its associated MFA pheromone precursor in *T. contraversa*. In the two strains of *T. caries* investigated, the putative 26S proteasome regulatory subunit Rpn1 is located in this region, while in *T. indica* and *T. walkeri*, this region contains no identifiable protein-coding genes. Upstream of the STE3 GPCR is a large gene of unknown function occupying the space where MAP-1 and CCR4 NOT 7 are found in *T. caries* 517. In the AAFC strain of *T. caries*, this upstream region cannot be determined from the available assembly. A major difference between the *T. caries* and *T. contraversa* isolates and the *T. indica* and *T. walkeri* isolates is the direction of transcription of the MFA gene relative to the STE3 GPCR. In the former group, this protein is transcribed in the same direction as STE3, while in the latter it is transcribed in the opposite direction. Downstream from STE3 and MFA is another variable region. In *T. indica* the putative ubiquinone biosynthesis protein COQ9 is present while in *T. walkeri* no syntenic gene could be identified. In the *T. caries* and *T. contraversa* isolates, this same locus encodes a protein of unknown function. Further downstream synteny is well conserved in all species investigated, though for the AAFC strains of *T. caries* and *T. contraversa*, the status of the assembly prevented annotation beyond the putative pescadillo protein.

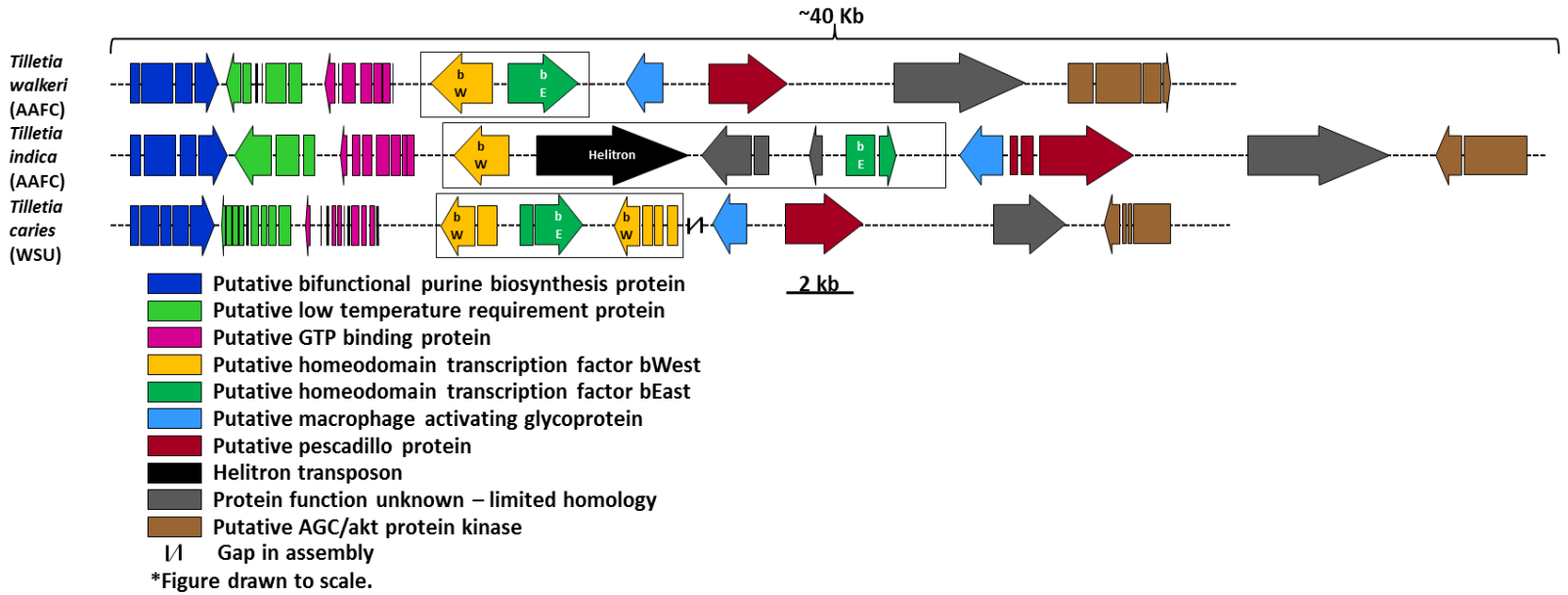

Figure 2-12. Synteny map of the MAT-b sublocus for *T. caries*, *T. indica* and, *T. walkeri*. Contigs containing the MAT-b genes were identified as described above.

Synteny around the mating-type-determining HDP genes at MAT-b is well conserved in these *Tilletia* species; however the number and order of the HDPs at each sublocus reveals some intriguing differences in b-type HDP copy number, for example, the already remarked upon duplication of the bW gene in *T. caries* 517. The gap in the *T. caries* 517 genome assembly just downstream from the second bW ortholog should be resolved to determine if duplication follows this pattern found in Agaricomycotina. *T. indica* demonstrates at least one transposon insertion between its bE and bW type HDPs. Synteny is otherwise entirely conserved around the HDP genes with one exception, an inversion in *T. walkeri* of the AGC/akt kinase found downstream of bW relative to *T. caries* and *T. indica*. *T. caries* 517 and *T. walkeri* demonstrate the classical divergently transcribed gene order originally discovered in *U. maydis*.

DISCUSSION

Duplication of Mating-Type HDPs and Other MAT-b Peculiarities

This study revealed tandem duplication of b-type HDPs at the MAT-b sublocus in *T. caries* 517. Thus far, the presence of duplicates of mating-type-determining homeodomain transcription factors is only known in Agaricomycotina (James et al. 2006; James et al. 2013; Ohm et al. 2010; Pardo et al. 1996); no members of Ustilaginomycotina have yet been shown to exhibit duplication of MAT-b type HDPs. The two bW-type HDP-encoding genes on the 5' and 3' distal ends of the bE gene are present in the haploid genome of *T. caries* strain 517 and Sanger sequencing of the locus containing the mating-type HDP genes confirms that the genome of this strain encodes both of these bW genes. cDNA sequence data provides additional evidence

that these tandem HDP-encoding genes at MAT-B are transcribed. It is unlikely that *T. caries* 517 is a dikaryon or mixed monokaryon because of the method by which it was isolated and its morphology and manner of growth in culture, which is typical of a monokaryon (data not shown). Extensive duplication of putative mating-type homeodomain transcription factors was identified in this study in all strains of systemically infecting bunts including the AAFC strains of *T. controversa* and *T. caries*; however colocation at the mating locus could only be verified for *T. caries* 517. MAT-b HDPs have previously been found in divergently transcribed pairs in other smut fungi (Bakkeren et al. 2008) as well as in Agaricomycotina, where tandem duplication of divergently transcribed MAT-type HDP cassettes is common, but seldom is just one of these genes duplicated rather than the whole cassette (Moore et al. 2011). In Agaricomycotina the single pair of HDPs at the A locus is seen in *Laccaria bicolor*, *Lentinula edodes*, *Pholiota macrospora* and *Postia placenta* (Niculita-Hirzel et al. 2008; Au et al. 2014; Martinez et al. 2009; Yi et al. 2009). Two subloci, each with a pair of HDPs, are seen in *Coprinellus disseminatus* and *Flammulina velutipes* (James et al. 2006; van Peer et al. 2011). Three pairs of HDPs are found at the A locus in *Coprinopsis cinerea* (Pardo et al. 1996) and eight total at both A subloci in *S. commune* (Ohm et al. 2010). A gap is present in the *T. caries* 517 assembly less than 1 kb from Tc517_bW_1.2. It is possible that further duplication, for example of the bE gene, is present in this gap. This arrangement would be more consistent with that found in Agaricomycotina and requires further investigation.

In most fungi investigated thus far, mating type HDPs are present in divergently transcribed pairs of bE and bW orthologs (Bakkeren et al. 2008). *T. indica* has a transposon insertion between its divergently transcribed homeodomain proteins. This helitron transposon,

commonly found in plants, replicates by a rolling circle mechanism and often takes adjacent genes along with it (Kapitonov and Jurka 2001; Lisch 2013). The two other genes in the region between bE and bW are of unknown function and may have hitched a ride along with the helitron. Transposon insertion between divergently transcribed MAT-b HDPs has not yet been reported in any other fungi investigated.

Mating Types Present

Based on these data, each STE3 and MFA ortholog represents a different mating type. Interspecific variation might account for differences between the *T. indica*-*T. walkeri* and dwarf and common bunt groups, but cannot explain the differences within the dwarf and common bunt group, particularly between the two *T. caries* isolates. If these were of the same mating type, 100% identity would be expected for both the STE3 pheromone receptor and the pheromone precursor genes. Sexual compatibility between *T. caries* and *T. contraversa* has been extensively documented (Holton 1954; Metzger and Hoffmann 1978; Silbernagel 1964; Trail and Mills 1990) as has conspecificity of *T. caries*, *T. contraversa* and, *T. laevis* (Banowitz et al. 1984; Castlebury et al. 2005; Kawchuk et al. 1988; Russel and Mills 1992). Conspecificity could explain the higher percent identity between the *T. caries* 517 and *T. contraversa* MFA peptides than between those of the two strains of *T. caries*. If these bunts are conspecific, and certain mating type alleles are always associated with specific sporidial morphological types, or infection phenotypes (dwarfing of host plants or vernalization requirements), these fungi would represent an interesting example of sex-linked phenotypic variation, similar to that seen in animals.

While the differences between the STE3 and MFA genes at MAT-a suggest that these three isolates are of different mating types, the relationships between the MAT-b HDPs contradict this. Figure 2-13 shows the pairwise percent identities per mating-type gene between dwarf and common bunt species. Two out of four homeodomain proteins in the AAFC *T. caries* and *T. contraversa* strains were identical (with the exception of different predicted start codons that did not shift the reading frame, which could simply represent an alternative translation initiation site). *T. caries* 517's HDPs showed less conservation than these first two species. The variation in sequence conservation of the b-type HDPs relative to the consistent lack of conservation between the MAT-a genes suggests that recombination events may have occurred between the MAT-a and MAT-b subloci before divergence. Although all *Tilletia* species that have been studied have bipolar mating systems (Durán and Cromarty 1977; Holton and Kendrick 1957; Hoffmann and Kendrick 1965; 1969), recombination between mating subloci in predominately bipolar species is known to occur (Coehlo et al. 2008; Coelho et al. 2011). The presence of a second bW-type HDP at MAT-B has been verified in *T. caries* 517. The presence of this gene could be the result of hybridization with another wheat bunt species such as *T. contraversa*. This would explain its low sequence identity to the other bW type gene in *T. caries* 517 and its greater similarity to one of the *T. controversa* bW genes than either of the AAFC *T. caries* bW genes. Its presence would increase the possibility of successful MAT-b regulated development in matings with other species as only one functional heterodimer is necessary for this process to occur (Gillissen et al. 1992). It is unclear whether the bW-type-HDP is conserved in the same arrangement in AAFC's *T. caries* and *T. contraversa* strains. The presence of less well conserved third, fourth or fifth bW-type HDPs in the two AAFC strains could both represent additional MAT-b ideomorphs. On the other hand, the differences in MAT-a genes may be

attributed to the quality of the genome assembly, dikaryotization of the sequenced cultures, a mix of different monokaryons, or other recombination between MAT-a and MAT-b. It should be noted that these are not mutually exclusive scenarios, and there could be multiple factors involved. More genome sequence data and population genetic analyses are needed to answer these questions.

Based on these data, it appears that there are three MAT-a alleles in the sequenced strains from the dwarf-common bunt group, one per strain investigated. There are two to three MAT-b alleles in this group, depending on whether the extraneous b-type HDPs from the AAFC group are actually associated with the allele and functional in mating.

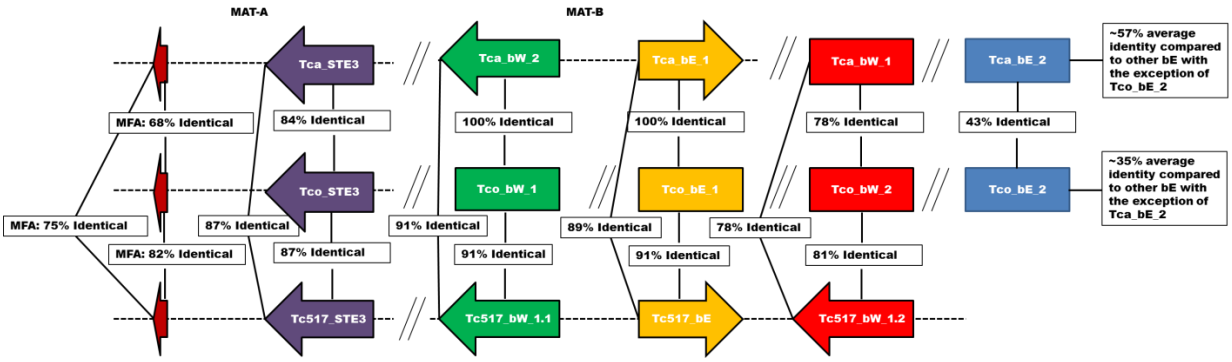

Figure 2-13. Graphical representation of pairwise similarities between mating genes at MAT-a (STE3 pheromone receptor and pheromone precursor) and MAT-b (HDPs). From top to bottom: *T. caries* AAFC, *T. contraversa* and *T. caries* 517. Hashmarks represent gaps in assembly. Dotted lines represent genomic contigs. Figure not drawn to scale and order of genes is arbitrary unless otherwise indicated.

***Relevance to Reports of Multiallelism in* T. contraversa**

It has been previously reported that *T. contraversa* demonstrates mating behavior characteristic of a bipolar MAT-locus with multiple alleles (Hoffmann and Kendrick 1965). Schirawski et al. (2005) previously described multiallelism at MAT-a in *Sporisorium reilianum* as relying on two MFA genes per allele, each specific to a different STE3 GPCR found in two other mating types. This arrangement was identified in other Ustilaginales by Kellner et al. (2011) but is not present in this strain of *T. contraversa* based on the results of the present study. A second MFA pheromone precursor gene could not be identified in the species included at the MAT-a subloci. On the other hand, in the sequenced dwarf and common bunt species, three distinct alleles (in the bipolar sense, with one allele referring to both linked MAT-a and MAT-b subloci) are present. The presence of three alleles at MAT-a means one of two things: either each

MAT-a type is capable of mating with only one other MAT-a type or each distinct pheromone is capable of activating two different STE3 GPCRs. In the former case, multiple compatibility groups would be apparent. In the latter case conjugation would always occur, though would not always be followed by b-regulated development. This situation would be analogous to the B(HDP)-compatible, A(STE3)-incompatible reaction observed in *S. commune* termed the "flat" phenotype (Vaillancourt et al. 1997). Because this latter case has not been recorded in *T. caries* or *T. contraversa,* for the bipolar multiallelism previously observed to actually occur, each pheromone must only be capable of activating one other STE3 GPCR. The presence of more than the traditional two b-type HDP proteins in each strain investigated further complicates the situation. If all b-type HDPs are functional, b-regulated development following conjugation would be almost certain to occur in each a-compatible mating interaction. Gillissen et al. (1992) showed that for *U. maydis*, only one heterodimer of b-type HDPs from different alleles is necessary to trigger b-regulated development. If each of these b-type HDPs is functional, then the relevance of b-compatibility as a checkpoint to pathogenic development could be removed, leaving a-compatibility as the only requirement to initiate mating and pathogenesis. While a similar phenomenon has been reported in Agaricomycotina where only the A-type HDPs determine mating type (Aimi et al. 2005; James et al. 2006), this would be a novel discovery in Ustilaginomycotina. Hoffmann and Kendrick's (1965) and Carris and Gray's (1994) observations that conjugation in pairings of *T. contraversa* x *T. caries* and *T. contraversa* x *Tilletia bromi* is not always followed by b-regulated development provides evidence against this last scenario. It is clear that b-type HDPs are not active in all matings.

Dikaryotization and its Potential Implications for Next Gen Sequencing

While it is unlikely that *T. caries* 517 is a dikaryon or culture containing more than one monokaryon, the nuclear condition of *T. contraversa* or *T. caries* strains from AAFC is not known (S. Hambleton, pers. comm.), and fragmentation of the genomic locus containing the MAT-b mating-type genes in the current genome assemblies precludes accurate analysis of the locus. Four MAT-b mating-type genes were identified in both AAFC assemblies. In the AAFC strain of *T. caries*, two MAT-b mating type genes can be assigned to the same contig (divergently transcribed Tca_bW_2 and Tca_bE_2), but this the case with *T. contraversa*. The contig from AAFC's strain of *T. caries* is too short determine if the additional duplication of bW found in *T. caries* 517 is present at the same sublocus. The duplication of bW is unlikely the result of dikaryotization or presence of mixed monokaryons in the strain 517 because only one ortholog of the STE3 pheromone receptor is present in the assembly. However, considering the higher similarity seen in alignments of the STE3 pheromone receptors relative to the HDP alleles, it is possible that the presence of one receptor per strain could be an artefact of the sequencing and assembly process. The inability to resolve larger sections of the MAT subloci in *T. caries* and *T. controversa* could also be attributed to this dikaryotization, for example in MAT loci where synteny between otherwise orthologous regions of the genome breaks down. Longer reads, such as those obtained using PacBio technology, in conjunction with the available short-read data is needed to confirm the organization of the mating loci. The construction of a BAC library, hybridization experiments to isolate colonies containing the MAT subloci, PCR and, Sanger sequencing, could be used to specifically investigate the MAT subloci instead of further whole-genome sequencing. Another approach would be to isolate single basidiospores and monokaryotic cultures from the AAFC collections and repeat genomic sequencing.

Future Research

In order to confirm the arrangement and function of these mating-type genes a number of additional assays are needed. First, further sequencing of the AAFC strains of *T. caries* and *T. contraversa* should be carried out, preferably on a long-read platform such as PacBio. This would help to link together the fragmented contigs containing the mating loci. Prior to this, ploidy should be confirmed with fluorescence microscopy such as DAPI to determine whether they are mono- or dikaryons. The dwarf and common bunt cultures should be paired in all combinations and pathogenicity tests carried out on wheat to determine if they are of compatible mating types. In addition, *in vitro* assays would determine which of the b-type HDPs are functional in mating. This could be accomplished using a heterologous system approach by transforming the genes into *U. maydis* or *U. hordei*, which are well developed model systems amenable to this type of experimentation (Feldbrügge et al. 2013). Functional b-type HDPs from *Tilletia* spp. would differ enough from endogenous *Ustilago* HDPs to trigger b-regulated development in the haploid *Ustilago* strains, rendering them capable of causing disease on maize (Guus Bakkeren, pers. comm.). Alternatively, haploid *Ustilago* strains with endogenous b-type HDP knockouts could be constructed or obtained and all b-type HDPs identified in *Tilletia* spp. could be introduced in pairwise combinations. This would be followed by growth and pathogenicity tests, thus confirming the compatibility (or lack thereof) between the various *Tilletia* b-type HDPs.

CHAPTER TWO REFERENCES

Aimi T, Yohida R, Ishikawa M, Bao D and, Kitamoto Y. 2005. Identification and linkage mapping of the genes for the putative homeodomain protein hox1 and the putative pheromone receptor protein homologue (rcb1) in a bipolar basidiomycete *Pholiota nameko*. Current Genetics 48: 184-194.

Altschul SF, Gish W, Miller W, Myers EW and, Lipman DJ. 1990. Basic Local Alignment Search Tool. Journal of Molecular Biology 215: 403-410.

———, Madden TL, Schäffer AA, Zhang Z, Miller W and, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25: 3389-3402.

Au CH, Wong MC, Bao D, Zhang M, Song C, Song W, Law PTW, Kües U and, Kwan H. 2014. The genetic structure of the *A* mating-type locus of *Lentinula edodes*. Gene 535: 184-190.

Bakkeren G, Kämper J and, Schirawski J. 2008. Sex in smut fungi: structure function and evolution of mating-type complexes. Fungal Genetics and Biology 45: 515-521.

——— and Kronstad JW. 1993. Conservation of the *b* mating-type gene complex among bipolar and tetrapolar smut fungi. The Plant Cell 5: 123-136.

Banowitz GM, Trione EJ and, Krygier BB. 1984. Immunological comparisons of teliospores of the two wheat bunt fungi, *Tilletia* species using monoclonal antibodies and antisera. Mycologia 76(1): 51-62.

Bölker M, Urban M and, Kahmann R. 1992. The *a* mating type locus of *U. maydis* specifies cell signaling components. Cell 68: 441-450.

Boyd ML and Carris LM. 1997. Molecular relationships among varieties of the *Tilletia fusca* (*T. bromi*) complex and related species. Mycological Research 101(3): 269-277.

Carris LM and Gray PM. 1994. The ability of *Tilletia fusca* to hybridize with the wheat bunt species under axenic conditions. Mycologia 86(1): 157-163.

Castlebury LA, Carris LM and, Vánky K. 2005. Phylogenetic analysis of *Tilletia* and allied genera in order Tilletiales (Ustilaginomycetes; Exobasidiomycetidae) based on large subunit nuclear rDNA sequences. Mycologia 97(4): 888-900.

Coelho MA, Gonçalves P and, Sampaio JP. 2011. Evidence for maintenance of sex determinants but not of sexual stages in red yeasts, a group of early diverged basidiomycetes. BMC Evolutionary Biology 11: 249-263.

———, Rosa A, Rodrigues N, Fonseca A and Gonçalves P. 2008. Identification of mating type genes in the bipolar basidiomycetous yeast *Rhodosporidium toruloides*: First insight into the *MAT* locus structure of the *Sporidiobolales*. Eukaryotic Cell 7(6):1053-1061.

Durán R and Cromarty R. 1977. *Tilletia indica*: a heterothallic bunt fungus with multiple alleles controlling incompatibility. Phytopathology 67: 812-815.Coelho MA, Gonçalves P and Sampaio JP. 2011. Evidence for maintenance of sex determinants but not of sexual stages in red yeasts, a group of early diverged basidiomycetes. BMC Evolutionary Biology 11: 249-263. Enright AJ,

Feldbrügge M, Kämper J, Steinberg G and, Kahmann R. 2004. Regulation of mating and pathogenic development in *Ustilago maydis*. Current Opinion in Microbiology 7: 666-672

Flor HH. 1932. Heterothallism and hybridization in *Tilletia tritici* and *T. levis*. Journal of Agricultural Research. 44: 49-58.

Froeliger EH and Leong SA. 1991. The *a* alleles of *Ustilago maydis* are ideomorphs. Gene 100:113-122.

Gillissen B, Bergemann J, Sandmann C, Shroeer B, Bölker M and, Kahmann R. 1992. A two-component regulatory system for self/non-self recognition in *Ustilago maydis*. Cell 68: 647-657.

Giraud T, Yockteng R, López-Villavicencio M, Refrégier G and, Hood ME. 2008. Mating system of the anther smut fungus *Microbotryum violaceum*: selfing under heterothallism. Eukaryotic Cell 7(5): 765-775. Goddard MR, Godfray CJ and Burt A. 2004. Sex increases the efficacy of natural selection in experimental yeast populations. Nature 434: 636-640.

Gold S, Gillian D, Barrett K and, Kronstad J. 1994. cAMP regulates morphogenesis in the fungal pathogen *Ustilago maydis*. Genes and Development 8: 2805-2816.

Halisky PM. 1965. Physiologic specialization and genetics of the smut fungi. III. The Botanical Review 31(1): 114-150.

Heitman J. 2006. Sexual reproduction and the evolution of microbial pathogens. Current biology 16: R711-R725.

Hoff KJ and Stanke M. 2013. WebAUGUSTUS – a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Research 41(W1): W123-W128.

Hoffmann JA and Kendrick El. 1965. Compatibility relationships in *Tilletia contraversa*. Phytopathology 55: 1061-1062 (abstr).

——— and Kendrick EL. 1969. Genetic control of compatibility in *Tilletia contraversa*. Phytopathology 59(1): 79-83.

Holton CS. 1954. Natural hybridization between common and dwarf bunt as related to the problem of delimitation of species of *Tilletia* occurring on wheat. Phytopathology 44: 493 (abstr).

——— and Kendrick EL. 1957. Fusion between secondary sporidia in culture as a valid index of sex compatibility in *Tilletia caries*. Phytopathology 47: 688-689.James TY, Srivilai P,

Huelsenbeck JP and Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17:754-755.

Kawchuk LM, Kim WK and, Nielsen J. 1988. A comparison of polypeptides from the wheat bunt fungi *Tilletia laevis*, *T. tritici*, and *T. contraversa*. Canadian Journal of Botany 66: 2367-2376.

Kües U and Vilgalys R. 2006. Evolution of the bipolar mating system of the mushroom *Coprinellus disseminatus* from its tetrapolar ancestors involves loss of mating-type-specific pheromone receptor function. Genetics 172: 1877-1891.

James TY, Sheng S, Li W, Heitman J, Kuo H, Lee Y, Asiegbu FO, Olson Å and others. 2013. Polyporales genomes reveal the genetic architecture underlying tetrapolar and bipolar mating systems. Mycologia 105(6): 1374-1390. Kahmann R and Kämper J. 2004. *Ustilago maydis*: how its biology relates to pathogenic development. New Phytologist 164: 31-42.

———, Srivilai P, Kües U and Vilgalys R. 2006. Evolution of the bipolar mating system of the mushroom *Coprinellus disseminatus* from its tetrapolar ancestors involves loss of mating-type-specific pheromone receptor function. Genetics 172: 1877-1891.

Kämper J, Reichmann M, Romels T, Bölker M and Kahmann R. 1995. Multiallelic recognition: nonself-dependent dimerization of the bE and bW homeodomain proteins in *Ustilago maydis*.

Kapitonov VV and Jurka J. 2001. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America. 98(15): 8714-8719.

Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R and, Salzberg SL. 2013. TopHat2: Accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biology 14: R36.

Kronstad JW and Leong SA. 1989. Isolation of two alleles of the *b* locus of *Ustilago maydis*. Proceedings of the National Academy of Sciences of the United States of America 86: 978-982.

Lee SB, Taylor JW. 1990. Isolation of DNA from fungal mycelia and single spores. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ, eds. PCR protocols: a guide to methods and applications. Orlando, Florida: Academic Press. p 282–287.

Lisch D. 2013. How important are transposons for plant evolution? Nature Reviews Genetics 13: 49-61.

Martinez D, Challacombe J, Morgenstern I, Hibbett D, Schmoll M, Kubicek CP, Ferreira P, Ruiz-Duenas FJ, Martinez AT, Kersten P and others. 2009. Genome, transcriptome and secretome analysis of wood decay fungus *Postia placenta* supports unique mechanisms of lignocellulos conversion. Proceedings of the National Academy of Sciences of the United States of America 106(6): 1954-1959.

Marcheler-Bauer A, Lu S, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR and others. 2011. CDD: a conserved domain database for the functional annotation of proteins. Nucleic Acids Research 39(Database issue): D225-229.

Metzger RJ and Hoffmann JA. 1978. New races of common bunt useful to determine resistance of wheat to dwarf bunt. Crop Science 18: 49-51.

McWilliam H, Li W, Uludag M, Squizzato S, Park YM, Buso N, Cowley AP and, Lopez R. 2013. Analysis Tool Web Services from the EMBL-EBI. Nucleic Acids Research 41(Web Server issue)" W597-600.

Moore D, Robson GD and Trinci APJ. 2011. 21st Century Guidebook to Fungi. Cambridge University Press. New York, NY.

Müller P, Weinzierl G, Brachmann A, Feldbrügge M and, Kahmann R. 2003. Mating and pathogenic development of the smut fungus *Ustilago maydis* are regulated by one mitogen-activated protein kinas cascade. Eukaryotic Cell 2(6): 1187-1199.

Niculita-Hirzel H, Labbé J, Kohler A, le Tacon F, Martin R, Sanders IR and, Kües U. 2008. Gene organization of the mating type regions in the ectomycorrhizal fungus *Laccaria bicolor* reveals distinct evolution between the two mating type loci. New Phytologist 180(2): 329-342.

Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, Stajich JE, de Vries RP, Record E, de Vries RP, Record E and, others. 2010. Genome sequence of the model mushroom *Schizophyllum commune*. Nature Biotechnology 28: 957-963.

Pardo EH, O'Shea SF and, Casselton LA. 1996. Multiple versions of the *A* mating type locus of *Coprinus cinereus* are generated by three paralogous pairs of multiallelic homeobox genes. Genetics 144: 87-94.

Petit E, Giraud T, de Vienne DM, Coelho MA, Aguileta G, Amselem J, Kreplak J, Poulain J, Gavori F, Wincker P and others. 2012. Linkage to the mating-type locus across the genus *Microbotryum*: Insights into nonrecombining chromosomes. Evolution 66(11): 3519-3533.

Puhalla JE. 1968. Compatibility reactions on solid medium and interstrain inhibition in *Ustilago maydis*. Genetics 60: 461-474.

Raudaskoski M and Kothe E. 2010. Basidiomycete mating type genes and pheromone signaling. *Eukaryotic Cell* 9(6): 847-859.

Roberts A, Trapnell C, Donaghey J, Rinn JL and, Pachter L. 2011. Improving RNA-seq expression estimates by correcting for fragment bias. Genome Biology 12: R22.

Ronquist F and Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 9: 1572-2574.

Rowell JB and DeVay JE. 1954. Genetics of *Ustilago zeae* in relation to basic problems of its pathogenicity. Phytopathology 44: 356-362.

Rozen S and Skaletsky HJ. 2000. Primer3 on the WWW for General Users and for Biologist Programmers. In: Misener S and Krawetz SA eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Humana Press Inc., Totowa (NJ). pp 365-386. Source code available at http://fokker.wi.mit.edu/primer3/

Russel BW and Mills D. 1992. Electrophoretic karyotypes of *Tilletia caris*, *T. contraversa*, and their F1 progeny: Further evidence for conspecific status. Molecular Plant-Microbe Interactions 6(1): 66-74.

Schirawski J, Heinze B, Wagenknecht M and, Kahmann R. 2005. Mating type loci of *Sprisorium reilianum:* Novel pattern with three *a* and multiple *b* specificities. Eukaryotic Cell 4(8): 1317-1327.

Schulz B, Banuett F, Dahl M, Schlesinger R, Schäfer W, Martin T, Herskowitz I and, Kahmann R. 1990. The *b* alleles of *U. maydis*, whose combinations program pathogenic development, code for polypeptides containing a homeodomain-related motif. Cell 60: 295-306.

Scott MP, Tamkun JW and, Hartzell GW III. 1989. The structure and function of the homeodomain. Biochimica et Biophysica Acta 989: 295-306.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J and, others. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7: 539.

Silbernagel M. 1964. Compatibility between *Tilletia caries* and *T. contraversa*. Phytopathology 54: 1117-1120.

Snetselaar KM, Bölker M and, Kahmann R. 1996. *Ustilago maydis* mating hyphae orient their growth toward pheromone sources. Fungal Genetics and Biology 20: 299-312.

Solovyev V, Kosarev P, Seledsov I and, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biolology 7, Suppl 1: 10.1-10.12.

Spellig T, Bölker M, Lottspeich F, Frank RW and, Kahmann R. 1994. Pheromones trigger filamentous growth in *Ustilago maydis*. The EMBO Journal 13(7): 1620-1627.

Trail F and Mills D. 1990. Growth of haploid *Tilletia* strains in planta and genetic analysis of a cross of *Tilletia caries* X *Tilletia contraversa*. Genetics 80(4): 367-370.

Tusnády GE and Simon I. 1998. Principles governing amino acid composition of integral membrane proteins: Applications to topology prediction. Journal of Molecular Biology 283: 489-506.

——— and ———. 2001. The HMMTOP transmembrane topology prediction server. Bioinformatics 17: 849-850.

Trapnell C, Hendrickson DG, Sauvageau M, Goff L, Rinn JL and, Pachter L. 2012. Differential analysis of gene regulation at transcript resolution with RNA-seq. Nature Biotechnology 31: 46-53.

———, Pachter L and, Salzberg SL. 2009. TopHat: Discovering splice junctions with RNA-seq. Bioinformatics 25(9): 1105-1111.

———, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, Salzberg SL, Wold BJ and, Pachter L. 2010. Transcript assembly and quantification by RNA-seq reveals unannotated transcripts and isoform switching during cell differentiation. Nature Biotechnology 28(5): 511-515.

Trione EJ. 1964. Isolation and in vitro culture of the wheat bunt fungi *Tilletia caries* and *T. contraversa*. Phytopathology 54: 592-596.

Toome M, Ohm RA, Riley RW, James TW, Lazarus KL, Henrissat B, Albu S, Boyd A, Chow J, Clum A and, others. 2013. Genome sequencing provides insight into the reproductive biology, nutritional mode and ploidy of the fern pathogen *Mixia osmundae*. New Phytologist 202: 554-564.

Vaillancourt LJ, Raudaskoski M, Specht CA and, Raper CA. 1997. Multiple genes encoding pheromones and a pheromone receptor define the Bβ1 mating-type specificity in *Schizophyllum commune*. Genetics 146: 541-551.

Van Dongen S. 2000. Graph clustering by flow simulation. PhD thesis , University of Utrecht.

——— and Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Research 30(7): 1575-1584.

Van Peer AF, Park S, Shin P, Jang K, Yoo Y, Park Y, Lee B, Sung G, James TY and, Kong W. 2011. Comparative genomics of the mating-type loci of the mushroom *Flammulina velutipes* reveals widespread synteny and recent inversions. PLoS One 6, e22249.

Yi R, Tachikawa M, Mukaiyama H, Bao D and, Aimi T. 2009. Genomic structure of the *A* mating-type locus in a bipolar basidiomycete *Pholiota nameko*. Mycological Research 113: 240-248.

CHAPTER THREE

INTRODUCTION

Tilletia caries (syn. *T. tritici*) and *Tilletia laevis* (syn. *T. foetida*) (Tilletiales, Exobasidiomycetidae, Ustilaginomycotina, Basidiomycota), also known as common bunt or stinking smut, are two of the world's most devastating pathogens of cultivated wheat (*Triticum aestivum*) (Fischer and Holton 1957). Few *Tilletia* genomes have been sequenced and none are publically available, thus information on gene and protein sequences from this group are limited. The genome of *T. caries* strain 517 was sequenced in 2012 at the Washington State University Laboratory for Biotechnology and Bioanalysis. Access to the *T. caries* genome presents an opportunity to learn more about cellular processes in this biotrophic pathogen.

Ribonucleic acid (RNA) is a type of nucleic acid integral to cellular function. Each ribonucleotide is composed of a pentose sugar (ribose) with a covalently bonded nitrogenous base and a 5' phosphate group. Unlike the pentose sugars of deoxyribonucleotides comprising DNA, those of RNA possess a 2' hydroxyl group, and instead of the base thymine (T), uracil (U) is incorporated to complement adenosine (A). Polymers of multiple RNA molecules perform various functions in the cell; however one of the most important is messenger RNA or mRNA. mRNA is transcribed by RNA polymerase II from the coding sequence of a gene, modified with a poly-adenylate tail and 5' cap and then transported out of the nucleus to the ribosome where the gene's coding sequence is translated into protein (Nelson and Cox 2008). Regulation of gene expression may occur at the transcriptional or post-transcription level. RNA silencing is a form of post-transcriptional regulation of gene expression (Matzke and Birchler 2005). RNA silencing was first demonstrated in plants (Napoli et al. 1990, van der Krol 1990) and has been shown to be a

common adaptation across Eukarya, occurring in mammals (Fire et al. 1998), Protoctista (Ngô et al. 1998) as well as fungi (Romano and Macino 1992). The basic mechanism of RNA silencing relies on small, double-stranded RNAs (21-26 nt) which interact with a number of proteins to degrade complementary RNA molecules. These RNAs are termed small interfering RNA (siRNA) if they originate from double-stranded (dsRNA) and micro-RNA (miRNA) if they originate from hairpin RNAs (Matzke and Birchler 2005). RNA silencing in eukaryotes is involved in cellular functions such as regulation of developmental transcripts, transcript degradation, translation inhibition, defense against viruses and other foreign RNAs, as well as chromatin remodeling and control of transposable elements (Matzke and Birchler 2005).

Most fungi investigated thus far appear to contain genes needed for some form of RNA silencing with selective instances of loss of RNA silencing capabilities (Nunes et al. 2011). Examples of fungi that have lost RNA silencing capabilities include *Ashbya gossypii, Candida glabrata, Debaryomyces hansenii, Hanseniaspora uvarum, Kluyveromyces waltii, K. lactis, Saccharomyces cerevisiae, S. kluyveri, S. kudriavzevii, S. mikatae, S. paradoxus, Ustilago maydis* and *Zygosaccharomyces bailii* (Drinnenberg et al. 2011).

Fungi carry out post-translational gene regulation using siRNA-mediated RNA silencing in a manner similar to that found in plants and animals. This function is termed quelling (Romano and Macino 1992, Cogoni and Macino 1997) and is involved in control of the cell cycle and regulation of developmental processes in both fungi (Nunes et al. 2011) and eukaryotes in general (Wilson and Doudna 2013).

The genes involved in RNA silencing have diverged among the fungi (Nakayashiki et al. 2005), giving rise to additional unique roles including the silencing of unpaired DNA during

meiosis in *Neurospora crassa* (Shiu et al. 2001) and epigenetic roles such as the formation of heterochromatin in *Schizosaccharomyces pombe* (Verdel et al. 2004). Recently a much wider variety of small RNAs performing various roles have been discovered in fungi, yet the hairpin forming miRNAs described in other eukaryotes have yet to be identified (Nunes et al. 2011).

A particularly promising area of fungal RNA silencing research involves host-induced gene silencing, an RNA silencing mediated control strategy for fungal pathogens (Nunes and Dean 2012). In this process fungal transgenes are introduced into a host plant in order to silence target transcripts in the fungal pathogen via RNA silencing during infection (Nunes and Dean 2012). The role of RNA silencing in plant pathogenic fungi is complex but is not essential for pathogenesis. For example, the corn smut pathogen *Ustilago maydis* does not possess any of the genes necessary for RNA silencing (Kämper et al. 2004) and the genes found in the sorghum head smut pathogen *Sporisorium reilianum* are apparently dispensable for infection, with deletion mutants not differing from wild type strains in their pathogenic phenotypes (Shirawski et al. 2010).

There are several key proteins involved in RNA silencing (Figure 3-1): Dicer, a nuclease, produces small dsRNA out of larger dsRNA molecules and helps guide them to complementary RNA targeted for degradation; the RNA-dependent RNA polymerase (RdRP) present in plants and fungi, which transcribes ssRNA to form duplexes, amplifying the RNA silencing signal; and Argonaute, which carries out a key nuclease function in silencing-protein complexes such as the RNA-induced silencing complex (RISC) and is instrumental in quelling. The dsRNA duplex migrates with Dicer to RISC but only one strand, termed the guide strand, is involved in target RNA degradation while the other is lost. These proteins are often present as paralogs in the cell and are implicated in varying RNA silencing functions (Tomari and Zamore 2005). Orthologs are

often located in different parts of the cell depending on their function (Tomari and Zamore 2005). When dsRNA is processed within the nucleus at least one additional protein in the exportin family is required for export of the duplex to the cytoplasm. In mammals this protein is Exportin-5 and is related to miRNA export (Yi et al. 2003), however fungi possess the ortholog Exportin-1/CRM-1 that is also implicated in nuclear export of small RNAs (Bai et al. 2006). The diversification of RNA silencing-related genes in fungi is reflected in the varying numbers of orthologs present in different fungal genomes (Nakayashiki et al. 2005). The copy numbers of RNA silencing related genes in a variety of fungi are shown in Table 3-1.

Table 3-1. The copy numbers of genes involved in RNA silencing in a variety of ascomycetes and basidiomycetes (Nunes et al. 2011; Schirawski et al. 2010).

Species	**Dicer**	**Argonaute**	**RdRP**	**Exportin-1**
Saccharomyces cerevisiae	0	0	0	1
Aspergillus niger	3	3	2	2
Neurospora crassa	2	3	4	4
Coprinopsis cinerea	3	8	8	2
Cryptococcus neoformans	4	2	2	4
Postia placenta	1	3	8	0
Malassezia globosa	0	0	0	2
Ustilago maydis	0	0	0	2
Sporisorium reilianum	1	1	3	1
Tilletia caries	3	6	1	1

The proteins involved in RNA silencing possess functional domains required for their operation. Figure 3-1 shows some of these genes from the maize pathogen *Sporisorium reilianum*. Dicer class nucleases possess a number of important functional domains. These include RNase III domains, helicase motifs, a dsRNA binding site and a PAZ domain, also found in Argonaute proteins (Bernstein et al. 2001). While the PAZ (Ma et al. 2004) and dsRNA binding domains

(Dlakić 2006) function in dsRNA binding, the ATP-dependent helicase unwinds the duplex (Kim 1998, Theis 1999) and the RNase III domains process the duplex into small, dsRNA (Blaszczyk et al. 2001). The Argonaute family of nucleases forms the catalytic component of RISC and it is therefore instrumental in the actual degradation of target RNA. Domains characteristic of Argonaute proteins include the Piwi domain, a PAZ domain and a conserved Domain of Unknown Function (DUF1785). The Piwi domain functions both as an anchoring site for the guide RNA and the catalytic site for slicing the target while the PAZ domain performs similar functions to those found in Dicer class proteins, helping to bind dsRNA (Ji-Joon and Joshua-Tor 2006). RdRP enzymes transcribe errant ssRNA in the cell to form duplexes and also amplify the RNA silencing signal, transcribing dsRNA templates (Dalmay et al. 2000).

Exportin-1 or CRM-1 is part of a protein complex involved in nuclear export. It associates with a number of other proteins including those with a nuclear export signal. It includes an Xpo1 domain which interacts with the leucine-rich nuclear export signal (Ullman et al. 1997), an importin-beta N-terminal domain which interacts with the nuclear pore (Vetter et al. 1999), and a CRM1 C-terminal domain which interacts with several other proteins such as GTPase Ran and a nuclear export signal containing protein (Petosa et al. 2004). Its involvement in nuclear RNA export in fungi has yet to be elucidated but if it is involved, one of its partner proteins must have RNA binding capability.

Given that RNA silencing is a nearly ubiquitous phenomenon in eukaryotes, it is likely that *T. caries* utilizes some form of RNA-silencing in cellular processes. It could also have lost this capability, as has occurred in a small number of other fungi. Given the increasing potential of using RNA-silencing-related approaches such as host-induced gene silencing for the control of pathogenic

fungi (Nunes and Dean 2012), determining the presence or absence of RNA-silencing in *T. caries* is warranted. Because of the well characterized nature of RNA silencing-associated proteins in eukaryotes, it is possible to mine the genome sequence of *T. caries* for homologous genes by comparing the conserved domains present in predicted peptides. In this study, multiple putative RNA silencing genes are identified *in silico* and annotated. Predicted protein products are phylogenetically compared to homologous proteins found in related organisms.

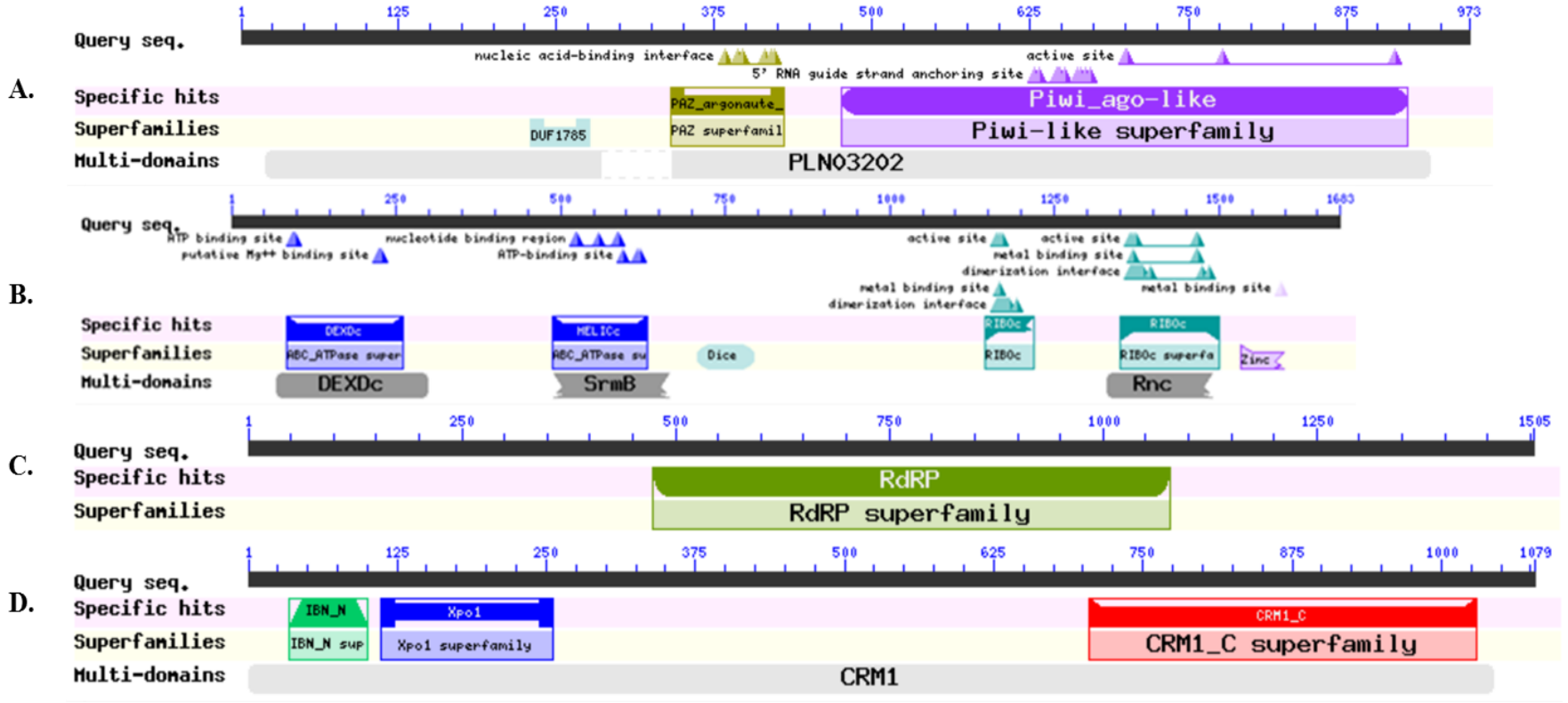

Figure 3-1. RNA silencing-related proteins from the genome of *Sporisorium reilianum* and their conserved functional domains. A. CBQ68774: Argonaute. B. CBQ71581: Dicer. C. CBQ71530: RdRP D. CBQ73251: Exportin-1. Numbers refer to the length in amino-acid residues of the proteins. Graphics from the NCBI Conserved Domain Database (Marchler-Bauer et al. 2011).

METHODS

Tilletia caries strain 517 (WSP72095) was collected at a Washington State University wheat research plot in Whitman Co., WA, USA in July 2006 by Xianming Chen. A haploid monosporidial line was established in culture using the methods described in Boyd and Carris (1997). Secondary sporidia and mycelial fragments scraped from cultures grown on M19, a semi-defined medium (Trione 1969), were suspended in 200 µL sterile distilled de-ionized water and streak-plated onto 100 x 15 mm Petri dishes of M19 amended with 40 U/mL penicillin and 40 µg/mL streptomycin (Life Technologies, Grand Island, NY) overlain with sterile SpectraPor dialysis membrane (VWR Scientific, Radnor, PA). Plates were incubated in the dark at 15C for 2 – 4 weeks in unsealed plastic culture boxes that allowed air exchange.

Mycelium for DNA extraction was removed from dialysis membranes with a sterile metal spatula, placed in a 1.5 mL microfuge tube, frozen at -80 C for 24 hours, and lyophilized overnight. Lyophilized mycelium was ground using a sterile mortar and pestle and stored at -20 C prior to use. DNA was extracted from approximately 50 mg of ground, lyophilized mycelium using a modified phenol-chloroform extraction procedure (Lee and Taylor 1990).

Whole Genome Sequencing

Tilletia caries monokaryotic strain 517 (strain details in Table S1), isolated as described in Carris and Gray (1994), was sequenced at the Washington State University Laboratory for Biotechnology and Bioanalysis on PacBio RS and Roche 454 sequencing platforms. A hybrid assembly of both datasets was carried out by Dr. Jodi Humann in the WSU Horticulture Dept.

utilizing Allora (SMRT analysis: http://www.pacificbiosciences.com/products/software/algorithms/) to create corrected continuous long reads. Genome assembly statistics are available in Table S2.

Gene Annotation

Sequences encoding the RNA silencing related genes Dicer, Argonaute, RdRP and exportin-1/CRM-1 in the genome of the related biotrophic fungus *Sporisorium reilianum* (Shirawski et al. 2010) were used to conduct pairwise alignment with the BLAST algorithm (Altschul et al. 1990) against the *T. caries* 517 assembly (accessions CBQ71530, CBQ70480, CBQ68774, CBQ71581, CBQ68774, CBQ73251 and, CBQ73426). A subset of contigs containing matches with Expectation-values (E-value) less than $1e^{-6}$ were selected for gene annotation. Annotation was carried out utilizing FGENESH (Solovyev et al. 2006; http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind&advanced=on) under *Ustilago* parameters. The NCBI Conserved Domain Search Tool (Marcheler-Bauer et al. 2011; http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) set to default parameters was used to check each annotation for homology to RNA silencing related genes. Annotations were visualized in the Genome Sequence Annotation Server (GenSAS) (Lee et al. 2011).

Phylogenetic Analyses

Homologous amino acid sequences were obtained from GenBank (accessions listed in Table S5). Sequence alignment utilized Clustal Omega (Sievers et al. 2011). Alignments were curated in Geneious v7.0.4 (BioMatters Ltd.). Models of amino acid substitution were selected using ProtTest 3.4 (Darriba et al. 2011). Phylogenetic analyses utilized PhyML (Guindon and Gascuel 2003) and MrBayes 3.2.3 (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003). Primers to

amplify full length clones from genomic DNA were designed using Primer3 v0.4.0 (Rozen and Skaletsky 2000). Primers are listed in Table S4.

RESULTS

Conserved Domains Present in T. caries 517's *Putative RNA silencing-Associated Genes*

Putative homologs were identified for Argonaute, Dicer, Exportin-1/CRM-1 and RdRP within the *T. caries* 517 genome assembly. *T. caries* appears to have a relatively large number of Argonaute and Dicer orthologs compared to the average number found in its relatives (Table 3-1). Contigs from the *T. caries* 517 genome assembly with significant homology (E-value < $1e^{-6}$) and the number of full length orthologs identified are shown in Table 3-2. For details on BLAST hits, see Table S6. Annotation information for predicted orthologs is shown in Table 3-3 and FGENESH (Solovyev et al. 2006) annotations for all predicted RNA silencing related genes are in RNA silencing_FGENESH_AA_SUPP.fasta and RNA silencing_FGENESH_CDS_SUPP.fasta in FASTA format including both amino acid and nucleotide sequences. Details on contig hits in the genome assembly are provided in Table S6. The match for Dicer CBQ71581 that was not annotated was ignored because it was too close to the end of a contig to be verified based on its conserved domain architecture. The conserved domain layout of each gene listed in Table 3-3 is depicted in Figure 3-2, Figure 3-3, Figure 3-4, and Figure 3-5. The conservation of domain structure remarked on previously and shown for *S. reilianum* orthologs in Fig. 3-1 is evident in these images.

Table 3-2. Numbers of tBLASTn hits on the genome and the respective numbers of RNA-I orthologs annotated.

Query Sequence	**Function**	**High Scoring Contigs (E-value < 1e-6)**	**Orthologs Annotated**
CBQ71581	Dicer	4	3
CBQ68774	Argonaute	6	6
CBQ71530	RdRP	1	1
CBQ73251	Exportin-1	1	1

Table 3-3. Putative RNA silencing-related genes predicted in *T. caries* 517 Genomic DNA.

Gene	**Contig**	**Length (nt/aa)**	**Exons**	**Complete ORF***	**Notes**
Ago1	75716	3084/1027	14	Y	
Ago2	76014	3009/1002	17	Y	
Ago3	76414	1338/445	8	Y	Tandem Duplicates
Ago4	76414	702/223	4	N	
Ago5	76036	1023/340	5	N	
Ago6.0	75596	426/141	2	Y	Identical: contigs overlap
Ago6.1	74923	426/141	2	Y	
Dicer1	37836-36863	4443/1480	20	Y	Resequence: spans overlapping contigs
Dicer2	75969	3533/1692	25	Y	
Dicer3	37419	954/318	5	N	
RdRP	76383	3831/1276	9	Y	
Exportin-1	75955	3234/1077	14	Y	

*Not all predicted RNA silencing gene orthologs were present as complete open reading frames (ORF); some were incomplete in the current genome assembly and are indicated in the table.

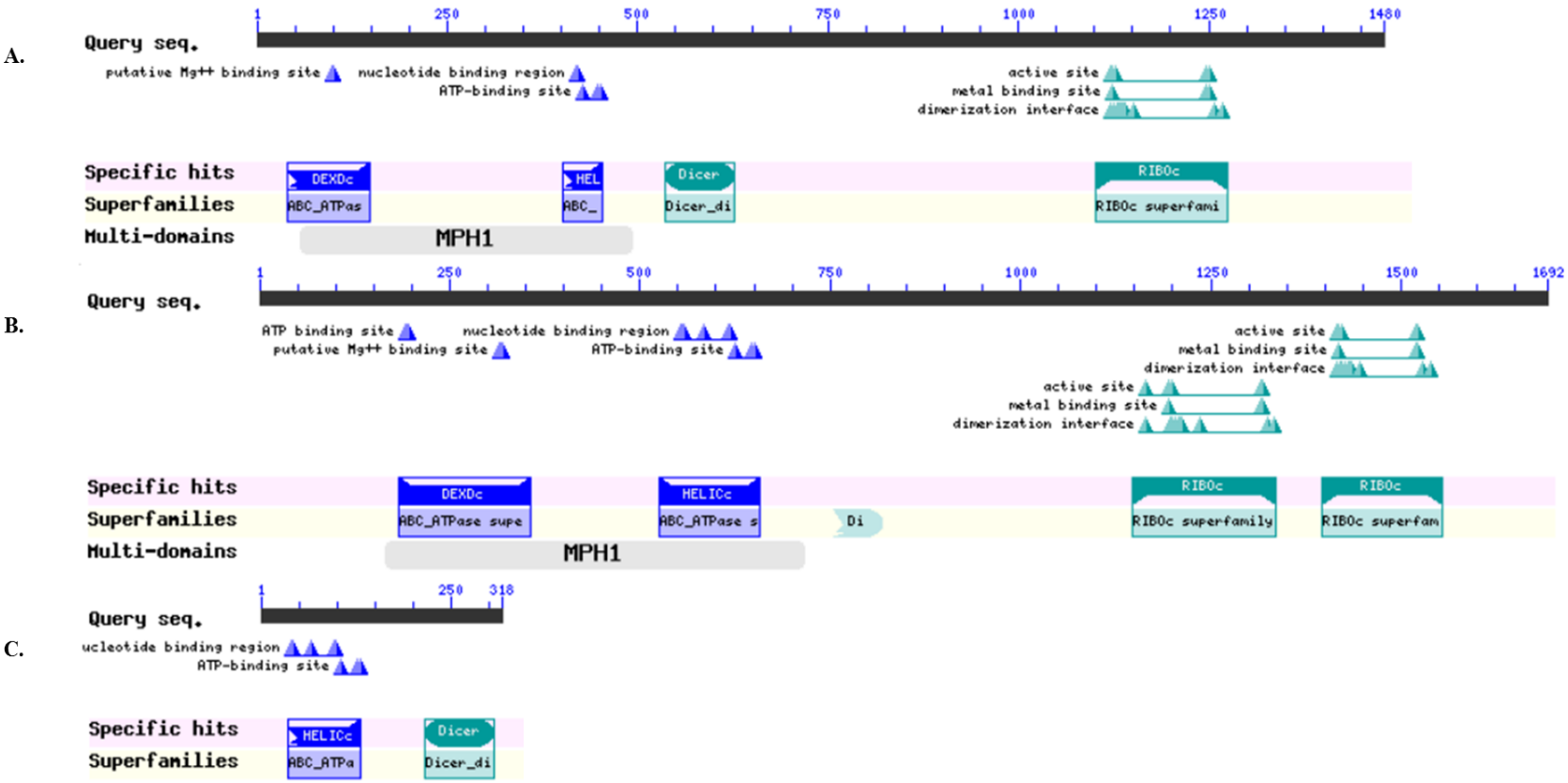

Figure 3-2. Conserved domain sketches of putative Dicer genes predicted in the *T. caries* 517 genome. A. Dicer1 B. Dicer2 C. Dicer3 (Table 3-3). Numbers refer to the length in amino-acid residues of the proteins. Compare to Fig. 3-1. Blue represents DEXDc and HELICc domains; teal represents Dicer dimerization domains and RIBOc domains. Graphics from the NCBI Conserved Domain Database (Marchler-Bauer et al. 2011).

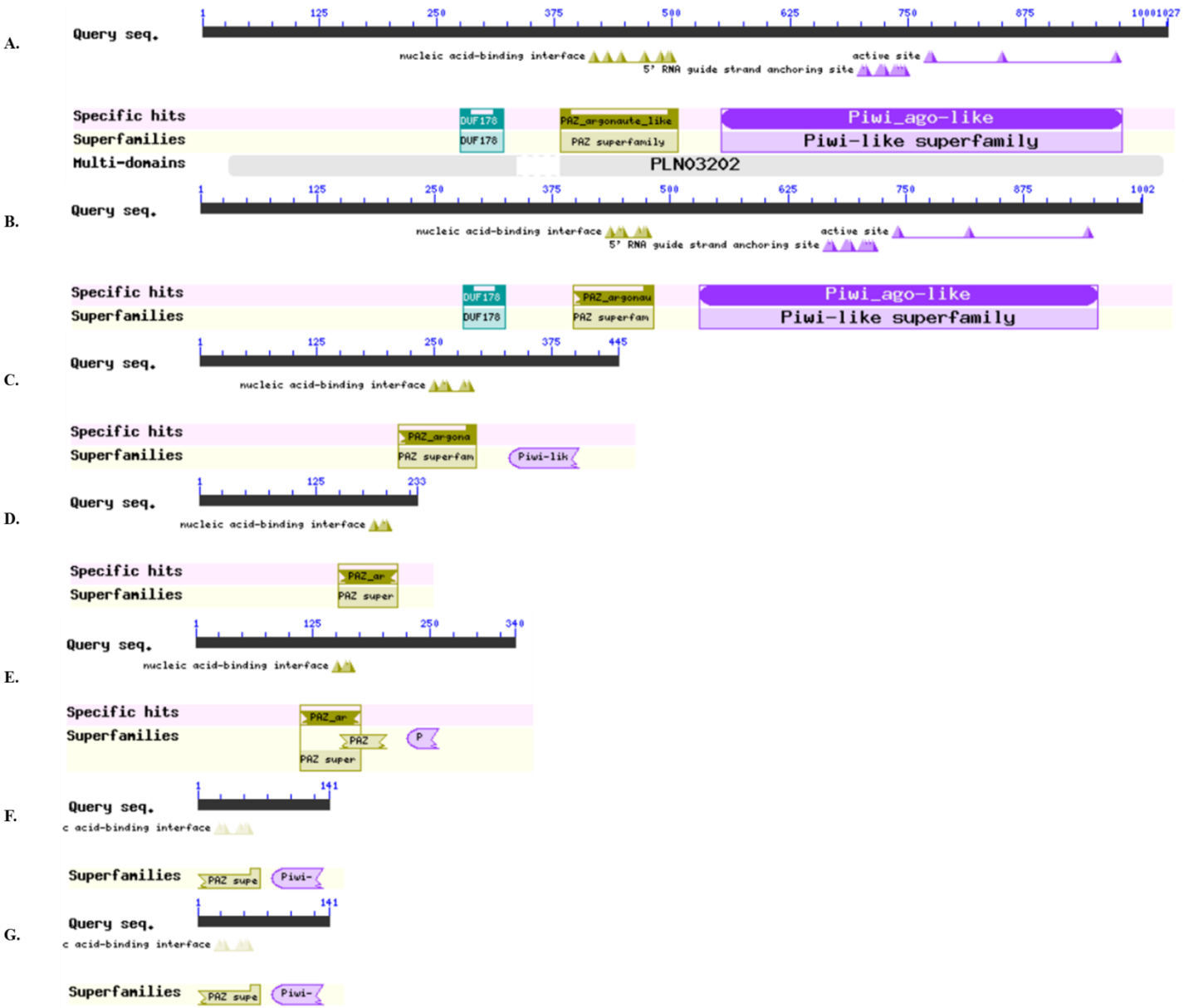
A.
Query seq.
1
125
250
375
500
625
750
875
1000
1027
nucleic acid-binding interface
active site
5' RNA guide strand anchoring site
Specific hits
Superfamilies
Multi-domains
DUF178
DUF178
PAZ_argonaute_like
PAZ superfamily
Piwi_ago-like
Piwi-like superfamily
PLN03202
B.
Query seq.
1
125
250
375
500
625
750
875
1002
nucleic acid-binding interface
active site
5' RNA guide strand anchoring site
Specific hits
Superfamilies
DUF178
DUF178
PAZ_argonau
PAZ superfam
Piwi_ago-like
Piwi-like superfamily
C.
Query seq.
1
125
250
375
445
nucleic acid-binding interface
Specific hits
Superfamilies
PAZ_argona
PAZ superfam
Piwi-lik
D.
Query seq.
1
125
233
nucleic acid-binding interface
Specific hits
Superfamilies
PAZ_ar
PAZ super
E.
Query seq.
1
125
250
340
nucleic acid-binding interface
Specific hits
Superfamilies
PAZ_ar
PAZ
P
PAZ super
F.
Query seq.
1
141
c acid-binding interface
Superfamilies
PAZ supe
Piwi-
G.
Query seq.
1
141
c acid-binding interface
Superfamilies
PAZ supe
Piwi-

Figure 3-3. Conserved domain sketches of putative Argonaute genes predicted in the *T. caries* 517 genome. A. Ago1 B. Ago2 C. Ago3, D. Ago4, E. Ago5, F. Ago6, G. Ago6.1 (Table 3-3). Numbers refer to the length in amino-acid residues of the proteins. Compare to Fig. 3-1. Gold represents the PAZ domain; purple represents the PIWI domain; teal represents DUF1785. Graphics from the NCBI Conserved Domain Database (Marchler-Bauer et al. 2011).

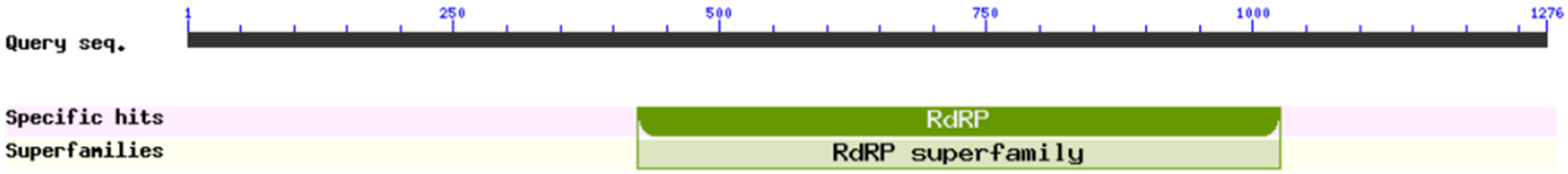

Figure 4. Conserved domain sketch of putative RdRP gene predicted in the *T. caries* 517 genome (Table 3-3). Numbers refer to the length in residues of the protein. Compare to Fig. 3-1. Graphics from the NCBI Conserved Domain Database (Marchler-Bauer et al. 2011).

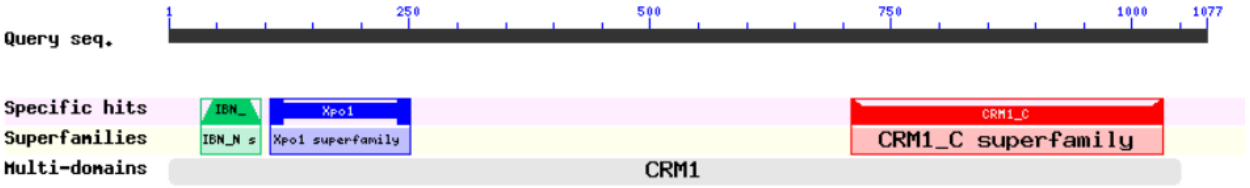

Figure 3-5. Conserved domain sketch of putative Exportin-1/CRM-1 gene predicted in the *T. caries* 517 genome (Table 3-3). Numbers refer to the length in amino-acid residues of the protein. Compare to Fig. 3-1. Graphics from the NCBI Conserved Domain Database (Marchler-Bauer et al. 2011)

Three ORFs encode putative Dicer peptides in *T. caries* 517; however only one is present in the current assembly as a complete ORF (Dicer 2). A second complete ORF (Dicer 1) was constructed from putatively overlapping contigs. There is some variation in the conserved domain structure of the Dicer peptides in the *T. caries* 517 genome assembly relative to the Dicer gene present in *S. reilianum*. None of the putative Dicer-encoding genes in the *T. caries* 517 assembly (Fig 3-2) have conserved domain architecture identical to that found in *S. reilianum*'s Dicer (Fig 3-1B). In *S. reilianum* Dicer encodes, from N- to C-terminus a DEXDc domain, HELICc domain, Dicer dimerization domain, two RIBOc domains and a zinc finger domain. Dicer 2 has identical conserved domain architecture to *S. reilianum* with the exception of the C-terminal zinc finger domain (Fig. 3-1). Dicer 1 contains one less C-terminal RIBOc domain and also lacks the zinc finger domain. Dicer 3 contains only the Dicer dimerization domain and a single HELICc domain and lacks all other conserved domains present in *S. reilianum*; however it is not a complete ORF in the current *T. caries* 517 assembly.

There is some variation in conserved domain structure among the Argonaute proteins predicted in the *T. caries* 517 genome (Fig. 3-3) relative to the architecture of the Argonaute gene found in *S. reilianum* (Fig. 3-1). Argonaute 3 and Argonaute 4 are not complete open reading frames and Argonaute 6 and 6.1 represent the same peptide. A reassembly of their contigs of origin revealed that they overlap and the matches identified in BLAST searches and subsequently annotated using FGENESH originated from the same sequence within the overlapping region. Moving from its N-terminus, *S. reilianum* contains a DUF1785 (conserved domain of unknown function), followed approximately 100 residues downstream by a PAZ_Argonaute_1 nucleic acid binding domain. Approximately100 residues following the PAZ domain is the larger Piwi

Argonaute-like domain, which includes RNA-binding and catalytic domains. DUF1785 is found in Ago 1 and Ago 2 approximately 100 residues towards the N-terminus of the PAZ domain. DUF1785 is frequently found in conjunction with the PIWI domain and in Argonaute proteins (Hall 2005). Ago1 and Ago2 are identical in conserved domain architecture to the Argonaute protein found in *S. reilianum*. While Ago4 and Ago5 in *T. caries* 517 differ in conserved domain architecture from *S. reilianum*'s Argonaute, they are incomplete ORFs in the current assembly. Ago3 and Ago6 are complete ORFs as predicted by FGENESH and thus represent marked departures from the conserved domains present in *S. reilianum*'s Argonaute, possessing only the PIWI and PAZ domains. Additionally, they are significantly shorter in length than that found in *S. reilianum*. A neighbor joining phylogenetic tree comparing complete *T. caries* 517 Argonaute predictions is shown in Figure 3-6. Unlike the Dicer and Argonaute genes, both Exportin-1/CRM-1 and RdRP exhibit identical conserved domain structure to that found in *S. reilianum*.

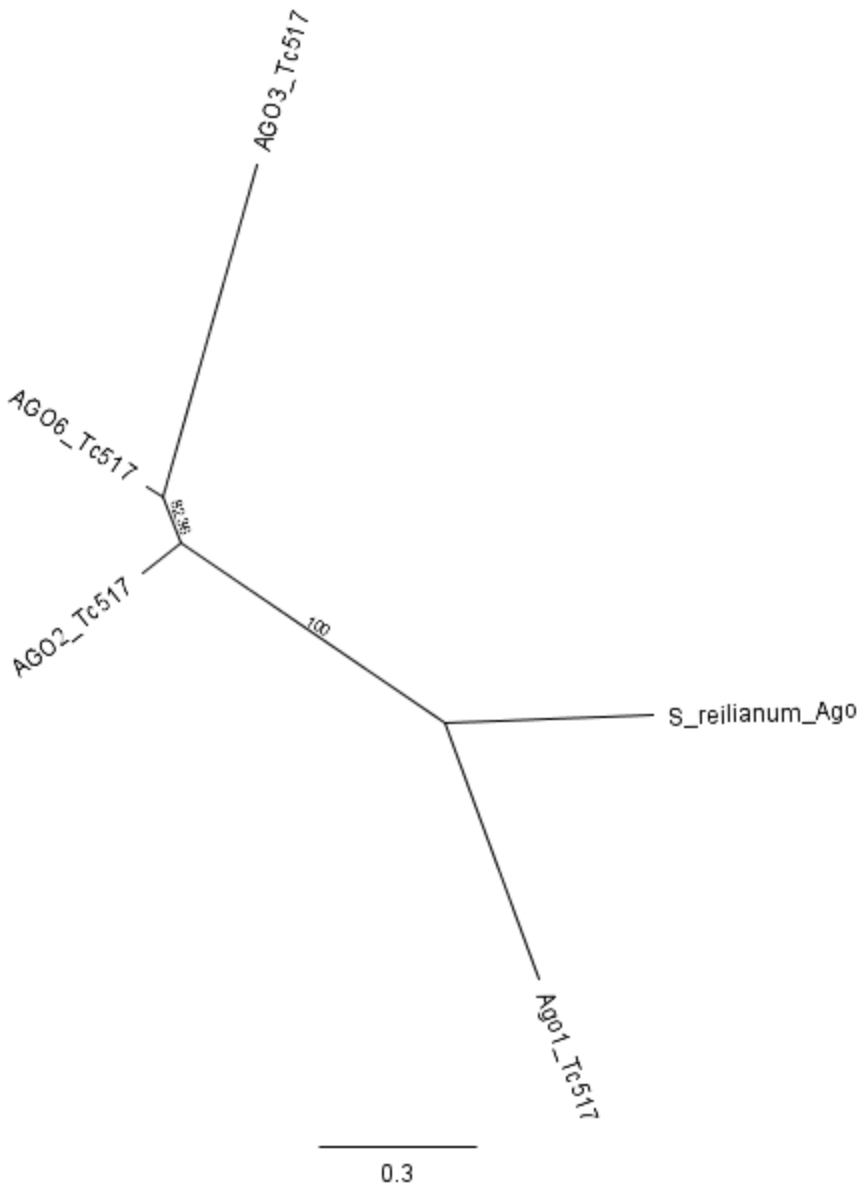

Figure 3-6. NNI tree of *T. caries* (Tc517) complete predicted Argonaute (Ago) peptides. *S. reilianum*'s Ago gene is included as an outgroup. Distances are in substitutions per site. Branch labels represent bootstrap support out of 10,000 replicates.

Phylogenetic Comparisons with RNA silencing-Associated Genes in Other Fungi

Figures 3-7, 3-8, 3-9 and, 3-10 show maximum likelihood (A) and Bayesian Markov Chain Monte Carlo (B) phylogenetic analyses of the putative RNA silencing associated genes in the *T. caries* 517 genome (Argonaute, Dicer, RdRP and Exportin-1/CRM-1, respectively) in comparison with those found in a variety of basidiomycete fungi representing all three major lineages of Basidiomycota: Ustilaginomycotina, Pucciniomycotina and Agaricomycotina. Two Argonaute orthologs (Ago1 and Ago2) and one Dicer ortholog (Dicer2) were selected for inclusion in the analyses. These were chosen because their lengths in amino acids were representative of other Basidiomycete Argonaute and Dicer orthologs identified in the NCBI non-redundant protein sequence database and they are present as full ORFs in the current *T. caries* 517 assembly. The two methods produce identical topologies for Argonaute and Dicer; however this is not the case for the remaining comparisons. Well supported clades in the Argonaute analyses include Ustilaginomycotina, the rust fungi and *Mixia osmundae*, the red yeasts (Pucciniomycotina), *Cryptococcus neoformans* and *Cryptococcus gattii*, and *Schizophyllum commune* and *Laccaria bicolor*. Monophyly is only exhibited by Ustilaginomycotina for Argonaute. Agaricomycotina and Pucciniomycotina are polyphyletic in this analysis. For Dicer, Ustilaginomycotina and Pucciniomycotina are monophyletic and well supported in both analyses (100% bootstrap and posterior probability support). Some of the taxa from Pucciniomycotina that contributed to its polyphyly in the Argonaute tree were not included, for example the red yeasts.

Maximum likelihood and Bayesian trees for RdRP differ only in the branching order of the clade containing Pucciniomycotina. In the Bayesian analysis, Pucciniomycotina forms a well-supported sister clade to the mushroom forming fungi in Agaricomycotina; however in the

maximum likelihood analysis, the red yeast *Rhodosporidium toruloides* is paraphyletic and basal to both clades. Bootstrap support for the sister clade relationship between the remaining members of Pucciniomycotina and the mushroom forming Agaricomycetes is weak. Other clades are identical in both analyses. *C. neoformans* and *C. gattii* are identified as basal to Ustilaginomycotina, Pucciniomycotina and remaining Agaricomycotina with poor bootstrap support and 94% posterior probability. Ustilaginomycotina forms a well-supported clade with the exception of *T. caries*, which is identified as basal to all other basidiomycete lineages in both methods. Monophyly of RdRP per accepted subphyla distinctions in Basidiomycota is not apparent in either analysis.

Maximum likelihood and Bayesian phylogenetic analyses of Exportin-1/CRM-1 differ only in the branching order for Pucciniomycotina. In the maximum likelihood analysis, *M. osmundae* and *R. toruloides* form a basal clade with poor bootstrap support to the cereal rusts; in the Bayesian analysis they are both paraphyletic to the cereal rust group with 100% posterior probability and *R. toruloides* occupying the basal node. Monophyly of the CRM-1 proteins is clear and well supported in both analyses along accepted subphyla patterns for Agaricomycotina and Ustilaginomycotina.

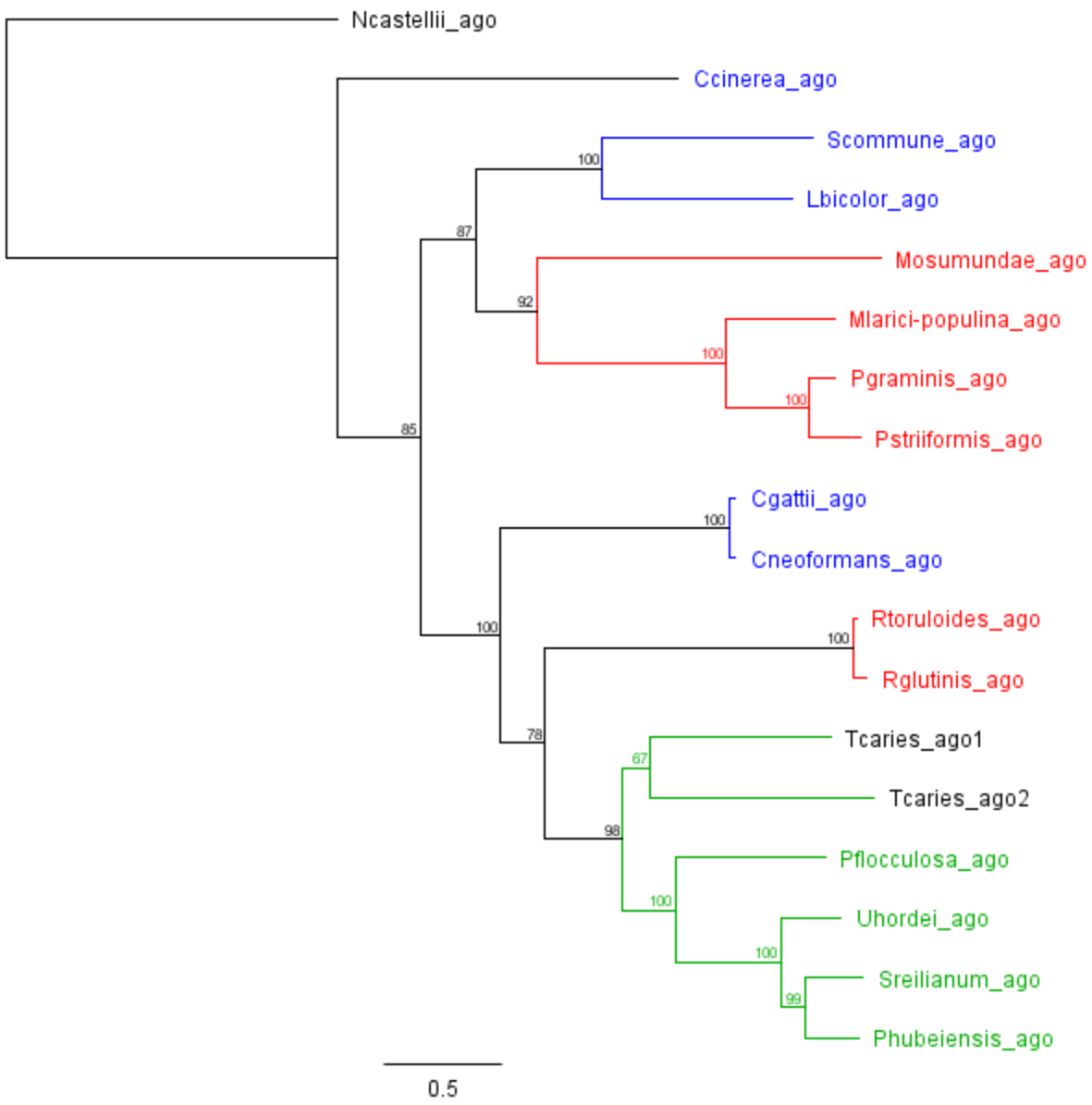
Ncastellii_ago
Ccinerea_ago
Scommune_ago
Lbicolor_ago
Mosumundae_ago
Mlarici-populina_ago
Pgraminis_ago
Pstriiformis_ago
Cgattii_ago
Cneoformans_ago
Rtoruloides_ago
Rglutinis_ago
Tcaries_ago1
Tcaries_ago2
Pflocculosa_ago
Uhordei_ago
Sreilianum_ago
Phubeiensis_ago
100
87
92
100
100
85
100
100
100
78
67
98
100
100
99
0.5

Figure 3-7 A. 100 bootstrap-replicate maximum likelihood phylogeny (LG+I+G+F model) of Argonaute proteins from representative taxa in Basidiomycota including putative Argonaute orthologs identified in the *T. caries* 517 genome (black taxon labels). The ascomycetous yeast *Naumovozyma castellii* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and node labels denote bootstrap support. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

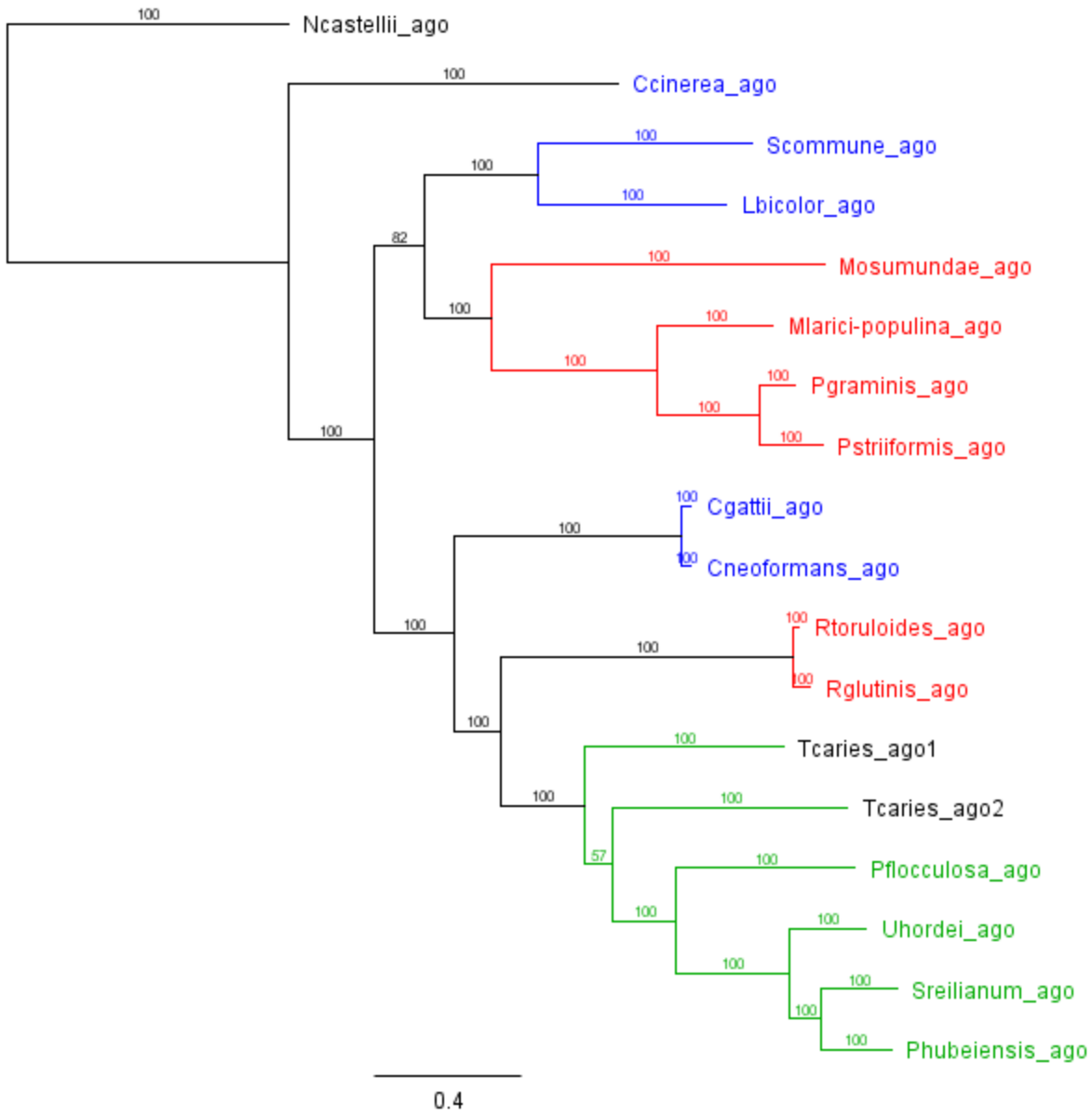
Ncastellii_ago
Ccinerea_ago
Scommune_ago
Lbicolor_ago
Mosumundae_ago
Mlarici-populina_ago
Pgraminis_ago
Pstriiformis_ago
Cgattii_ago
Cneoformans_ago
Rtoruloides_ago
Rglutinis_ago
Tcaries_ago1
Tcaries_ago2
Pflocculosa_ago
Uhordei_ago
Sreilianum_ago
Phubeiensis_ago
0.4

Figure 3-7 B. Bayesian Markov Chain Monte Carlo phylogenetic analysis (LG+G model) of Argonaute proteins from representative taxa in Basidiomycota including putative Argonaute orthologs identified in the *T. caries* 517 genome (black taxon labels). The ascomycetous yeast *Naumovozyma castellii* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and branch labels denote posterior probabilities. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

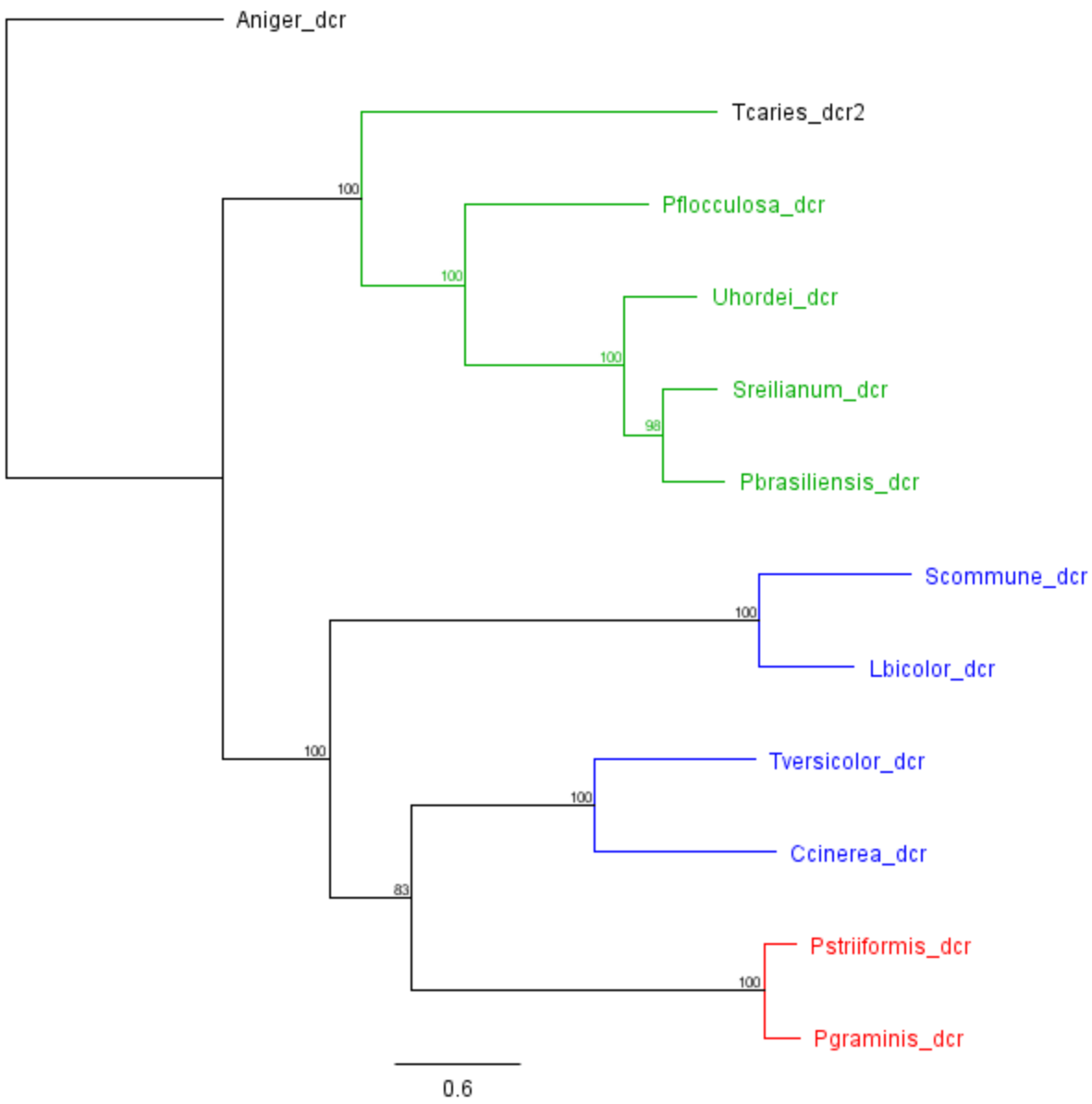
Aniger_dcr
Tcaries_dcr2
Pflocculosa_dcr
Uhordei_dcr
Sreilianum_dcr
Pbrasiliensis_dcr
Scommune_dcr
Lbicolor_dcr
Tversicolor_dcr
Ccinerea_dcr
Pstriiformis_dcr
Pgraminis_dcr
100
100
100
98
100
100
100
83
100
0.6

Figure 3-8 A. 100 bootstrap-replicate maximum likelihood phylogeny (WAG+I+G+F model) of Dicer proteins from representative taxa in Basidiomycota including a putative Dicer ortholog identified in the *T. caries* 517 genome (black taxon label). The filamentous ascomycete *Aspergillus niger* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and node labels denote bootstrap support. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

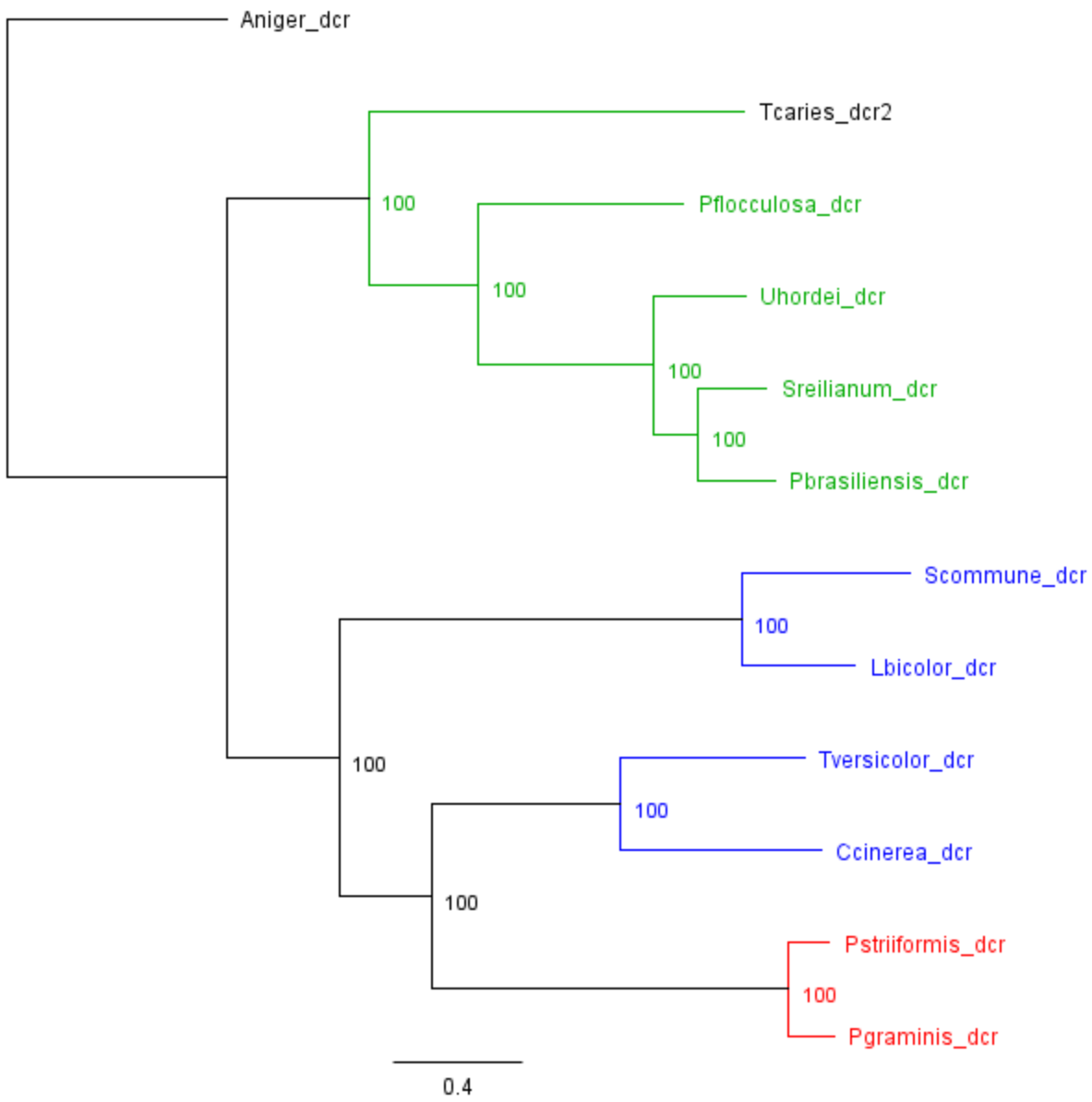
Aniger_dcr
Tcaries_dcr2
100
Pflocculosa_dcr
100
Uhordei_dcr
100
Sreilianum_dcr
100
Pbrasiliensis_dcr
Scommune_dcr
100
Lbicolor_dcr
100
Tversicolor_dcr
100
Ccinerea_dcr
100
Pstriiformis_dcr
100
Pgraminis_dcr
0.4

Figure 3-8 B. Bayesian Markov Chain Monte Carlo phylogenetic analysis (WAG+G model) of Dicer proteins from representative taxa in Basidiomycota including a putative Dicer ortholog identified in the *T. caries* 517 genome (black taxon label). The filamentous ascomycete *Aspergillus niger* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and branch labels denote posterior probabilities. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

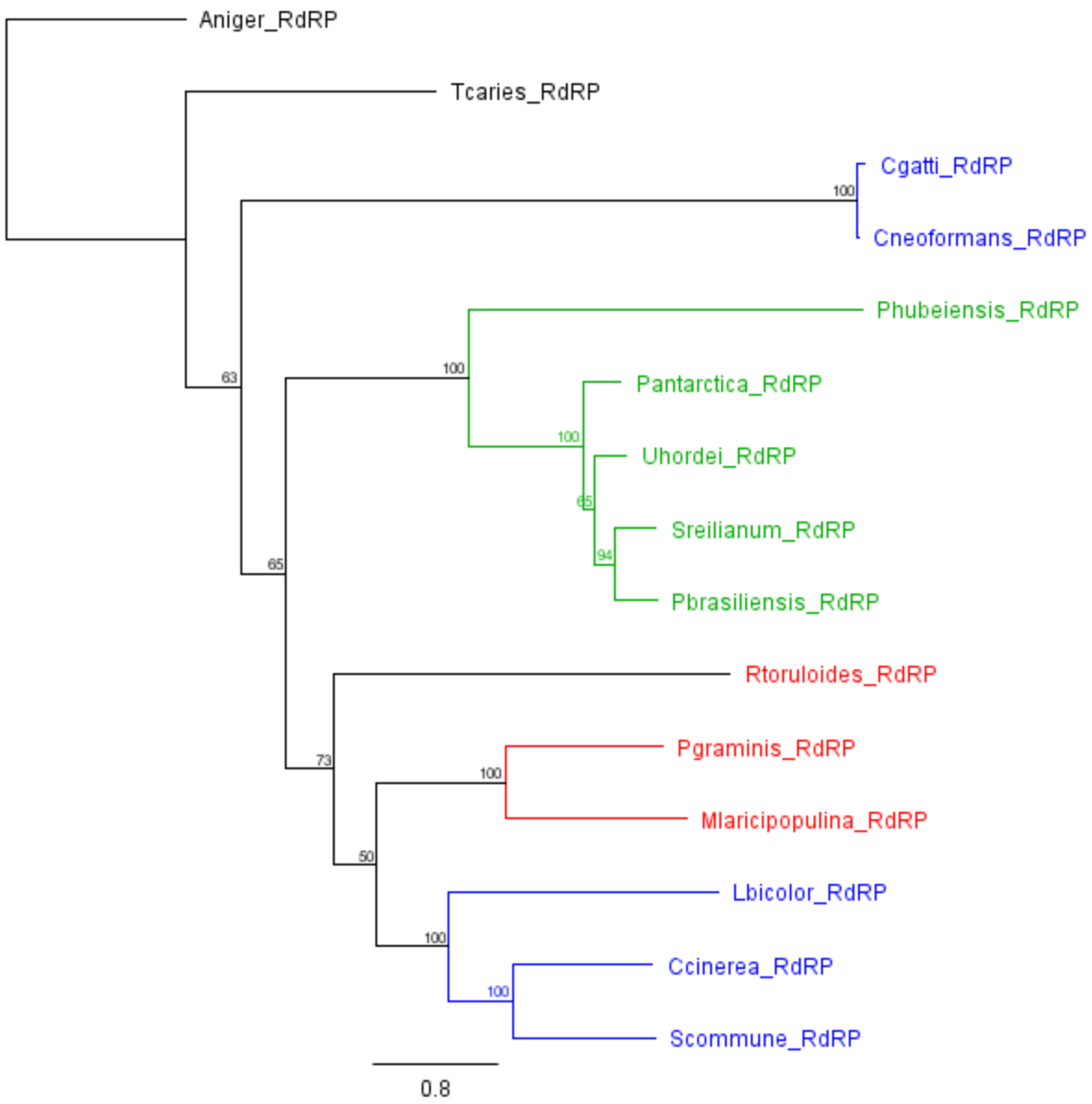
Aniger_RdRP
Tcaries_RdRP
Cgatti_RdRP
Cneoformans_RdRP
Phubeiensis_RdRP
Pantarctica_RdRP
Uhordei_RdRP
Sreilianum_RdRP
Pbrasiliensis_RdRP
Rtoruloides_RdRP
Pgraminis_RdRP
Mlaricipopulina_RdRP
Lbicolor_RdRP
Ccinerea_RdRP
Scommune_RdRP
100
63
100
100
65
94
65
73
100
50
100
100
0.8

Figure 3-9 A. 100 bootstrap-replicate maximum likelihood phylogeny (LG+I+G+F model) of RdRP proteins from representative taxa in Basidiomycota including the putative RdRP ortholog identified in the *T. caries* 517 genome (black taxon label). The filamentous ascomycete *Aspergillus niger* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and node labels denote bootstrap support. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

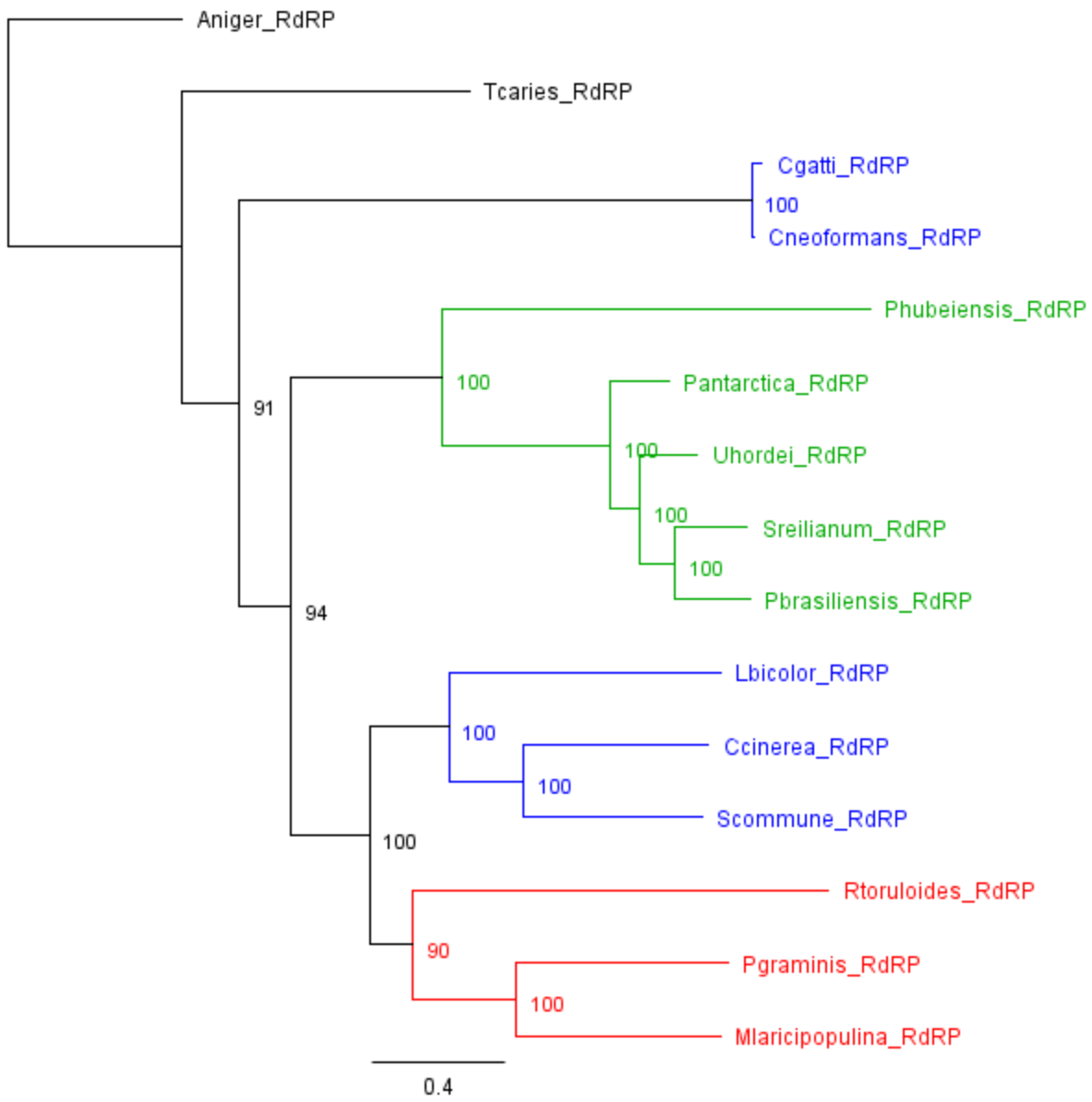
Aniger_RdRP
Tcaries_RdRP
Cgatti_RdRP
100
Cneoformans_RdRP
Phubeiensis_RdRP
100
Pantarctica_RdRP
91
100
Uhordei_RdRP
100
Sreilianum_RdRP
100
Pbrasiliensis_RdRP
94
Lbicolor_RdRP
100
Ccinerea_RdRP
100
Scommune_RdRP
100
Rtoruloides_RdRP
90
Pgraminis_RdRP
100
Mlaricipopulina_RdRP
0.4

Figure 3-9. B. Bayesian Markov Chain Monte Carlo phylogenetic analysis (LG+G model) of RdRP proteins from representative taxa in Basidiomycota including the putative RdRP ortholog identified in the *T. caries* 517 genome (black taxon label). The filamentous ascomycete *Aspergillus niger* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and branch labels denote posterior probabilities. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

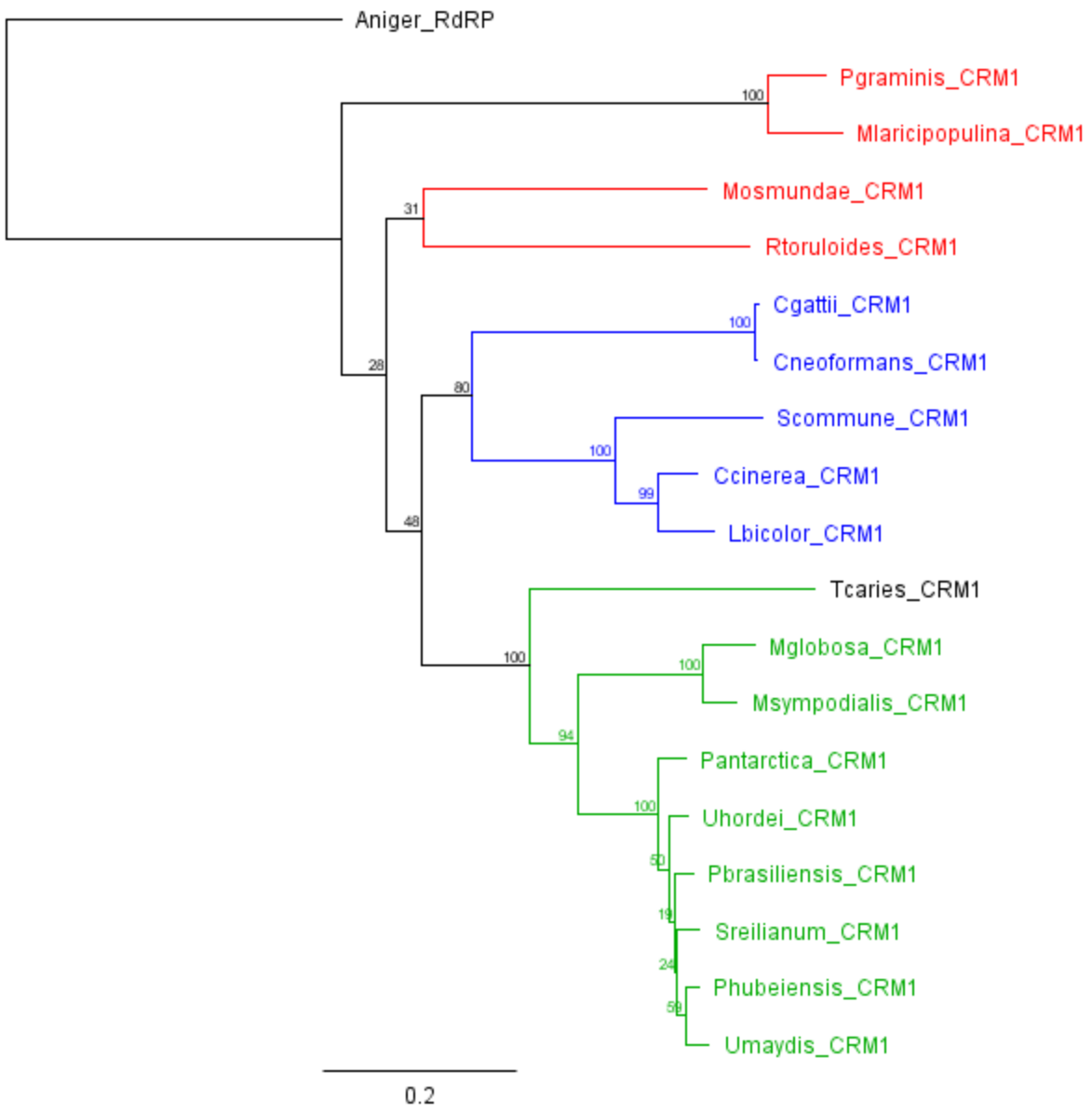
Aniger_RdRP
Pgraminis_CRM1
Mlaricipopulina_CRM1
Mosmundae_CRM1
Rtoruloides_CRM1
Cgattii_CRM1
Cneoformans_CRM1
Scommune_CRM1
Ccinerea_CRM1
Lbicolor_CRM1
Tcaries_CRM1
Mglobosa_CRM1
Msympodialis_CRM1
Pantarctica_CRM1
Uhordei_CRM1
Pbrasiliensis_CRM1
Sreilianum_CRM1
Phubeiensis_CRM1
Umaydis_CRM1
100
31
28
80
100
99
48
100
100
94
100
50
19
24
59
0.2

Figure 3-10 A. 100 bootstrap-replicate maximum likelihood phylogeny (LG+I+G model) of Exportin-1/CRM-1 proteins from representative taxa in Basidiomycota including the putative Exportin-1/CRM-1 ortholog identified in the *T. caries* 517 genome (black taxon label). The filamentous ascomycete *Aspergillus niger* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and node labels denote bootstrap support. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

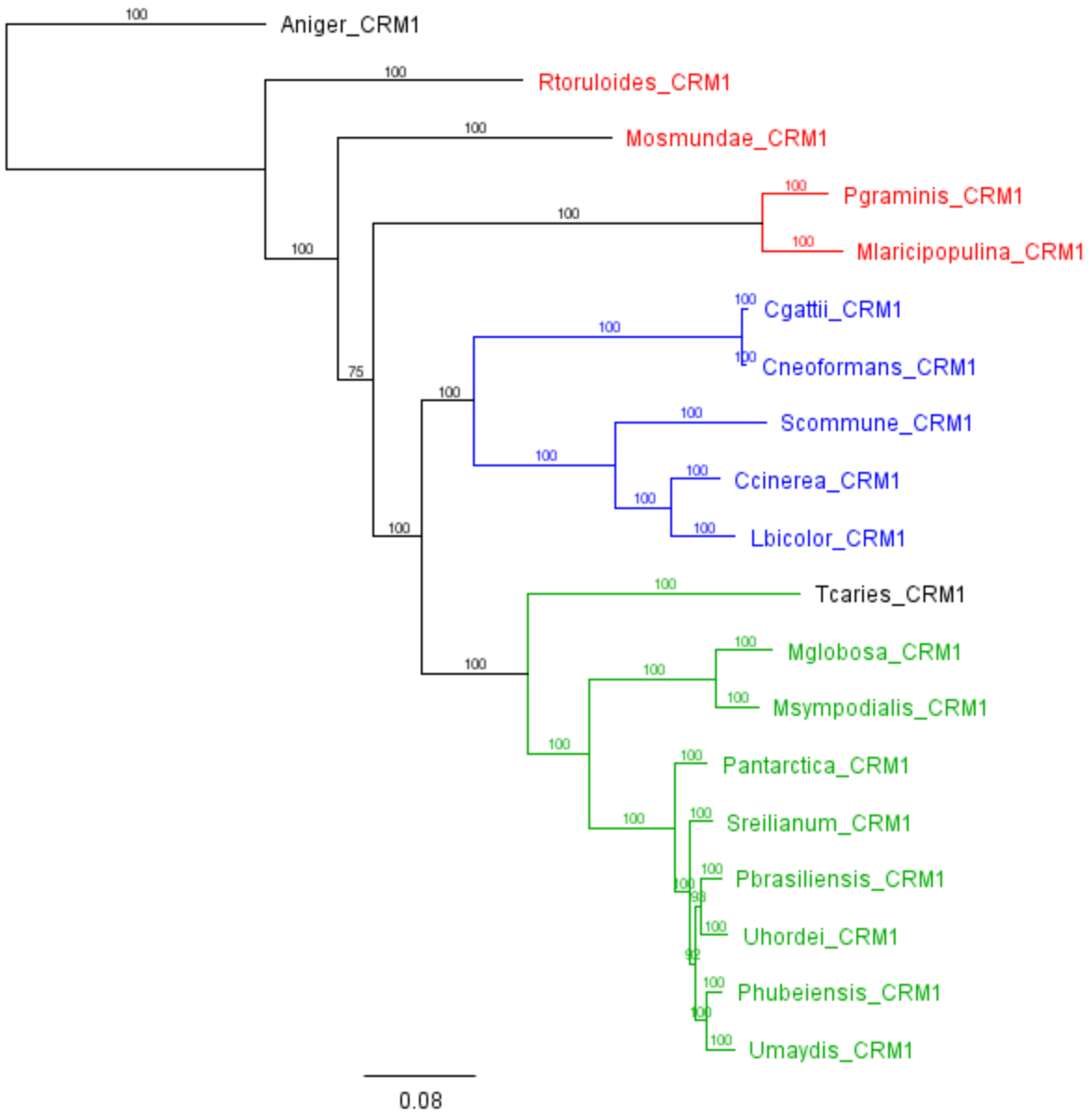
Aniger_CRM1
Rtoruloides_CRM1
Mosmundae_CRM1
Pgraminis_CRM1
Mlaricipopulina_CRM1
Cgattii_CRM1
Cneoformans_CRM1
Scommune_CRM1
Ccinerea_CRM1
Lbicolor_CRM1
Tcaries_CRM1
Mglobosa_CRM1
Msympodialis_CRM1
Pantarctica_CRM1
Sreilianum_CRM1
Pbrasiliensis_CRM1
Uhordei_CRM1
Phubeiensis_CRM1
Umaydis_CRM1
0.08

Figure 3-10 B. Bayesian Markov Chain Monte Carlo phylogenetic analysis (LG+G model) of Exportin-1/CRM-1 proteins from representative taxa in Basidiomycota including the putative Exportin-1/CRM-1 ortholog identified in the *T. caries* 517 genome (black taxon label). The filamentous ascomycete *Aspergillus niger* (black taxon label) is included as an outgroup. Green branches and taxa labels represent Ustilaginomycotina, red, Pucciniomycotina and blue, Agaricomycotina. Distances are in substitutions per site and branch labels denote posterior probabilities. GenBank accession numbers for non-*Tilletia* taxa are shown in Table S5.

DISCUSSION

This evidence suggests that there has been an expansion of the Argonaute and Dicer protein families in *T. caries*, as reported by Nakayashiki et al. (2005) for other fungi. Conversely, there are a lower number of RdRP and Exportin-1/CRM-1 genes present than in many other fungi (Table 3-1). This proliferation of Dicer and Argonaute genes may be associated with regulation of developmental processes as in other eukaryotes (Wilson and Doudna 2013; Nunes et al. 2011).

The wide variation in structure and length of proteins such as Dicer and Argonaute within *T. caries* and across species boundaries make them difficult to compare phylogenetically. Complete annotations need to be obtained for all of the putative RNA silencing orthologs identified here before an exhaustive phylogenetic analysis can be carried out. Additionally, orthologs to the shorter RNA silencing proteins identified in the *T. caries* genome (if any) need to be identified in other taxa to facilitate comparison.

The neighbor joining cladogram of Argonaute sequences shown in Fig. 3-6 demonstrates the wide variation among the complete putative Argonaute proteins in this version of the *T. caries* 517 genome assembly; this suggests nuanced roles for these various Argonaute proteins despite their conserved RNA silencing-related functional domains. Note that Tc517_Ago1 has greater similarity to the *S. reilianum* outgroup than the other Argonaute sequences. This suggests that this particular Argonaute ortholog could fulfill a similar biological function to that of Argonaute in *S. reilianum*.

Varying conserved domains are present in the Argonaute and Dicer orthologs when compared to *S. reilianum*. This supports the notion that these genes play diversified roles in the cell. Many of the previously annotated Argonaute genes from other taxa contain DUF1785. This domain is present in only two of the Argonaute orthologs. Given that Argonaute is known to associate with silencing-related protein complexes such as RISC and RIST (Verdel et al. 2004) it is possible that this DUF plays a role in the protein-protein interactions necessary for complex formation. This variation in conserved domain structure suggests different roles for the genes in the set of Argonaute orthologs.

Polyphyly is common for orthologs from the same subphyla. None of the subphyla is monophyletic in every tree, though Ustilaginomycotina is only polyphyletic in the RdRP tree. This suggests that the genes involved in RNA silencing are not evolving at the same rates as genes traditionally included in fungal phylogenies (Spatafora 2005). This supports the notion that RNA silencing- related genes have diversified greatly in structure and function amongst the fungi (Nakayashiki et al. 2005) and provides further credence to the presence of multiple orthologs in *T. caries* 517.

Drinnenberg et al. (2011) showed that the complete loss of RNA silencing systems in species such as *Saccharomyces cerevisiae*, *Malassezia globosa*, *U. maydis* and others is associated with the presence of the dsRNA Killer Virus, which encodes a toxin and immunity system, killing uninfected fungi in the surrounding area (Drinnenberg et al. 2011). Because the species that have lost RNA silencing are infected with the killer virus, strong selective pressure favors jettison of RNA silencing machinery. Without RNA silencing the fungus no

longer degrades Killer transcripts, allowing the toxin and immunity system's expression thus providing a selective advantage. Most fungi which have lost RNA silencing, including *U. maydis* and *S. cerevisiae*, demonstrate evidence of a killer infection (Drinnenberg et al. 2011). Killer virus infections have not been reported in *T. caries*; however further RNA sequencing data is necessary to verify the absence of killer virus infection and if Drinnenberg et al.'s (2011) hypothesis is correct, it should not be present.

A number of assembly problems were encountered in the course of annotation, for example ORFs duplicated on overlapping contigs or ORFs spanning gaps between contigs in the assembly. Additional PacBio sequencing data should be obtained to reduce the number of gaps in the assembly. Each gene should also be resequenced with Sanger sequencing (Primer_SUPP.xlsx). Transcriptome data recently available for this and an additional strain of *T. caries* will aid in this process. Proof-of-function assays with a heterologous system-approach should be carried out using *U. maydis*. *U. maydis* is well suited for studies of RNA silencing because it has endogenously lost its own RNA silencing machinery (Laurie et al. 2008) and is a well characterized model system (Feldbrügge et al. 2013). Introduction of foreign RNA silencing machinery would create effects that would be readily measurable relative to the wild type, for example loss of killer virus infection (Drinnenberg et al. 2011) or suppression of ectopic GFP expression with an introduced GFP transgene.

CHAPTER THREE REFERENCES

Altschul SF, Gish W, Miller W, Myers EW and, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403-410.

Bai Y, Lu H and Machida CA. 2006. CRM 1- mediated degradation and agonist-induced down-regulation of β-adrenergic receptor mRNAs. Molecular Cell Research 1763(1): 1076-1089.

Bernstein E, Caudy AA, Hammond SM and, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. Nature 409: 363-366.

Blaszczyk J, Tropea JE, Bubunenko M, Routzahn KM, Waugh DS, Court DL and, Ji X. 2001. crystallographic and modeling studies of RNase III suggest a mechanism for double-s tranded RNA cleavage. Structure 9(12): 1225-1236.

Carris LM and Gray PM. 1994. The ability of *Tilletia fusca* to hybridize with the wheat bunt species under axenic conditions. Mycologia 86(1): 157-163.

Cogoni C and Macino G. 1997. Isolation of quelling-defective (qde) mutants impaired in posttranscriptional transgene-induced gene silencing in *Neurospora crassa*. Proceedings of the National Academy of Sciences of the United States of America 94: 10233-10238.

Dalmay T, Hamilton A, Rudd S, Angell S and, Baulcombe DC. 2000. An RNA-Dependent RNA Polymerase gene in *Arabidopsis* is required for posttranscriptional gene silencing mediated by a transgene but not by a virus. Cell 101(5): 543-554.

Darriba D, Taboada GL, Doalla R and, Posada D. 2011. ProtTest 3: Fast selection of best-fit models of protein evolution. Bioinformatics 27: 1164-1165.

Dlakić M. 2006. DUF283 domain of Dicer proteins has a double-stranded RNA-binding fold. Bioinformatics 22(22): 2711-2714.

Drinnenberg IA, Fink GR and, Bartel DP. 2011. Compatibility with Killer explains the rise of RNA silencing-deficient fungi. Science 333: 152.

Feldbrügge Kellner R and, Schipper K. 2013. The biotechnological use and potential of plant pathogenic smut fungi. Applied Microbiology and Biotechnology 97: 3253-3265.

Fire A, Xu S, Montgomery MK, Kostas SA and, Driver SE. 1998. potent and specific genetic interference by double-stranded RNA in *C. elegans*. Nature 391(6669): 806-811.

Fischer GW and Holton CS. 1957. Biology and Control of the Smut Fungi. New York, NY: Ronald Press Co. 622 p.

Guindon S and Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large scale phylogenies by maximum likelihood. Systematic Biology. 52. 696-704.

Ji-Joon S and Joshua-Tor L. 2006. Argonaute and RNA - getting into the groove. Current Opinion in Structural Biology 16(1): 5-11.

Hall, TMT. 2005. Structure and function of Argonaute proteins. Structure 13(10): 1403-1408.

Huelsenbeck JP and Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17: 754-755.

Kämper J, Kahmann R, Bölker M, Ma L, Brefort T, Saville T, Banuette F, Kronstad JW, Gold SE, Müller O and, others. 2004. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature Letters 444(2): 97-101.

Kim JL, Morgenstern KA, Griffith JP, Dwyer MD, Thomson JA, Murchko MA, Lin C and, Caron PR. 1998. Hepatitis C virus NS3 RNA helicase domain with a bound oligonucleotide: The crystal structure provides insights into the mode of unwinding. Structure 6(1): 89-100.

Laurie JD, Linning R and, Bakkeren G. Hallmarks of RNA silencing are found in the smut fungus *Ustilago hordei* but not in its close relative *Ustilago maydis*. Current Genetics 53: 49-58.

Lee T, Cho I, Peace C, Jung S, Zheng P and, Main D. 2011. Development approach and architecture of GenSAS: the Genome Sequence Annotation Server. http://www.lidi.info.unlp.edu.ar/WorldComp2011-Mirror/BIC5013.pdf

Ma JB, Ye K and, Patel DJ. 2004. Structural basis for overhang-specific small interfering RNA recognition by the PAZ domain. Nature 429(6989): 318-322.

Marcheler-Bauer A, Lu S, Anderson JB, Chitsaz F, debryshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC Gonzales NR and, others. 2011. CDD: A conserved domain database for the functional annotation of proteins. Nucleic Acids Research 39(Database Issue) D225-229.

Millership JJ, Cai X and, Zhu G. 2004. Functional characterization of replication protein A2 (RPA2) from *Cryptosporidium parvum*. Microbiology 150(Pt5): 1197-11205.

Nakayashiki H, Kadotani N and, Mayama S. 2005. Evolution and diversification of RNA silencing proteins in fungi. Journal of Molecular Evolution 63: 127-135.

Napoli C, Lemieux C and, Jorgenson R. 1990. Introduction of a chimeric chalcone synthase gene into petunia results in reversible co-suppression of homologous genes in trans. The Plant Cell 2(4): 279-289.

Nelson DL and Cox MM. 2008. Lehninger Principles of Biochemistry 5th Ed. New York, NY: W.H. Freeman and Company. 1158 p.

Ngô H, Tschudi C, Gull K and, Ullu E. 1998. Double-stranded RNA induces mRNA degradation in *Trypanosoma brucei*. Proceedings of the National Academy of Sciences of the United States of America 95(25): 14678-14692.

Nunes CC and Dean RA. 2012. Host-induced gene silencing: A tool for understanding fungal host interaction and for developing novel disease control strategies. Molecular Plant Pathology 13(5): 519-529.

Nunes CC, Sailsbery JK and Dean RA. 2011. Characterization and application of small RNAs and RNA silencing mechanisms in fungi. *Fungal Biology Reviews* 25(4): 172-180.

Petosa C, Schoehn G, Askjaer P, Bauer U, Moulin M, Steuerwald U, Soler-López M, Baudin F, Mattaj IW and, Müller CW. 2004. Architecture of CRM1/Exportin1 suggests how cooperativity is achieved during formation of a nuclear export complex. Molecular Cell 16(5): 761-765.

Romano N and Macino G. Quelling: Transient inactivation of gene expression in *Neurospora crassa* by transformation with homologous sequences. Molecular Microbiology 6(22): 3343-3353.

Ronquist F and Huelsenbeck JP. 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19: 1572-1574.

Rozen S and Skaletsky HJ. 2000. Primer3 on the WWW for general users and for biologist programmers. In: Misener S and Krawetz SA eds. Bioinformatics Methods and Protocols: Methods in Molecular Biology. Totowa, NJ: Humana Press Inc. 500 p. Source code available at http://fokker.wi.mit.edu/primer3/

Shirawski J, Mannhaupt G, Munch K, Brefort T, Schipper K, Doehlemann G, Di Stasio M, Rossel N and, others. 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330(6010): 1546-1548.

Shiu PK, Raju NB, Zickler D and, Metzenberg RL. 2001. Meiotic silencing by unpaired DNA. Cell 107(7): 915-916.

Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, Lopez R, McWilliam H, Remmert M, Söding J and, others. 2011. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Molecular Systems Biology 7: 539.

Solovyev V, Kosarev P, Seledsov I and, Vorobyev D. 2006. Automatic annotation of eukaryotic genes, pseudogenes and promoters. Genome Biology 7, Suppl 1: 10.1-10.12

Spatafora JW. 2005. Assembling the fungal tree of life (AFTOL). Mycological Research 109: 755-756.

Theis K, Chen PJ, Skorvaga M, Van Houten B and, Kisker C. 1999. Crystal structure of UvrB, a DNA helicase adapted for nucleotide excision repair. EMBO J 18(24): 6899-6907.

Tomari Y and Zamore PD. 2005. Perspective: Machines for RNA silencing. Genes & Development 19: 517-529.

Ullman KS, Powers MA and, Forbes DJ. 1997. Nuclear export receptors: From Importin to Exportin. Cell 90(6): 967-970.

Van der krol, AR. 1990. Flavonoid genes in petunia: Addition of a limited number of gene copies may lead to suppression of gene expression. The Plant Cell Online 2(4): 291- 299.

Verdel A, Jia S, Gerber S, Sugiyama T, Gygi S, Grewal SIS and, Moazed D. 2004. RNA silencing- mediated targeting of heterochromatin by the RITS complex. Science 303(5658): 672-676.

Vetter IR, Arndt A, Kutay U, Görlich D and, Wittinghofer A. 1999. Structural view of the Ran-Importin Beta interaction at 2.3 Å resolution. Cell 97(5): 636-646.

Wilson RC and Doudna JA. 2013. Molecular mechanisms of RNA interference. Annual Review of Biophysics 42: 217-239.

Yi R, Qin Y Macara IG and, Cullen BR. 2003. Exportin-5 mediates the nuclear export of pre-microRNAs and short hairpin RNAs. Genes & Development 17(24): 3011-3016.

SUPPLEMENTAL MATERIALARY

SUPPLEMENTARY TABLES

Table S1. Strain origins

Species	Herbarium	Accession	Date collected	Collector	Origin
Tilletia caries	WSU Herbaria	WSP72095	06-16-2007	X. Chen	Pullman, WA, USA
Tilletia caries	AAFC National Mycological Herbarium	DAOM238032	1996	B. Goates	ID, USA
Tilletia contraversa	AAFC National Mycological Herbarium	DAOM236426	1998	Unk.	ON, CA
Tilletia indica	AAFC National Mycological Herbarium	DAOM23641R	1996	Unk.	Pakistan
Tilletia walkeri	AAFC National Mycological Herbarium	DAOM236422	1997	Unk.	OR, USA

Table S2. Genome assembly information

Strain	Sequenced by	# contigs	N50	# Bases	Longest contig	L50 contig #
Tilletia caries 517	WSU	3607	16.9 Kb	35.8 Mb	90.8 Kb	624
Tilletia caries	AAFC	10997	9.6 Kb	31.6 Mb	79.5 Kb	925
Tilletia contraversa	AAFC	11712	8.7 Kb	30.2 Mb	68.2 Kb	969
Tilletia indica	AAFC	7563	19.6 Kb	31.2 Mb	92.3 Kb	452
Tilletia walkeri	AAFC	2671	31.8 Kb	24.7 Mb	216.3 Kb	229

Table S3. Genes annotated per genome

strain	# gene predictions
Tilletia caries 517	11,805
Tilletia caries	10,773
Tilletia contraversa	10,451
Tilletia indica	10,393
Tilletia walkeri	8,798

Table S4. Primers

Primers at MAT-a				contig_74556-73883_resequenced	
Primer Name	**Direction**	**Sequence**	**Length**	**Minimum**	**Maximum**
Shared_right	F	TGCTCCCGATGCCATGAACGAGGGT	25	23,820	23,844
73883_right	F	GGTCATCCTGTGAAGAGATAAGGTG	25	22,363	22,387
STE3_FG_R	F	CCATATGAAGGCATTGATGG	20	19,339	19,358
ste3_19120-19620F	F	TCTTTGGCAGTGATCCGAAC	20	19,120	19,139
ste3_18659-19152F	F	GCTCTTGGATGAGGGTGAAC	20	18,659	18,678
STE3_FG_F	F	TCATTTCTCGAACTTCTCCA	20	18,127	18,146
ste3_18106-18722F	F	ATGTGAGAGAAGCCTATGCG	20	18,106	18,125
ste3_17626-18164F	F	CTGTTCCTCTCGCTTGACTT	20	17,630	17,649
ste3_17125-17686F	F	TCTTTGGCGCTGATTCTTGT	20	17,125	17,144
ste3_16920-17266F	F	CATGATCTTCCTGGTTGCCA	20	16,920	16,939
Mfa1_gDNA_R	F	GTAGAGCTGCAGAGGTGGAA	20	15,506	15,525
L_in_ortho_1	F	GGCTCCTTCAAGAAGCGCGAGCAGG	25	11,024	11,048
L_end_ortho_1	F	CATGCTCGCACTCGTCTCCAACACC	25	10,883	10,907
R_in_CAF1	R	GATGTGATCATGAGCTCGCCAAAGT	25	26,291	26,315
R_after_CAF1	R	GGTTGTCGGCCCACACATCGCAGAT	25	25,842	25,866
R_in_VWFA	R	GACAACTCGGAGTGGATGCGCAACG	25	25,391	25,415
ste3_16920-17266R	R	CTCCACTCTTCCAAGATCGC	20	25,109	25,128
74556_left	R	ATGACGTCGAAATGAGCACGGCAGA	25	24,283	24,307
Shared_left	R	TCGTCGTCGAGACCGAGTGCGCCGT	25	22,703	22,727
ste3_19120-19620R	R	CGCTCAGTTCTCCAAGGTC	19	19,602	19,620
ste3_18659-19152R	R	CGACAACTCGAAGAGATTGGA	21	19,152	19,172
ste3_18106-18722R	R	GCAATTTGGTCATTTCCGGT	20	18,703	18,722
ste3_17626-18164R	R	GGTGTAGAGACTCCGACTCA	20	18,164	18,183

ste3_17125-17686R	R	CGCTCTGTTCGACTCTCATC	20	17,667	17,686
ste3_16920-17266R	R	CTCCACTCTTCCAAGATCGC	20	17,207	17,226
Mfa1_gDNA_L	R	CCCACCGAACCTCTCAACTA	20	15,729	15,748
Primers at MAT-b				**contig 73754**	
Full_HD_F_73754	F	AGAAGAACGTTCGAGAGTCCA	21	15,315	15,335
FGENESH_8_F_73754	F	CTTTCCGAGGACAGAAGGCT	20	20,765	20,784
FGENESH_7_F_73754	F	ACGATGAGACTATGGGCGAG	20	17,748	17,767
FGENESH_6_F_73754	F	TCGAGGAAAGAAAGCCGAGA	20	15,298	15,317
FGENESH7-8F	F	TTGGGCAGAACTGAGGATGT	20	17,270	17,289
FGENESH6-7F	F	TTGGGCAGAACTGAGGATGT	20	17,270	17,289
bW2_partial_R	F	GGGAAGAGGCTCAGTACGAA	20	21,268	21,287
bW1_partial_R	F	TCTTGACCGTACTGAGCGAG	20	15,594	15,613
bE_partial_L	F	CCCTTCATGGCTTCGACAAC	20	19,295	19,314
Full_HD_R_73754	R	ACCATATCGGAACTTACCACCA	22	22,954	22,975
FGENESH_8_R_73754	R	ATGTGCTTCCTGTCCACCTT	20	23,000	23,019
FGENESH_7_R_73754	R	CGGCGAACAGGAAACTGAAA	20	20,229	20,248
FGENESH_6_R_73754	R	CACTCACCGCTCATCACTTG	20	17,468	17,487
FGENESH7-8R	R	TTTGAGATTGGGGAAGGGGA	20	17,916	17,935
FGENESH6-7R	R	TTTGAGATTGGGGAAGGGGA	20	17,916	17,935
bW2_partial_L	R	CTCATCAAGCTGGCGATGAC	20	21,889	21,908
bW1_partial_L	R	GCTTGGCCGAATCTTTCCAA	20	16,013	16,032
bE_partial_R	R	TCAGGCTGAGGTTGAGGTTC	20	19,710	19,729
RNA silencing gene primers				**contig 75716**	
Ago1_F	F	TCGCGTCACCTACACACAAA	20	10239	10259
Ago1_R	R	ATTCACTGCCAAGCGAAGGA	20	14252	14272
				contig 76014	
Ago2_F	F	CTTGAACCGAGCTCAGTCGT	20	17693	17713
Ago2_R	R	CCTCGCTCCTGCAGATCCTA	20	22027	22047

				contig 76414	
Ago3_F	F	CCTCATTGACTCAGGCGTGT	20	12249	12269
Ago3_R	R	CTCAACTCCGACGCTGAGAC	20	14121	14141
					contig 75596
Ago6_F	F	CTGCGTATGAGTGGGACTGG	20	889	**909**
Ago6_R	R	ATCGCCTCATTCAGCACCAA	20	1392	**1412**
				contig 37836-36863	
Dicer1_F	F	ACACTTAACCCTCCTTGCCG	20	163	183
Dicer1_R	R	GGATCTCGGTCATGGTCGTT	20	6055	6075
				contig 75969	
Dicer2_F	F	GGCGGTGTCCAGATGTATGT	20	3616	3636
Dicer2_R	R	CGGGAAGGGAACTCGGAAAG	20	10404	10424
				contig 76383	
RdRP_F	F	TGTCCTTCCGCGTCCTTAAC	20	34857	34877
RdRP_R	R	TGACAAAGGCAAGGATCCCC	20	39368	39388
				contig 75955	
CRM-1_F	F	ACACACCCCCATTCCCAATC	20	34857	34877
CRM-1_R	R	CTCACCCCTTGTCCTTCTCG	20	39368	39388

Table S5. Accessions of taxa used in Chapter Three phylogenies

Phylogeny: CRM1		
Accession	**Taxon Code**	**Organism**
None	Tcaries_CRM1	Tilletia caries
CBQ73251	Sreilianum_CRM1	Sporisorium reilianum
ADV23550	Cgattii_CRM1	Cryptococcus gattii
XP_001833702	Ccinerea_CRM1	Coprinopsis cinerea
EFJ01067	Scommune_CRM1	Schizophyllum commune
XP_003321285	Pgraminis_CRM1	Puccinia graminis
EDR11305	Lbicolor_CRM1	Laccaria bicolor
AFR96642	Cneoformans_CRM1	Cryptococcus neoformans
EST05152	Pbrasiliensis_CRM1	Pseudozyma brasiliensis
GAC97427	Phubeiensis_CRM1	Pseudozyma hubeiensis
EMS21461	Rtoruloides_CRM1	Rhodosporidium toruloides
EGG06975	Mlaricipopulina_CRM1	Melampsora larici-populina
CCF52436	Uhordei_CRM1	Ustilago hordei
XP_759006	Umaydis_CRM1	Ustilago maydis
GAC76163	Pantarctica_CRM1	Pseudozyma antarctica
EDP41778	Mglobosa_CRM1	Malassezia globosa
CCU99144	Msympodialis_CRM1	Malassezia sympodialis
GAA99903	Mosmundae_CRM1	Mixia osmundae
XP_001392127	Aniger_CRM1	Aspergillus niger
Phylogeny: RdRP		
Accession	**Taxon Code**	**Organism**
None	Tcaries_RdRP	Tilletia caries
CBQ71530	Sreilianum_RdRP	Sporisorium reilianum
XP_003195419	Cgattii_RdRP	Cryptococcus gattii
XP_001828874	Ccinerea_RdRP	Coprinopsis cinerea
XP_003031050	Scommune_RdRP	Schizophyllum commune
XP_003324286	Pgraminis_RdRP	Puccinia graminis
XP_001878528	Lbicolor_RdRP	Laccaria bicolor
XP_774375	Cneoformans_RdRP	Cryptococcus neoformans
EST06848	Pbrasiliensis_RdRP	Pseudozyma brasiliensis
GAC97192	Phubeiensis_RdRP	Pseudozyma hubeiensis

EMS19925	Rtoruloides_RdRP	Rhodosporidium toruloides
EGF97840	Mlaricipopulina_RdRP	Melampsora larici-populina
CCF48827	Uhordei_RdRP	Ustilago hordei
GAC77133	Pantarctica_RdRP	Pseudozyma antarctica
EHA18336	Aniger_RdRP	Aspergillus niger
Phylogeny: Dicer		
Accession	**Taxon Code**	**Organism**
None	Tcaries_dcr	Tilletia caries
XP_001398755	Aniger_dcr	Aspergillus niger
AEM61141	Pstriiformis_dcr	Puccinia striiformis
EIW60891	Tversicolor_dcr	Trametes versicolor
EPQ27664	Pflocculosa_dcr	Pseudozyma flocculosa
CBQ71581	Sreilianum_dcr	Sporisorium reilianum
CCF48410	Uhordei_dcr	Ustilago hordei
EST06983	Pbrasiliensis_dcr	Pseudozyma brasiliensis
XP_003337961	Pgraminis_dcr	Puccinia graminis
XP_003033469	Scommune_dcr	Schizophyllum commune
XP_001881178	Lbicolor_dcr	Laccaria bicolor
XP_002911949	Ccinerea_dcr	Coprinopsis cinerea
Phylogeny: Argonaute		
Accession	**Taxon Code**	**Organism**
None	Tcaries_AGO	Tilletia caries
CBQ68774	Sreilianum_AGO	Sporisorium reilianum
XP_003194007	Cgatti_AGO	Cryptococcus gattii
XP_001837864	Ccinerea_AGO	Coprinopsis cinerea
XP_003031830	Scommune_AGO	Schizophyllum commune
XP_003329577	Pgraminis_AGO	Puccinia graminis
XP_001878380	Lbicolor_AGO	Laccaria bicolor
XP_567314	Cneoformans_AGO	Cryptococcus neoformans
EST09042	Pbrasiliensis_AGO	Pseudozyma brasiliensis
GAC98167	Phubeiensis_AGO	Pseudozyma hubeiensis
EMS24354	Rtoruloides_AGO	Rhodosporidium toruloides
EGU12104	Rglutinis_AGO	Rhodotorula glutinis
GAA94730	Mosmundae_AGO	Mixia osmundae

AEM61140	Pstriiformis_AGO	Puccinia striiformis
EGG01332	Mlaricipopulina_AGO	Melampsora larici-populina
EPQ25705	Pflocculosa_AGO	Pseudozyma flocculosa
CCF50705	Uhordei_AGO	Ustilago hordei
EHA21487	Aniger_AGO	Aspergillus niger

Table S6. High scoring contigs from *T. caries* 517 for RNA-silencing queries

Query Gene Ortholog	***S. reilianum* Accession**	**Contig Hit 1**	**E-val**	**Contig Hit 2**	**E-val**	**Contig Hit 3**	**E-val**	**Contig Hit 4**	**E-val**	**Contig Hit 5**	**E-val**	**Contig Hit 6**	**E-val**
RDRP I	CBQ71530	76383_contig	1e^-53										
RDRP II	CBQ70480	76383_contig	7e^-42										
RDRP III	CBQ68774	76383_contig	2.00e^-9										
Argonaut	CBQ71581	75716_contig	1.00e^-111	76014_contig	1.00e^-81	76414_contig	5.00e^-20	76036_contig	4.00e^-18	75596_contig	5.0e^-14	74923_contig	5.00e^-14
Dicer	CBQ68774	36863_contig	3.00e^-29	75969_contig	2.00e^-28	37419_contig	4.00e^-17	37386_contig	1.00e^-15				
Exportin-5	CBQ73251	75955_contig	0										

```
#-------------------------------------------------------------------------
#!/usr/bin/perl
#http://code.izzid.com/2011/10/31/How-to-read-a-fasta-file-in-perl.html
#How to read a fasta file in Perl
#Created: Oct 31, 2011
#By: Jeremiah Faith
#Modified by: Sean McCotter
#2013

my $fasta_file=shift;

my $fh;
open($fh, $fasta_file) or die "can't open $fasta_file: $!\n";
open (OUT, ">out_fasta.fasta") or die "can't open $out_fasta: $!\n";

my %sequence_data;
#modify the following regex to extract fasta files based on sequence or
header/description contents. Current setting is to extract protein sequences
with a farnesylation/prenylation motif.
while (read_fasta_sequence($fh, \%sequence_data)) {
   #if ($sequence_data{header} =~ /my header regex/
   if ($sequence_data{seq} =~ /.*?C[GAVLI][GAVLI].$/) {
     print OUT ">$sequence_data{header}\n$sequence_data{seq}\n\n";
   }
}

sub read_fasta_sequence {
   my ($fh, $seq_info) = @_;

   $seq_info->{seq} = undef; # clear out previous sequence

   # put the header into place
   $seq_info->{header} = $seq_info->{next_header} if $seq_info-
>{next_header};

   my $file_not_empty = 0;
   while (<$fh>) {
      $file_not_empty = 1;
      next if /^\s*$/;  # skip blank lines
      chomp;

      if (/^>/) { # fasta header line
         my $h = $_;
         $h =~ s/^>//;
         if ($seq_info->{header}) {
            $seq_info->{next_header} = $h;
            return $seq_info;
         }
         else { # first time through only
            $seq_info->{header} = $h;
         }
      }
```

```
      else {
         s/\s+//;  # remove any white space
         $seq_info->{seq} .= $_;
      }
   }

   if ($file_not_empty) {
      return $seq_info;

   }
   else {
      # clean everything up
      $seq_info->{header} = $seq_info->{seq} = $seq_info->{next_header} =
undef;

      return;
   }
}
#-------------------------------------------------------------------------
#!/usr/bin/perl
#
# assemblathon_stats.pl
#
# A script to calculate a basic set of metrics from a genome assembly
#
# Author: Keith Bradnam, Genome Center, UC Davis
# This work is licensed under a Creative Commons Attribution-NonCommercial-
ShareAlike 3.0 Unported License.
#
# Last updated by: $Author: keith $
# Last updated on: $Date: 2011/10/13 00:07:00 $

use strict;
use warnings;
use FAlite;
use Getopt::Long;
use List::Util qw(sum max min);

#############################################
#
#  C o m m a n d   l i n e   o p t i o n s
#
#############################################

my $limit;       # limit processing of data to first $limit sequences (for
quick testing)
my $graph;       # produce some output ready for Excel or R
my $csv;         # produce CSV output file of results
my $n_limit;     # how many N characters should be used to split scaffolds
into contigs
my $genome_size; # estimated or known genome size (will be used for some
stats)

GetOptions ("limit=i"       => \$limit,
```

```
                  "csv"           => \$csv,
                  "graph"         => \$graph,
                  "n=i"           => \$n_limit,
                  "genome_size=i" => \$genome_size);

# set defaults
$limit = 1000000000 if (!$limit);
$n_limit = 25       if (!$n_limit);


# check we have a suitable input file
my $usage = "Usage: assemblathon_stats.pl <assembly_scaffolds_file>
options:
      -limit <int> limit analysis to first <int> sequences (useful for
testing)
      -csv         produce a CSV output file of all results
      -graph       produce a CSV output file of NG(X) values (NG1 through to
NG99), suitable for graphing
      -n <int>     specify how many consecutive N characters should be used
to split scaffolds into contigs
      -genome_size <int> estimated or known genome size
";

die "$usage" unless (@ARGV == 1);
my ($file) = @ARGV;

#############################################
#
#  S o m e   G l o b a l   v a r i a b l e s
#
#############################################

my $scaffolded_contigs = 0;                     # how many contigs that are
part of scaffolds (sequences must have $n_limit consecutive Ns)
my $scaffolded_contig_length = 0;         # total length of all scaffolded
contigs
my $unscaffolded_contigs = 0;             # how many 'orphan' contigs, not
part of a scaffold
my $unscaffolded_contig_length = 0;       # total length of all contigs not
part of scaffold
my $w = 60;                                             # formatting width for
output
my %data;                                               # data structure to
hold all sequence info key is either 'scaffold', 'contig' or intermediate',
values are seqs & length arrays
my (@results, @headers);                        # arrays to store results
(for use with -csv option)



# make first loop through file, capture some basic info and add sequences to
arrays
process_FASTA($file);
```

```
print "\n---------------- Information for assembly \'$file\' ----------------
\n\n";

if(defined($genome_size)){
      my $mbp_size = sprintf("%.2f", $genome_size / 1000000);
      printf "%${w}s %10s\n", "Assumed genome size (Mbp)", $mbp_size;
}

# produce scaffold statistics
sequence_statistics('scaffold');

# produce a couple of intermediate statistics based on scaffolded contigs vs
unscaffolded contigs
sequence_statistics('intermediate');

# finish with contig stats
sequence_statistics('contig');

# produce CSV output if required
write_csv($file) if ($csv);

exit(0);



##########################################
#
#
#    S  U  B  R  O  U  T  I  N  E  S
#
#
##########################################


##########################################
#    M A I N  loop through FASTA file
##########################################

sub process_FASTA{

      my ($seqs) = @_;

      my $input;

      # if dealing with gzip file, treat differently
      if($seqs =~ m/\.gz$/){
            open($input, "gunzip -c $seqs |") or die "Can't open a pipe to
$seqs\n";
      } else{
            open($input, "<", "$seqs") or die "Can't open $seqs\n";
      }

      my $fasta = new FAlite(\*$input);
```

```
	# want to keep track of various contig + scaffold counts
	my $seq_count = 0;

	while(my $entry = $fasta->nextEntry){
	    my $seq = uc($entry->seq);
	      my $length = length($seq);
	      $seq_count++;

	      # everything gets pushed to scaffolds array
	      push(@{$data{scaffold}{seqs}},$seq);
	      push(@{$data{scaffold}{lengths}},$length);

	      # if there are not at least 25 consecutive Ns in the sequence we
need to split it into contigs
	      # otherwise the sequence must be a contig itself and it still
needs to be put in @contigs array
	      if ($seq =~ m/N{$n_limit}/){

	            # add length to $scaffolded_contig_length
	            $scaffolded_contig_length += $length;

	            # loop through all contigs that comprise the scaffold
	            foreach my $contig (split(/N{25,}/, $seq)){
	                  $scaffolded_contigs++;
	                  my $length = length($contig);
	                  push(@{$data{contig}{seqs}},$contig);
	                  push(@{$data{contig}{lengths}},$length);
	            }
	      } else {
	            # must be here if the scaffold is actually just a contig
(or is a scaffold with < 25 Ns)
	            $unscaffolded_contigs++;
	            $unscaffolded_contig_length += $length;
	            push(@{$data{contig}{seqs}},$seq);
	            push(@{$data{contig}{lengths}},$length);
	      }
	      # for testing, just use a few sequences
	      last if ($seq_count >= $limit);

	}
	close($input);
}


########################################
#    Calculate basic assembly metrics
########################################

sub sequence_statistics{
	my ($type) = @_;

	print "\n";

	# need descriptions of each result
```

```
	my $desc;

	# there are just a couple of intermediate level statistics to print
	if($type eq 'intermediate'){
		my $total_size = sum(@{$data{scaffold}{lengths}});

		# now calculate percentage of assembly that is accounted for by
scaffolded contigs
		my $percent = sprintf("%.1f",($scaffolded_contig_length /
$total_size) * 100);
		$desc = "Percentage of assembly in scaffolded contigs";
		printf "%${w}s %10s\n", $desc, "$percent%";
		store_results($desc, $percent) if ($csv);

		# now calculate percentage of assembly that is accounted for by
unscaffolded contigs
		$percent = sprintf("%.1f",($unscaffolded_contig_length /
$total_size) * 100);
		$desc = "Percentage of assembly in unscaffolded contigs";
		printf "%${w}s %10s\n", $desc, "$percent%";
		store_results($desc, $percent) if ($csv);


		# statistics that describe N regions that join contigs in
scaffolds

		# get number of breaks
		my $contig_count = scalar(@{$data{contig}{lengths}});
		my $scaffold_count = scalar(@{$data{scaffold}{lengths}});
		my $average_contigs_per_scaffold = sprintf("%.1f",$contig_count /
$scaffold_count);
		$desc = "Average number of contigs per scaffold";
		printf "%${w}s %10s\n", $desc, $average_contigs_per_scaffold;
		store_results($desc, $average_contigs_per_scaffold) if ($csv);

		# now calculate average length of break between contigs
		# just find all runs of Ns in scaffolds (>=25) and calculate
average length
		my @contig_breaks;
		foreach my $scaffold (@{$data{scaffold}{seqs}}){
			while($scaffold =~ m/(N{25,})/g){
				push(@contig_breaks, length($1));
			}
		}
		# set break size to zero if there are no Ns in scaffolds
		my $average_break_length;

		if(@contig_breaks == 0){
			$average_break_length = 0;
		} else{
			$average_break_length = sum(@contig_breaks) /
@contig_breaks;
		}
```

```
            $desc = "Average length of break (>25 Ns) between contigs in
scaffold";
            printf "%${w}s %10d\n", $desc, $average_break_length;
            store_results($desc, $average_break_length) if ($csv);
            return();
      }


      # n
      my $count = scalar(@{$data{$type}{lengths}});
      $desc = "Number of ${type}s";
      printf "%${w}s %10d\n", $desc, $count;
      store_results($desc, $count) if ($csv);



      # more contig details (only for contigs)
      if ($type eq 'contig'){
            $desc = "Number of contigs in scaffolds";
            printf "%${w}s %10d\n",$desc, $scaffolded_contigs;
            store_results($desc, $scaffolded_contigs) if ($csv);

            $desc = "Number of contigs not in scaffolds";
            printf "%${w}s %10d\n", $desc,$unscaffolded_contigs;
            store_results($desc, $unscaffolded_contigs) if ($csv);
      }


      # total size of sequences
      my $total_size = sum(@{$data{$type}{lengths}});
      $desc = "Total size of ${type}s";
      printf "%${w}s %10d\n", $desc, $total_size;
      store_results($desc, $total_size) if ($csv);


      # For scaffold data only, can caluclate the percentage of known genome
size
      if ($type eq 'scaffold' && defined($genome_size)){
            my $percent = sprintf("%.1f",($total_size / $genome_size) * 100);
            $desc = "Total scaffold length as percentage of assumed genome
size";
            printf "%${w}s %10s\n", $desc, "$percent%";
            store_results($desc, $percent) if ($csv);
      }


      # longest and shortest sequences
      my $max = max(@{$data{$type}{lengths}});
      $desc = "Longest $type";
      printf "%${w}s %10d\n", $desc, $max;
      store_results($desc, $max) if ($csv);

      my $min = min(@{$data{$type}{lengths}});
      $desc = "Shortest $type";
```

```
      printf "%${w}s %10d\n", $desc, $min;
      store_results($desc, $min) if ($csv);


      # find number of sequences above certain sizes
      my %sizes_to_shorthand = (1000     => '1K',
                                10000    => '10K',
                                100000   => '100K',
                                1000000  => '1M',
                                10000000 => '10M');

      foreach my $size qw(1000 10000 100000 1000000 10000000){
            my $matches = grep { $_ > $size } @{$data{$type}{lengths}};
            my $percent = sprintf("%.1f", ($matches / $count) * 100);

            $desc = "Number of ${type}s > $sizes_to_shorthand{$size} nt";
            printf "%${w}s %10d %5s%%\n", $desc, $matches, $percent;
            store_results($desc, $matches)  if ($csv);

            $desc = "Percentage of ${type}s > $sizes_to_shorthand{$size} nt";
            store_results($desc, $percent)  if ($csv);      }


      # mean sequence size
      my $mean = sprintf("%.0f",$total_size / $count);
      $desc = "Mean $type size";
      printf "%${w}s %10d\n", $desc, $mean;
      store_results($desc, $mean) if ($csv);

      # median sequence size
    my $median = (sort{$a <=> $b} @{$data{$type}{lengths}})[$count/2];
      $desc = "Median $type size";
      printf "%${w}s %10d\n", $desc, $median;
      store_results($desc, $median) if ($csv);



      ##########################################################################
############
      #
      # N50 values
      #
      # Includes N(x) values, NG(x) (using assumed genome size)
      # and L(x) values (number of sequences larger than or equal to N50
sequence size)
      ##########################################################################
############

      # keep track of cumulative assembly size (starting from smallest seq)
      my $running_total = 0;

      # want to store all N50-style values from N1..N100. First target size
to pass is N1
      my $n_index = 1;
```

```
	my @n_values;
	my $n50_length = 0;

	my $i = 0;

	my $x = $total_size * 0.5;
	# start with longest lengths scaffold/contig
	foreach my $length (reverse sort{$a <=> $b} @{$data{$type}{lengths}}){
		$i++;
		$running_total += $length;

		# check the current sequence and all sequences shorter than
current one
		# to see if they exceed the current NX value
		while($running_total > int (($n_index / 100) * $total_size)){
			if ($n_index == 50){
				$n50_length = $length;
				$desc = "N50 $type length";
				printf "%${w}s %10d\n", $desc, $length;
				store_results($desc, $length) if ($csv);

				# L50 = number of scaffolds/contigs that are longer
than or equal to the N50 size
				$desc = "L50 $type count";
				printf "%${w}s %10d\n","L50 $type count", $i;
				store_results($desc, $i) if ($csv);
			}
			$n_values[$n_index] = $length;
			$n_index++;
		}
	}

	my @ng_values;

	# do we have an estimated/known genome size to work with?
	if(defined($genome_size)){
		my $ng_index = 1;
		my $ng50_length = 0;

		$running_total = 0;
		$i = 0;

		foreach my $length (reverse sort{$a <=> $b}
@{$data{$type}{lengths}}){
			$i++;
			$running_total += $length;

			# now do the same for NG values, using assumed genome size
			while($running_total > int (($ng_index / 100) *
$genome_size)){
				if ($ng_index == 50){
					$ng50_length = $length;
					$desc = "NG50 $type length";
					printf "%${w}s %10d\n", $desc, $length;
```

```
                              store_results($desc, $length) if ($csv);

                              $desc = "LG50 $type count";
                              printf "%${w}s %10d\n", $desc, $i;

                              store_results($desc, $i) if ($csv);
                        }
                        $ng_values[$ng_index] = $length;
                        $ng_index++;
                  }
            }


            my $n50_diff = abs($ng50_length - $n50_length);
            $desc = "N50 $type - NG50 $type length difference";
            printf "%${w}s %10d\n", $desc, $n50_diff;
            store_results($desc, $n50_diff) if ($csv);
      }
      # add final value to @n_values and @ng_values which will just be the
shortest sequence
#     $n_values[100] = $min;
#     $ng_values[100] = $min;


      # base frequencies
      my %bases;

    my $seq = join('',@{$data{$type}{seqs}});
      my $length = length($seq);

    # count mononucleotide frequencies
    $bases{A} = ($seq =~ tr/A/A/);
    $bases{C} = ($seq =~ tr/C/C/);
    $bases{G} = ($seq =~ tr/G/G/);
    $bases{T} = ($seq =~ tr/T/T/);
    $bases{N} = ($seq =~ tr/N/N/);

      my $base_count = 0;
      foreach my $base qw (A C G T N){
            my $percent = sprintf("%.2f", ($bases{$base} / $length) * 100);
            $desc = "$type %$base";
            printf "%${w}s %10s\n", $desc, $percent;
            store_results($desc, $percent) if ($csv);
            $base_count += $bases{$base};
      }

    # calculate remainder ('other) in case there are other characters present
      my $other = $length - $base_count;
      my $percent = sprintf("%.2f", ($other / $length) * 100);
      $desc = "$type %non-ACGTN";
      printf "%${w}s %10s\n",$desc, $percent;
      store_results($desc, $percent) if ($csv);

      $desc = "Number of $type non-ACGTN nt";
```

```
	printf "%${w}s %10d\n",$desc, $other;
	store_results($desc, $other) if ($csv);


	# anything to dump for graphing?
	if($graph){

		# create new output file name
		my $file_name = $file;
		$file_name =~ s/\.gz$//;
		$file_name =~ s/\.(fa|fasta)$//;
		$file_name .= ".${type}.NG50.csv";

		open(my $out, ">", "$file_name") or die "Can't create
$file_name\n";
		print $out join (',',"Assembly",1..99), "\n";

		# make some guesses of what might constitute the unique assembly
ID
		my $assembly_ID = $file;
		($assembly_ID) = $file =~ m/^([A-Z]\d{1,2})_/ if ($file =~ m/^[A-
Z]\d{1,2}_/);
		($assembly_ID) = $file =~ m/^((bird|snake|fish)_\d+(C|E))_/ if
($file =~ m/^(bird|snake|fish)_\d+C|E_/);

		# CSV file, with filename in first column
		print $out "$assembly_ID";

		for (my $i = 1; $i < 100; $i++){
			# higher NG values might not be present if assembly is poor
			if (defined $ng_values[$i]){
				print $out ",$ng_values[$i]";
			} else{
				print $out ",0";
			}
		}
		print $out "\n";
		close($out);
	}
}

# simple routine to add results to a pair of arrays that will be used for
printing results later on
# if -csv option is used
sub store_results{
	my ($desc, $result) = @_;

	push(@headers,$desc);
	push(@results,$result);
}

sub write_csv{
	my ($file) = @_;
```

```
	# create new output file name
	my $output = $file;
	$output =~ s/\.gz$//;
	$output =~ s/\.(fa|fasta)$//;
	$output .= ".csv";

	# make some guesses of what might constitute the unique assembly ID
	my $assembly_ID = $file;
	($assembly_ID) = $file =~ m/^([A-Z]\d{1,2})_/ if ($file =~ m/^[A-
Z]\d{1,2}_/);
	($assembly_ID) = $file =~ m/^((bird|snake|fish)_\d+(C|E))_/ if ($file
=~ m/^(bird|snake|fish)_\d+C|E_/);

	open(my $out, ">", $output) or die "Can't create $output\n";

	print $out "Assembly,";
	foreach my $header (@headers){
		print $out "$header,";
	}
	print $out "\n";

	print $out "$assembly_ID,";
	foreach my $result (@results){
		print $out "$result,";
	}
	print $out "\n";


	close($out);
}
#-------------------------------------------------------------------------------
#!/usr/bin/perl
# # # # # #
# http://www.uppmax.uu.se/userscript/extract-specific-sequences-from-fasta-
#file
#extractFromFasta.pl
#written by Linnéa Smeds May 2010, mod Feb 2011
# ====================================================
# Extracts certain sequences from a fasta file. Either
# a single sequence name is given, or a text file with
# a list of names (type must be given as input, either
# "single" or "list" .
# ====================================================
# Usage: extractFromFasta.pl <seqfile.fa>
#                       <single|list> <name|list.txt>
#
# Example: extractFromFasta.pl mySeq.fa single "contig4" \
#           >contig4.fa

use strict;
use warnings;

# Input parameters
my $scaffold_file = $ARGV[0];
```

```
my $type = $ARGV[1];
my $query = $ARGV[2];

# Save wanted fasta headers
my %list=();
if ($type eq "list") {
      open(IN, $query);
      while(<IN>) {
            chomp($_);
            $_=~s/>//;
            $list{$_} = 1;
      }
}
elsif($type eq "single") {
      $query=~s/>//;
      $list{$query}=1;
}
else {
      die &usage();
}

#Go through fasta file, extract sequences
open(IN, $scaffold_file);
my $seq = "";
my $flag = "off";
while(<IN>) {
      if($_ =~ m/^>/) {

            my $head = $_;
            chomp($head);
            $head=~s/>//;

            if(defined $list{$head}) {
                  print $_;
                  $flag = "on";
            }
            else {
                  if($type eq "single" && $flag eq "on") {
                        exit;
                  }
                  $flag = "off";
            }
      }
      else {

            if($flag eq "on") {
                  print $_;
            }
      }
}

sub usage {
      print << "A";
\nextractFromFasta.pl
```

```
written by Linnéa Smeds May 2010, mod Feb 2011
=================================================
Extracts given sequences from a fasta file. Either a
single sequence name is given, or a text file with
a list of names (type must be given as input, either
 "single" or "list" .
=================================================
Usage: extractFromFasta.pl <seqfile.fa>
                  <single|list> <name|list.txt>
\nExample: extractFromFasta.pl mySeq.fa single "contig4" >contig4.fa\n
A

exit;
}
#-----------------------------------------------------------------------------
#http://bioinf.uni-greifswald.de
!/usr/bin/perl
# getAnnoFast.pl
# Creates fasta sequence files from the AUGUSTUS output.
#
# Mario Stanke, 10.05.2007
#

use strict;
use Getopt::Long;

my $usage = "getAnnoFasta.pl augustus.gff\n";
$usage .= "   Makes a fasta file with protein sequences (augustus.aa)\n";
$usage .= "   and one with coding sequences (augustus.codingseq)\n";
$usage .= "   from the sequences provided in the comments of the AUGUSTUS
output.\n";
$usage .= "   These sequence comments are turned on with --protein=on and --
codingseq=on, respectively\n";
$usage .= "Options:\n";
$usage .= "   --seqfile=s  Input a fasta file with the genomic sequences that
AUGUSTUS was run on.\n";
$usage .= "                When this option is given, an additional file with
the individual\n";
$usage .= "                coding exon sequences (augustus.cdsexons) is
output.\n";
$usage .= "                and a file with the complete mRNA including UTRs
(augustus.mrna) is output.\n";

my ($seqname, $trid, $status, $haveCod, $haveAA, $haveCDS, $haveRNA, $seq,
$seqfile);

GetOptions('seqfile=s'=>\$seqfile);

if ($#ARGV != 0) {
    print $usage;
    exit;
}

my $separator = ";";
```

```
my $augustusfilename = $ARGV[0];
open(AUG, "<$augustusfilename") || die "Couldn't open $augustusfilename\n";
my $stemfilename = $augustusfilename;
$stemfilename =~ s/(\.gff|\.gtf|.gff3|\.txt)//;
my %sequence = (); # Hash with the DNA sequence. The sequence names are the
keys.

# Read in the sequence file in one chunk.
# And sort it in the sequence hash.
# Yes, this requires a lot of memory for large genomes.
if ($seqfile){
    open (SEQ, "<$seqfile") or die ("Could not open sequence file
$seqfile\n");
    $/=">";
    while (<SEQ>){
      s/>$//;
      next unless /\S+/;
      /(.*)\n/;

      $seqname = $1;
      my $sequencepart = $'; #'
      $seqname =~ s/\s.*//; # seqname only up to first white space
      $sequencepart =~ s/\s//g;
      $sequence{$seqname} = $sequencepart;
    }
    print "Read in " . (scalar keys %sequence) . " sequence(s) from
$seqfile.\n";
}
$/="\n";

#
# Go through the augustus output transcript by transcript.
#
$haveCod = $haveAA = $haveCDS = $haveRNA = 0;
$status = 0;
my $exonUTRFormat = 0; # UTR implicitly given by exon features
my $UTRFormat = 0; # UTR explicitly given by *UTR features
my $cdsSeq = "";
my $aaSeq = "";
my %cdsnr = ();
my %cdsTx = (); #keys transcript id, values concatenated coding exon
sequences
my %mrnaTx = (); #keys transcript id, values concatenated exon sequences
(including UTR)
my %strandTx = (); #keys transcript id, values strands

while(<AUG>) {
    if ($seqfile &&
(/^(\S+)\t\S+\t(\S+)\t(\d+)\t(\d+)\t\S+\t(\S+)\t\S+\ttranscript_id "([^"]*)";
gene_id "([^"]*)";$/
                ||
/^(\S+)\t\S+\t(\S+)\t(\d+)\t(\d+)\t\S+\t(\S+)\t\S+\t.*Parent=([^;]+)/)){
      my $feat = $2;
      $seqname = $1;
```

```
      my $start = $3;
      my $end = $4;
      my $strand = $5;
      $trid=$6;
      $trid =~ s/\s$//;
      next unless ($feat eq "CDS" || $feat =~ /UTR/ || $feat eq "exon");
      # decide whether to use exon or UTR format for mRNA by whether we see
UTR or exon first
      $UTRFormat = 1 if (!$exonUTRFormat && $feat =~ /UTR/);
      $exonUTRFormat = 1 if (!$UTRFormat && $feat eq "exon");
      $cdsnr{$trid}++ if ($feat eq "CDS");
      my $seqpart = lc(substr($sequence{$seqname}, $start-1, $end - $start +
1));
      #print "$seqname $trid CDS $cdsnr{$trid} $start -> $end $seqpart\n";
      if ($seqpart ne "") {
          # add mRNA if applicable
          if (($exonUTRFormat && $feat eq "exon") ||
            ($UTRFormat && ($feat eq "CDS" || $feat =~ /UTR/))) {
            $mrnaTx{$trid} = "" if (!defined($mrnaTx{$trid}));
            $mrnaTx{$trid} .= $seqpart;
          }
          if ($feat eq "CDS"){
            if (!$haveCDS) {
                open (CDSEXON, ">$stemfilename.cdsexons");
                $haveCDS++;
            }
            $cdsTx{$trid} = "" if (!defined($cdsTx{$trid}));
            $cdsTx{$trid} .= $seqpart;
            if ($strand eq '-') {
                $seqpart = rc($seqpart);
                $strandTx{$trid} = $strand;
            }
            print CDSEXON ">$trid.cds" . $cdsnr{$trid} . "\n$seqpart\n";
          }
      }
    }
    if (/^(\S+)\t.*\ttranscript_id "([^"]*)"; gene_id "([^"]*)";$/ ||
      /^(\S+)\t.*Parent=([^;]+)/){
      $seqname=$1;
      $trid=$2;
      $trid =~ s/\s$//;
      $status=1;
    } elsif (/coding sequence = \[(.*)/ && $status == 1){
      if ($haveCod == 0) {
          open (COD, ">$stemfilename.codingseq");
      }
      $haveCod++;
      $seq = $1;
      $seq =~ s/\]$//;
      print COD ">$seqname.$trid\n$seq\n";
      $status=2;
    } elsif ($status == 2 && /^\# ([\w\]]*)$/){
      $seq = $1;
      $seq =~ s/\]$//;
```

```
      print COD "$seq\n";
      $status=2;
    } elsif (/protein sequence = \[(.*)/ && $status >= 1){
      if ($haveAA == 0) {
          open (AA, ">$stemfilename.aa");
      }
      $haveAA++;
      $seq = $1;
      $seq =~ s/\]$//;
#     print AA ">$seqname$separator$trid\n$seq\n";
#     print AA ">$trid\n";
      $aaSeq .= $seq;
      if (!/\]/){
          $status=3;
      } else {
          if ($aaSeq ne ""){
            print AA ">$trid\n";
            print AA getFa($aaSeq, 100);
          }
          $aaSeq = "";
          $status=1;
      }
    } elsif ($status == 3 && /^\# (.*)/){
      $seq = $1;
      $seq =~ s/\]$//;
#     print AA "$seq\n";
      $aaSeq .= $seq;
      if (!/\]/){
          $status=3;
      } else {
          if ($aaSeq ne ""){
            print AA ">$trid\n";
            print AA getFa($aaSeq, 100);
          }
          $aaSeq = "";
          $status=1;
      }
    }
}

#
# print coding sequences, if not already done (because included in output)
#
if (!$haveCod && scalar(keys %cdsTx)>0){
    open (COD, ">$stemfilename.codingseq") or die ("Could not open
$stemfilename.codingseq for writing.");
    foreach my $trid (sort by_id keys %cdsTx){
      print COD ">$trid\n";
      my $codingseq = $cdsTx{$trid};
      $codingseq = rc($codingseq) if ($strandTx{$trid} eq "-");
      print COD getFa($codingseq);
    }
}
```

```
#
# print mRNA sequences
#
if (scalar(keys %mrnaTx)>0){
    open (MRNA, ">$stemfilename.mrna") or die ("Could not open
$stemfilename.mrna for writing.");
    foreach my $trid (sort by_id keys %mrnaTx){
      print MRNA ">$trid\n";
      my $mrnaseq = $mrnaTx{$trid};
      $mrnaseq = rc($mrnaseq) if ($strandTx{$trid} eq "-");
      print MRNA getFa($mrnaseq);
    }
}


#
# sort by increasing transcript id
#
sub by_id{
    $a =~ /g(\d+)\.t(\d+)/;
    my ($ag,$at)=($1,$2);
    $b =~ /g(\d+)\.t(\d+)/;
    my ($bg,$bt)=($1,$2);
    if ($ag>$bg){
      return 1;
    } elsif ($bg>$ag){
      return -1;
    } else {
      return $at <=> $bt;
    }
}


# reverse complement
sub rc{
    my $s = shift;
    $s = reverse $s;
    $s =~ s/a/T/g;
    $s =~ s/c/G/g;
    $s =~ s/g/C/g;
    $s =~ s/t/A/g;
    $s = lc $s;
    return $s;
}


sub getFa{
    my $seq = shift;
    my $cols = 100;
    $cols = shift if (@_);

    my $start = 0;
    my $ret = "";
    while (length($seq)-$start >= $cols) {
```

```
      my $shortline = substr($seq, $start, $cols);
      $ret .= "$shortline\n";
      $start += $cols;
    }
    $ret .= substr($seq, $start, $cols) . "\n" if ($start<length($seq));
    return $ret;
}
#------------------------------------------------------------------------------
#!/usr/local/bin/perl -w
#used to parse cluster output from MCL
############################################################
# Author          : Ping Zheng
# Target          : Main Lab
# Name                  : createSubfiles.pl
# Date                  : 09-18-14
# Version         : 1.0.0
# Description   : Create a sub files
# <input>         : tab text
# Dependencies    : N/A
# Usage                 : %>*.pl <tab text>
############################################################
use warnings;
use strict;
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
#                   GLOBAL VARIABLES                       #
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
my %data =();
my $output = "output";
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
#                          MAIN                            #
#~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~~#
main();
exit;
#-------------------------------------------------------------#
# main function
#-------------------------------------------------------------#
sub main {
# Check command line
      die "\n\tInvalid command line\n\tUsage:\n\t>perl $0 tabfile  \n\n"if
($#ARGV != 0);

      open LIST_INPUT, "<$ARGV[0]" or die "Can't open $ARGV[0]: $!\n";
      my $count = 0;
      while (<LIST_INPUT>){
            /^\s*$/ and next;
            $count++;
            my @cols = split(/\t/,$_);
            $data{$count} =[@cols];
      }
      close(LIST_INPUT);

      mkdir($output) if !-e $output;

      for my $file (sort keys %data)
```

```
      {     open (OUTPUT, ">$output/line_$file.txt")|| die "Cannot create
file";
            print OUTPUT join("\n", @{$data{$file}}), "\n";
            close OUTPUT;
      }
}
```

HMMTOP OUTPUT

```
>Tc517_STE3 g3435
MKDNAYGPLGIVAIILAASVLPWHIKAKNTGVLILSAWLIVSVSCLTINAFIWWRNVDIRCQIWCDISTKLIFGAYM
GLP
CSSICIIRQLEEIGSTRRVRITAKDRKHKLYFDLGVGVAIPLLYMALHIVNQGHRFDVIESVGCFPTYYITPVAIVL
ILV
PPVVASAVALIYSFLALRWFIIRRRQFNAVLQNSHSGLDRSRYLRLMTMASTEAIWSFPVNMTILVTKFTLIQEPIN
PYV
SWQDTHFDFGRIGQFPAEFFDVSPDTKAYWRANMALGRYGALVGCFFLFVFFGTSQDALKFYGTILRKITSIFPQFQ
AQR
TMSPTTSSQADHWNIEVCVSKEAGVETPTQSEMDRMEKFEK

Protein: noname
Length:  367
N-terminus: OUT
Number of transmembrane helices: 7
Transmembrane helices: 11-29 36-54 77-93 116-136 159-183 221-238 280-300

Total entropy of the model:  17.0111
Entropy of the best path:  17.0150

The best path:

     seq  TCSTEGMKDN AYGPLGIVAI ILAASVLPWH IKAKNTGVLI LSAWLIVSVS    50
     pred OOOOooooooo HHHHHHHHHH HHHHHHHHHi iiiiiHHHHH HHHHHHHHHH

     seq  CLTINAFIWW RNVDIRCQIW CDISTKLIFG AYMGLPCSSI CIIRQLEEIG   100
     pred HHHHooooooo ooooooooooo oooooHHHH HHHHHHHHHH HHHiiiiiii

     seq  STRRVRITAK DRKHKLYFDL GVGVAIPLLY MALHIVNQGH RFDVIESVGC   150
     pred iiiiiiiiii iiiiiHHHHH HHHHHHHHHH HHHHHHoooo oooooooooo

     seq  FPTYYITPVA IVLILVPPVV ASAVALIYSF LALRWFIIRR RQFNAVLQNS   200
     pred ooooooooHH HHHHHHHHHH HHHHHHHHHH HHHiiiiiii iiiiiiiiII

     seq  HSGLDRSRYL RLMTMASTEA IWSFPVNMTI LVTKFTLIQE PINPYVSWQD   250
     pred IIIIIiiiii iiiiiiiiii HHHHHHHHHH HHHHHHHHoo oooooooooo

     seq  THFDFGRIGQ FPAEFFDVSP DTKAYWRANM ALGRYGALVG CFFLFVFFGT   300
     pred oooOOOOOOO OOOOooooooo ooooooooooH HHHHHHHHHH HHHHHHHHHH

     seq  SQDALKFYGT ILRKITSIFP QFQAQRTMSP TTSSQADHWN IEVCVSKEAG   350
     pred iiiiiiiiii iiiiiIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

     seq  VETPTQSEMD RMEKFEK  367
     pred IIIIIIIIII IIIIIII


If you are going to use these results in your work, please cite:

    G.E Tusnády and I. Simon (1998) J. Mol. Biol. 283, 489-506.
```

G.E Tusnády and I. Simon (2001) Bioinformatics 17, 849-850.

```
  >Tca_STE3 g9885
MKDNAYGPLGIVAIILAASVLPWHIKAKNTGVLILSAWLIVSVSCLTINAFIWWRNVDIRCQIWCDISTKLIFGAYM
GLP
CSSICIIRQLEEIGSTRRVRITAKDRKHKLYFDLGVGVAIPLLYMALHIVNQGHRFDVIESVGCFPTYYLTPVAIVL
ILV
PPVLASSVALIYSFLALRWFFIRRRQFQAVLQNSHSGLNRSRYLRLMAMAGTEALWSFPVNVTILVTKFTLLKLPTY
PYI
SWEDTHFNFGRADQFDSTFFDSDPDLKAYWRSNIALGRYGVLVGCFFLFVFFGTSQDALDFYGGVLRQLTFFSPRRQ
AQH
TTSPITSSQGDHWNIEICVSKEAGIETPTQSEMDRMEKFDPSK

Protein: noname
Length:  370
N-terminus: OUT
Number of transmembrane helices: 7
Transmembrane helices: 12-30 37-55 78-94 117-137 160-184 215-239 281-300

Total entropy of the model:  17.0094
Entropy of the best path:  17.0133

The best path:

     seq  TCASTEGMKD NAYGPLGIVA IILAASVLPW HIKAKNTGVL ILSAWLIVSV    50
     pred OOOOOooooo oHHHHHHHHH HHHHHHHHHH iiiiiiHHHH HHHHHHHHHH

     seq  SCLTINAFIW WRNVDIRCQI WCDISTKLIF GAYMGLPCSS ICIIRQLEEI   100
     pred HHHHHooooo oooooooooo oooooooHHH HHHHHHHHHH HHHHiiiiii

     seq  GSTRRVRITA KDRKHKLYFD LGVGVAIPLL YMALHIVNQG HRFDVIESVG   150
     pred iiiiiiiiii iiiiiiHHHH HHHHHHHHHH HHHHHHHooo oooooooooo

     seq  CFPTYYLTPV AIVLILVPPV LASSVALIYS FLALRWFFIR RRQFQAVLQN   200
     pred ooooooooooH HHHHHHHHHH HHHHHHHHHH HHHHiiiiii iiiiiiiiii

     seq  SHSGLNRSRY LRLMAMAGTE ALWSFPVNVT ILVTKFTLLK LPTYPYISWE   250
     pred iiiiiiiiii iiiiHHHHHH HHHHHHHHHH HHHHHHHHHo oooooooooo

     seq  DTHFNFGRAD QFDSTFFDSD PDLKAYWRSN IALGRYGVLV GCFFLFVFFG   300
     pred ooooOOOOOO OOOOOooooo oooooooooo HHHHHHHHHH HHHHHHHHHH

     seq  TSQDALDFYG GVLRQLTFFS PRRQAQHTTS PITSSQGDHW NIEICVSKEA   350
     pred iiiiiiiiii iiiiiIIIII IIIIIIIIII IIIIIIIIII IIIIIIIIII

     seq  GIETPTQSEM DRMEKFDPSK   370
     pred IIIIIIIIII IIIIIIIIII


If you are going to use these results in your work, please cite:

    G.E Tusnády and I. Simon (1998) J. Mol. Biol. 283, 489-506.
    G.E Tusnády and I. Simon (2001) Bioinformatics 17, 849-850.
```

```
      >Tco_STE3_ Tco_g7596 extraction
MKDNAYGPLGIVAIILAASVLPWHIKAKNTGVLILSAWLIVSVSCLTINAFIWWRNVDIRCQIWCDISTKLIFGAYM
GLP
CSSICIIRQLEEIGSTRRVRITAKDRKHQLYFDLGVGVAIPVLYMALHIVNQGHRFDIFESVGCYPTFYMTPVALVL
ILI
PPVVASVVALVYSFLALRWFVIRRRQFNAVLQNSHSGLNRSRYLRLMAMSSTEALWSFPVNVTVLASKFTLQHQPVF
PYI
SWEDTHYNFSAIGQYPAAFWDSSPDLQGYWRASVSLGRYGVLVGCFFLFVFFGTSQDALKFYGAVLRKITSIFSR

Protein: noname
Length:  331
N-terminus: OUT
Number of transmembrane helices: 7
Transmembrane helices: 21-39 46-64 87-103 126-146 169-193 224-243 290-310

Total entropy of the model:  17.0081
Entropy of the best path:  17.0121

The best path:

     seq  TCSTETCGET RACTINMKDN AYGPLGIVAI ILAASVLPWH IKAKNTGVLI    50
     pred OOOOOooooo oooooooooo HHHHHHHHHH HHHHHHHHHi iiiiiHHHHH

     seq  LSAWLIVSVS CLTINAFIWW RNVDIRCQIW CDISTKLIFG AYMGLPCSSI   100
     pred HHHHHHHHHH HHHHoooooo oooooooooo ooooooHHHH HHHHHHHHHH

     seq  CIIRQLEEIG STRRVRITAK DRKHQLYFDL GVGVAIPVLY MALHIVNQGH   150
     pred HHHiiiiiii iiiiiiiiii iiiiiHHHHH HHHHHHHHHH HHHHHHoooo

     seq  RFDIFESVGC YPTFYMTPVA LVLILIPPVV ASVVALVYSF LALRWFVIRR   200
     pred oooooooooo ooooooooHH HHHHHHHHHH HHHHHHHHHH HHHiiiiiii

     seq  RQFNAVLQNS HSGLNRSRYL RLMAMSSTEA LWSFPVNVTV LASKFTLQHQ   250
     pred iiiiiiiiii iiiiiiiiii iiiHHHHHHH HHHHHHHHHH HHHooooooo

     seq  PVFPYISWED THYNFSAIGQ YPAAFWDSSP DLQGYWRASV SLGRYGVLVG   300
     pred ooooooooOO OOOOOOOOOO OOOOoooooo ooooooooooH HHHHHHHHHH

     seq  CFFLFVFFGT SQDALKFYGA VLRKITSIFS R  331
     pred HHHHHHHHHH iiiiiiiiii iiiiiIIIII I



If you are going to use these results in your work, please cite:

    G.E Tusnády and I. Simon (1998) J. Mol. Biol. 283, 489-506.
    G.E Tusnády and I. Simon (2001) Bioinformatics 17, 849-850.

>Tin_STE3_ Tin_g7227
MKDHAYGPLGIVAVILAASVLPWHIKAKNSGVLILCSWLIISVACLTINSFIWWNDTEIRCQIWCDISTKITFGAFM
GLP
CSSVCILRQLEEIGSTRRVRSTAKDRKHQLYFDLGVGVGIPLVYMILHIVNQGHRFDISESIGCFPTYYLTPVAIVL
VLV
```

PPVVASTVALIYSFLALRWFVIRRRQFNAVLQNSHTGLNRNRYLRLMAMAGTEALWSFPINVTILVSKFTLQDKPIH
PYL
SWSDTHFQFSRIDQYTSAFWDSAPGLKAYWKASLCLGRYGALVGCFFLFVFFGTGNDALKFYGAVLRKILPIPSRTQ
ATR
KPTMSSSQDHWNIEVCVSKEAGLVTPTRSEMDQLEKVDSLS

Protein: noname
Length: 371
N-terminus: OUT
Number of transmembrane helices: 7
Transmembrane helices: 17-33 40-58 81-97 120-140 163-185 216-236 289-305

Total entropy of the model: 17.0085
Entropy of the best path: 17.0116

The best path:

```
     seq  TINSTETING MKDHAYGPLG IVAVILAASV LPWHIKAKNS GVLILCSWLI    50
     pred Ooooooooo oooooHHHH HHHHHHHHH HHHiiiiiiH HHHHHHHHHH

     seq  ISVACLTINS FIWWNDTEIR CQIWCDISTK ITFGAFMGLP CSSVCILRQL   100
     pred HHHHHHHHoo ooooooooo ooooooooo HHHHHHHHH HHHHHHHiii

     seq  EEIGSTRRVR STAKDRKHQL YFDLGVGVGI PLVYMILHIV NQGHRFDISE   150
     pred iiiiiiiii iiiiiiiiiH HHHHHHHHH HHHHHHHHH ooooooooo

     seq  SIGCFPTYYL TPVAIVLVLV PPVVASTVAL IYSFLALRWF VIRRRQFNAV   200
     pred ooooooooo ooHHHHHHH HHHHHHHHH HHHHHiiiii iiiiiiiii

     seq  LQNSHTGLNR NRYLRLMAMA GTEALWSFPI NVTILVSKFT LQDKPIHPYL   250
     pred iiiiiiiii iiiiiHHHH HHHHHHHHH HHHHHHoooo ooooooooo

     seq  SWSDTHFQFS RIDQYTSAFW DSAPGLKAYW KASLCLGRYG ALVGCFFLFV   300
     pred oOOOOOOOO OOOOOOOOO OOOooooooo ooooooooHH HHHHHHHHH

     seq  FFGTGNDALK FYGAVLRKIL PIPSRTQATR KPTMSSSQDH WNIEVCVSKE   350
     pred HHHHHiiiii iiiiiiiii IIIIIIIII IIIIIIIII IIIIIIIII

     seq  AGLVTPTRSE MDQLEKVDSL S  371
     pred IIIIIIIII IIIIIIIII I
```

If you are going to use these results in your work, please cite:

    G.E Tusnády and I. Simon (1998) J. Mol. Biol. 283, 489-506.
    G.E Tusnády and I. Simon (2001) Bioinformatics 17, 849-850.

>Twa_STE3 Twa_g7605 (reversed) extraction
MKDNAYGPLGIVAVILAASVLPWHIRAKNSGVLILSSWLIISVACLTINAFIWWSDTEIRCQIWCDISTKIAFGAFM
GLP
CSSICILRQLEEIGSTRRVRTTAKDRKHQLYFDLGVGVVIPVIYMILHIVNQGHRFDIFESVGCFPTYYLTPLAIVL
VSV

PPVVASAVALIYSFLALRWFVIRRRQFNTVLQNSHSGLNRSRYLRLMAMAGTEALWSFPINVIVLVSKYTLLKLPVY
RYV
SWSDTHFDFGRIDRFPSTFWDASPNLKAYWVASLNLGRYGPLVGCLFLFFFFGTSRDALKWYGALLRKICPMPSRNQ
AKR
NIPGASQEDHWNIEVCVSKEAGLATPTHSEMDQLDKVDSLN

Protein: noname
Length: 387
N-terminus: OUT
Number of transmembrane helices: 7
Transmembrane helices: 31-49 56-80 97-113 136-156 179-203 234-253 296-320

Total entropy of the model: 17.0095
Entropy of the best path: 17.0128

The best path:

```
     seq  TWASTETWAG REVERSEDET RACTINMKDN AYGPLGIVAV ILAASVLPWH    50
     pred OOOOOOOOOO OOOOOooooo oooooooooo HHHHHHHHHH HHHHHHHHHi

     seq  IRAKNSGVLI LSSWLIISVA CLTINAFIWW SDTEIRCQIW CDISTKIAFG   100
     pred iiiiiHHHHH HHHHHHHHHH HHHHHHHHHH oooooooooo ooooooHHHH

     seq  AFMGLPCSSI CILRQLEEIG STRRVRTTAK DRKHQLYFDL GVGVVIPVIY   150
     pred HHHHHHHHHH HHHiiiiiii iiiiiiiiii iiiiiHHHHH HHHHHHHHHH

     seq  MILHIVNQGH RFDIFESVGC FPTYYLTPLA IVLVSVPPVV ASAVALIYSF   200
     pred HHHHHHoooo oooooooooo ooooooooHH HHHHHHHHHH HHHHHHHHHH

     seq  LALRWFVIRR RQFNTVLQNS HSGLNRSRYL RLMAMAGTEA LWSFPINVIV   250
     pred HHHiiiiiii iiiiiiiiii iiiiiiiiii iiiHHHHHHH HHHHHHHHHH

     seq  LVSKYTLLKL PVYRYVSWSD THFDFGRIDR FPSTFWDASP NLKAYWVASL   300
     pred HHHooooooo ooooooooOO OOOOOOOOOO oooooooooo oooooHHHHH

     seq  NLGRYGPLVG CLFLFFFFGT SRDALKWYGA LLRKICPMPS RNQAKRNIPG   350
     pred HHHHHHHHHH HHHHHHHHHH iiiiiiiiii iiiiiIIIII IIIIIIIIII

     seq  ASQEDHWNIE VCVSKEAGLA TPTHSEMDQL DKVDSLN  387
     pred IIIIIIIIII IIIIIIIIII IIIIIIIIII IIIIIII
```

If you are going to use these results in your work, please cite:

G.E Tusnády and I. Simon (1998) J. Mol. Biol. 283, 489-506.
G.E Tusnády and I. Simon (2001) Bioinformatics 17, 849-850.

FASTA ANNOTATIONS

>Tco_g4688_HDP
MQASTGASIWANRGSSSSSSDITSSAHERDRNPPNHSANSSYVPRSIPSIPGVDLSGASSAS
SRPFPPITTGSSNTSSSMSSFYPADFAHSAPGSASGSTNSAYPPRAGAQEGSRPYTMASYL
GPSGGGGEGPAGAGSGNTSGPSSSSYHGSGSEYVSGGPSYQSAAGPGNGSSAAVLAGYR
PATSHETFSGANYSYPSSRSSFSAYGSAPVTGAYGSSGVPSSSSTAAPYHHGGYPHPAGY
GAPAGPGGSGPSGIPSEGYPSYPPQTSNTSSDPSATYNSGVPGYAHPNSAGYTYGGSQPG
FYPNPHLGVNAAAAPVPGAPGAYGGFSHPAGPAIHYLNPFEVKHRRRTTKTQFRVLEGT
FKENPKPNAAVRKALAQQLDMPGRAVQIWFQNRRAKAKAQAKKEEQGTRERAATSEG
STSTSSKTQQQPQQHSRSVTAPQPMAGGGSSTDSYSTRSSATSHGNGGGGSGGGGSGDY
HRHSQAQQQGPAGYSQAGHHASSHHGPSGSGAWHASHSSASRN
>Tca_g5325_HDP
MQASTGASIWANRGSSSSSSDITSSAHERDRNPPNHSANSSYVPRSIPSIPGVDLSGASSAS
SRPFPPITTGSSNTSSSMSSFYPADFAHSAPGSASGSTNSAYPPRAGAQEGSRPYTMASYL
GPSGGGGEGPAGAGSGNTSGPSSSSYHGSGSEYVSGGPSYQSAAGPGNGSSAAVLAGYR
PATSHETFSGANYSYPSSRSSFSAYGSAPVTGTYGSSGVPSSSSTAAPYHHGGYPHPAGY
GAPAGPGGSGPSGIPSEGYPSYPPQTSNTSSDPSATYNSGVPGYAHPNSAGYTYGGSQPG
FYPNPHLGVNAAAAPVPGAPGAYGGFSHPAGPAIHYLNPFEVKHRRRTTKTQFRVLEGT
FKENPKPNAAVRKALAQQLDMPGRAVQIWFQNRRAKAKAQAKKEEQGTRERAATSEG
STSTSSKTQQQPQQHSRSVTAPQPMAGGGSSTDSYSTRSSATSHGNGGGGSGGGGSGDY
HRHSQAQQQGPAGYSQAGHHASSHHGPSGSGAWHASHSSASRN
>Tc517_g394_HDP
MQASTGASIWANRGSSSSSSDITSSAHERDRNPPNHSANSSYVPRSIPSIPGVDLSGASSAS
SRPFPPITTGSSNTSSSMSSFYPADFAHSAPGSASGSTNSAYPPRAGAQEGSRPYTMASYL
GPSGGGGEGPAGAGSGNTSGPSSSSYHGSGSEYVSGGPSYQSAAGPGNGSSAAVLAGYR
PATSHETFSGANYSYPSSRSSFSAYGSAPVTGTYGSSGVPSSSSTAAPYHHGGYPHPAGY
GAPAGPGGSGPSGIPSEGYPSYPPQTSNTSSDPSATYNSGVPGYAHPNSAGYTYGGSQPG
FYPNPHLGVNAAAAPVPGAPGAYGGFSHPAGPAIHYLNPFEVKHRRRTTKTQFRVLEGT
FKENPKPNAAVRKALAQQLDMPGRAVQIWFQNRRAKAKAQAKKEEQGTRERAATSEG
STSTSSKTQQQPQQHSRSVTAPQPMAGGGSSTDSYSTRSSATSHGNGGGGSGGGGSGDY
HRHSQAQQQGPAGYSQAGHHASSHHGPSGSGAWHASHSSASRN
>Twa_g3288_HDP
MFATPIRTSRTSFFSVPSIASLDAPAINSTYTTATGSSTSSSPSLGYPSMQDNASAGIWANR
GSATSSNTTTSVHERERTNTSSAYIPRSIPPVPGVDLSGTSSASSRFFPPITAGSNNSSSSMS
SFYPADFAHSAPGSASGSGSTGSAYPNPARAGPEGSRPYTMAGYLGPSASGGDGPAGTG
SGSASGPSSSTYHGGPEYSSGPHSYATGPQGNTSVLTGYRPATSHETFSGANYSYPSSRSS
FSTYGSAPGSGSYNSSAVQSSSSTATPYHHSAYQTHPAYGAPTGAGPSGIQPEGYPSYQP
PQTANSSSDPSSTYSSGVPGYANPTSGSYTYGSQPGVYPSSHIGQNAAGGPVSAAPGAY
NGGFNHPAGPAIHYLNPFEVKHRRRTTKTQFRVLEGTFKENPKPNAAVRKALAQQLDM
PGRAVQIWFQNRRAKAKAQAKKEEATTRERAATSEGSTSTTTSKAQQHTRSTTAPQPAT
GGSSSTDSYSTRSSASHGNGAGAGASSGGGGSGDYHRHSQAQQPQGSAGYGSQSQGHG
HHGPSHHGSGGSGGSGGWHNSHMSSSRN
>Tin_g6649_HDP

MFATPIRTSRTSFFSVPSIASLDAPAINSTYTTAAGSSTSSSPSLGYPSMQDNASAGIWANR
GSATSSNTTTSVLERERTNTSSAYIPRSIPPVPGVDLSGTSSASSRFFPPITAGSNNSSSSMS
SFYPADFAHSAPGSASGSGSTGSAYPNPARAGPEGSRPYTMAGYLGPSASGGDGHTGTG
SGSASGPSSSTYHGGPEYSSGPHSYATGPQGNTSVLTGYRPATSHETFNGANYSYPSSRS
SFSTYGSAPGSGSYNSSAVQSSSSTAAPYHHSAYQTHPAYGAPTGAGPSGIQPEGYPSYQ
PPQTANSSSDPSSTYSSGVPGYANPTSGSYTYGSQPGVDPSSHLGQNAAGGPVPAAPGA
YNGGFNHPAGPAIHYLNPFEVKHRRRTTKTQFRVLEGTFKENPKPNAAVRKALAQQLD
MPGRAVQIWFQNRRAKAKAQAKKEEATTRERAATSEGSTSTTTSKAQQHTRSTTAPQP
ATGGSSSTDSYSTRSSASHGNGAGAGASSGGGGSGDYHRHSQAQQPQGSAGYGSQSQG
HGHHGPSHHGSGGSGGSGGWHNSHMSSSRN
>Tc517_g1873_HDP
MSKHPEQPEPQDVNVAYNQQQRRYSDTTDGALKRGFDEVLDDRHQRATSKGIKLPSLS
DLALGAGQWAASSRSKRPTLGSAALLPSFAGTRRQSSSQHTAAQFPTRSASDSKNMYA
VNANHASSAISRMVRNDPVADYEENDAGPSKRRRSPTSYYEDHPMDSSSARTAHDTTP
AVLHSGTPSVLERSGWSYGSSESYDWRGNSSRQTLRTSPSINQMKRKFSQGDMSKSSEG
NRSTLASGLPSTSSSFSRILAMGSPPLGTFPDESRRIPSRYSDVRTTSAINSSSTSGSRRASP
GTSQPFSRYEPLSLGGCPGQRHPSQTSPVLIQSHRSSNEANQSYQQPLRAFSGHSSEAQQ
QLSPSTSRDGSDLAYHRHSSDYREERQYSSSSKSSATSPASLASPRVDCDSFLVNRSDSID
SGIVSRMSGLAHDEAERRGYHAHEAVTDTLAYHHRRMHTSTEPSSTALGSWPPLTMLP
LASTSYYHTSGAAPQSTLHSQTSPNFGPTSATPMPPLSSPTALVLSTGHAEVESDHNQMY
KNSSSSGSHHSTNTSPQANQTDPDSDGGPRRGNFPRPVTEYLRQWLFNHRDHPYPSDQE
KREMTQITKLDLRQLNNWFINARRRVLRHGDAQREQLSGHQGAVAAAEESLD
>Tca_g10368_HDP
MSKHPEQPEPQDVNVAYNQQQRRYSDTTDGALKRGFDEVLDDRHQRATSKGIKLPSLS
DLALGAGQWAASSRSKRPTLGSAALLPSFAGTRRQSSSQHTAAQFPTRSASDSKNMYA
VNANHASSAISRMVRNDPVADYEENDAGPSKRRRSPTSYYEDHPMDSSSARTAHDTTP
AVLHSGTPSVLERSGWSYGSSESYDWRGNSSRQTLRTSPSINQMKRKFSQGDMSKSSEG
NRSTLASGLPSTSSSFSRILAMGSPPLGTFPDESRRIPSRYSDVRTTSAINSSSTSGSRRASP
GTSQPFSRYEPLSLGGCPGQRHPSQTSPVLIQSHRSSNEANQSYQQPLRAFSGHSSEAQQ
QLSPSTSRDGSDLAYHRHSSDYREERQYSSSSKSSATSPASLASPRVDCDSFLVNRSDSID
SGIVSRMSGLAHDEAERRGYHAHEAVTDTLAYHHRRMHTSTEPSSTALGSWPPLTMLP
LASTSYYHTSGAAPQSTLHSQTSPNFGPTSATPMPPLSSPTALVLSTGHAEVESDHNQMY
KNSSSSGSHHSTNTSPQANQTDPDSDGGPRRGNFPRPVTEYLRQWLFNHRDHPYPSDQE
KREMTQITKLDLRQLNNWFINARRRVLRHGDAQREQLSGHQGAVAAAEESLD
>Tco_g8961_HDP
MSKHPEQPEPQDVNVAYNQQQRRYSDTTDGALKRGFDEVLDDRHQRATSKGIKLPSLS
DLALGAGQWAASSRSKRPTLGSAALLPSFAGTRRQSSSQHTAAQFPTRSASDSKNMYA
VNANHASSAISRMVRNDPVADYEENDAGPSKRRRSPTSYYEDHPMDSSSARTAHDTTP
AVLHSGTPSVLERSGWSYGSSESYDWRGNSSRQTLRTSPSINQMKRKFSQGDMSKSSEG
NRSTLASGLPSTSSSFSRILAMGSPPLGTFPDESRRIPSRYSDVRTTSAINSSSTSGSRRASP
GTSQPFSRYEPLSLGGCPGQRHPSQTSPVLIQSHRSSNEANQSYQQPLRAFSGHSSEAQQ
QLSPSTSRDGSDLAYHRHSSDYREERQYSSSSKSSATSPASLASPRVDCDSFLVNRSDSID
SGIVSRMSGLAHDEAERRGYHAHEAVTDTLAYHHRRMHTSTEPSSTALGSWPPLTMLP
LASTSYYHTSGAAPQSTLHSQTSPNFGPTSATPMPPLSSPTALVLSTGHAEVESDHNQMY

KNSSSSGSHHSTNTSPQANQTDPDSDGGPRRGNFPRPVTEYLRQWLFNHRDHPYPSDQE
KREMTQITKLDLRQLNNWFINARRRVLRHGDAQREQLSGHQGAVAAAEESLD
>Tin_g1181_HDP
MSKHPEQVNSAQGTTTFPHERRESRGSNGSLKRVYEDYDNDGHQQSTSKGVTLPSCSD
LALGAEQWVASGSSRPSSQGHPALLPSIISAADHSTPFYQSSYRRPSSHFLTRSPSEPENSY
AGHAQHASSTISRMVRTNTVPEYEEDDNVPSKRRRSSRESMRDSMADSFSTRTALNFPA
AVVRDESSSAPERSAWNHRTSDGFDWRSSSSRSYLRVSPNTSQMQRRFSREDTSRYSEG
FRPTLPSTTLFPSGLPTKHSPPSGAFADESRRITGKYSDLRTTSNSNNSSGDSRRASPDAPQ
PFNRYEPLSLSGCPAQRYSGRTSPAAAQTYRSSEAVDRQQPRSTGGHSSEHHAHISPSTS
RAGSDRDYDQRSPGYHEQHHFPSSSASGATSPPSLPSPHVDPSSFSRYRSGSMDSGRQLS
SLFVSHLSGLAHDEGGRRIRHVYDPPADASVNHASHEQYQQQQVRPSTNHSSNAHRSW
SYVNAPTAASASYHNNSGPTSNSAFRSQESSNPVMSSARILTHLTYPDHAAPHAQEPHG
MSNSSSKGSIGTNDDSQANQLEPDLDSDNGTVPRRGNFPRQVTEYLRQWLFTHRDHPYP
SDQEKRDMAQYTNLDLRQLNNWFINARRRVLRNAEDDAQQQGGVATAEE
>Twa_g197_HDP
MSKHPEQVNSPQGTTTFPHERRESRGSNGSLKRVYEDYDNDGHQQSTFKGVTLPSCSDL
ALGAEQWVASGSSRHSSQGHSALLPSIVSAADHSPPFYQSSNRRPSSHFLTRSPSDPENSY
AGHAQHASSTISRMVRTNTVPEYEEDDNVPSKRRRSSRESMRDSMADSFSTRTASNFPA
AVVRDESSSAPERSAWNHRTSDGFDWRSSSSRSYLRVSPNTSQMQRRFSREDTSRYSEG
FRPMLPSNTLFPSGLPTKHSPPSGAFADESRRITGKYSDLRTTSNSHSNNNSSGDSRRASP
DAPQPFNRYEPLSLSGCPAQRYSGRTSPAAAQAYRSSEAADRQQPRSTGGHSSEHHAHIS
PSTSRAGSDRDYDQRSPGYHEQHHFPSSSASGATSPPSLPSPHGDPSSFSRYRSGSMDSGR
QLSSLFVSHLSGLAHDEGGRRIRHVYDPPADASVNHASHEQYQQQQVRPSTNHSSNAH
RSWSYVNALPAASASYHNNSGPTSNSAFRSQESSNPVMSSARILTHLTYPDHAAPHAQE
PHGMSNSSSKGSIGTNDDSQANQLEPDLDSDNGTVPRRGNFPRQVTEYLRQWLFTHRD
HPYPSDQEKRDMAQYTNLDLRQLNNWFINARRRVLRNAEDDAQQQGGVATAEE
>Tca_bE1
MTMDDIAFLRQADTEILLALSNGDEGIIDFNRRFTQALEAVQNLSIDQRPNSVQIKMIIRN
VRDTVAYLLEFEDKYQQKVRWVHSKTTLTIRDWVRSKAIQAIREEGAAVAATDIQANL
TPTTDSQTNITPATDIEANPTSFKFFNTQILRDWFVGNLDNPFPDTKTKDDLVAETNALP
GDGAGGRASKPMNTIQLQNWFGNMRRRSGWLSFKKTYARNNSVAFQGLIDALRRQGQ
APPAGVMDDVAELMSAWNVTMMVEKDEVDDDDEEGQNQQARTKTDSKGKEASASRS
NQSTANPRSSVNPATADPTGAIRAKACRDAYSKMMGHIAGKYTEEGAAVAATETEANP
TSSKVFNTKILRDWFVDNLVNPFLTPRPRMTSLRKRMRCPVTEMVAEHQSR
>Tco_bE2
MDDIAFFLQADTQVLAISSREDEIVDFNRQLTQAQEVVENLSIDQGLGAVQIEISVRYVR
DSIAFFVGVEGRYQQKLQCIRTKTALAIRDWVRAEATLAIREETAAVTATSIQANPTPAT
RSEANNTPETGSQASIITSGVFNAQFMRQWFAKSIVHPYFRIKMMQAGVAIANAFPSGEL
GGREQRPINYKRLWTNILRHQSGWTSFKNIYARNDRVSLEALISVLRRQGQAPPAGEMD
DIAMLMSAWDVPMMVEGEGADDIDDYDDDEEQQQQVRANIEKRSRGEETLVSRSSQS
TATPRSSVNPAAADPIGTIRAKACRNAYSKMMDSISAGSKTRANDWLDEISETVFEGEEE
KSKIVAQKGPVALQTRASERQDPLFDHSANSTPSQSNASSSSASIDELPSSGSIVLSPRKA
ARLPGSELGPRLPSTVASESTLVPTNFWQASLASNPSLSSFITPSANTSLPAASMNFTLSM
PLHGFDNGVPAQLRNSSIDSSFSGITDITQSSQSTDSSATTQSSQASYAYFASGMGMGDSS

LVRSNCPSEASPSVGQSWSPAAARHGEVRARDDNSVESAPKRFRAGNRPRHTDGGAPM
NMTNLSMMPVDQAPCTARTMNPQPGMGGPSTQ
>Tin_bE
MDAVSFLRQADTHLFRVISDGRGIQDYLKRIEQVREIVDKNGVPGPDAVQVWASLQCV
RDSAAFFLGIESISRQIADSTASEMSSALQSKGSSASETIISADVTPSDSLNAQVLRRWFAD
NIVDPSPDTKTKHTLVARTNALPGGGVGGRAQKPIDYKQVMLWFINMRRRSGWTSFYK
TFARKDKAKLERLVNTLKRQGQVPPPGVKDDIAALMVSWTVPMVSEEEDDEDDDEEE
QQQKGRGNKGSKAKKAPASKSRRSNKPITIPLSSQNPALADPTGTLRAKACRSAYSKM
MDFISAGAKERVGDWLDEVLEGVEEKPKIVKQTEAVSFSKVSSSSMSTGRKIAALPSRA
SKRQTSSHFAFPEASTSSASIVELPSSSVSSFVPRKSARLPGSELGPRLPSTLASDTPLVPTN
FWQSSLSSNSSSGSFIPPSIGSSLPTASMSLTLSMPFHGLNNGVPAPPRNASGQSTFSSITDI
SHSSQSTDSSATTQSSQPSCAYFASGVGMGDSSLVRSACSSEASPVMGSESPTRQVSSPA
GIKSRSTCDDGMLESTPKRFRPGNRPRHTDGGAPTSITNLPMMTLHQTPFPVSTMNLNLS
LSMNGEAGIIPFGAPQQTYYPGMSTTTTNVPMPQVPAMAAADLAAWYYQTVVQQNGH
QMPGTALYSSTMPGTSAMGMRSTSGPERFMESQSHGSSGPSFYGPGHP
>Twa_bE
MDAVSFLRQADTHLFQVLSDGRGIEDFRRRIEQVRELVEKTGVSGLDSFQIWTNIQCVR
DSAAFFLGIERISRQIADRSASEMSSALQNQALSAPETIVPADVTPSDSLNAQVLRRWFA
DNIVDPSPDTKTKHTLVARTNALPGGGVGGRAQKPIDYKQVMLWFINMRRRSGWTSFY
KTFARKDKAKLERLVNTLRRQGQAPPPGVKDDIAALMAAWTVPMVSEEEDDEDDDEE
EQQPKGRGNKGSKAKKAPASKSRRSNKPITIPLSSQNPALADPTGTLRAKACRSAYSKM
MDFISAGAKERVGDWLDEVLEGVEEKPKIVKQTEAVSFSKGSSSSMSTERKIAALPSRTS
KRQASSHFAFPEASTSGASIVELPSSSVSSFVPRKSARLPGSELGPRLPSTLASDTPLVPTN
FWQSSLSSNSSSGSFIHSSIGSSLPTASMSLTLSMPFHGLNNGVPAPPRNASGQSTFSSITDI
SHSGQSTDSSVTTQSSQPSCAYFASGVGMGDSSLVRSACSSEASPIVGSESPTRQVSSPAG
MKSRSACDDGMLESAPKRFRPGNRPRHTDGGAPTSITNLPMMTLHQTPFPVSTMNLNLS
LSMNGQAGIIPFGAPQQNYYPGMSTTNVPMPQVPAMAAADLAAWYYQTVVQQNGHQ
MPGTALYNSAMPGTSAVGMRSTSGPERFMEHHGSSGSSFYGPGVP
>Tc517_bE
MDTVSFLQEADTHLCRALSKPNGLEEFSRWSTEVREAVEKNGIQGLDSLHIRKVVQRVR
DSAAFFLGVEKINQRVGHWMLSQVTLTLREERTAVLATNAEIDITPSDSLNSKVLRQWF
ADNIVNPFLDTKTKNAVVAKTNALPGGGRGGRAQKPIDGRQVMLWFINMRRRSGWTS
FRKVYARKDTARLDRLISALRRQGEAPPAGVPDGLASLMCAWNVPMVVERAKKSNQS
VAIPLSTVNPATADPTGAIRAKACRNAYSKMMDFISAGAKERVGDWLDEILEGVEEKPK
IVAQTGPVTSLNVKSERKIAVLPGRASKRQNSLMDHSANPSSASIVELPSSGSIVLSPRKA
ARLPGSELGPRLPSTVASESTLVPTNFWQASLASNPSLSSFITPSASTSLPTASINFTLSMPL
HGFDNGVPAQPRNASSDSSFSGITDITQSSQSTDSSATTQSSQASYAYFASGMGMGDSSL
VRSNCPSEASPLVGQSSSPAAARHGEVRARDDNSVESAPKRFRAGNRPRHTDGGAPMS
MTNLSMMPVDQAPFAAPTMNLNLSLSMNPQPGMGPFGAPQQQFHPGMSTPVMQMPN
VPAAAAADLAAWYYQTLAQQHQQHQNEMPVTALHNSTMPPGSAGGMRSFSDPQRFT
DAHGPSVSSFNGLYMR
>Tco_bE1
MEKHGIQGFDSVQVQRNIQHVRDSAAFLLGVEDINQQTAHWIHSKVTLTFKRLEPASLA
TNSEVDLTPSDSLNSKVLRQWFADNIVNPFLDTKTKNAVVAKTNALPGGGRGGRAQKP

IDGRQVMLWFINMRRRSGWTSFRKVYARKDTARLDRLVSALRRQGEAPPAGVTDGLA
SLMCAWNVPMVVERAKKSNQSVAIPLSTVNPATADPTGAIRAKACRNAYSKMMDFISA
GAKERVGDWLDEILEGVEEKPKIVAQTGPVTSSNVKSERKIAVLPGRASKRQTSLMDHS
ANSSSASIVELPSSGSVVLSPRKAARLPGSEIGPRLPSTVASESTLVPTNFWQASLASNPSL
SSFITPSSNTSIPTASMNFTLSMPLHGFDNGVPAQPRNASSDSSFSGITDITQSSQSTDSSAT
TQSSQASYAYFASGMGMGDSSLVRSNCPSEASPLVGQSSSPAATRHGEVRARDDNTVES
APKRFRAGSRPRHTDGGAPMNMTNLSMMPVDQAPFVAPTMNLNLSLSMNPQPGMGPF
GAPQQQFHPGMSTPVMQMPNGPAAAAADLAAWYYQTLAQQHQQHQNEMSGAVLHS
STMAPGSAGGMRSFSDPQRFTDTHGPSVSSFNGPYMR
>Tca_bE2
MDSLSFLLQADTHLYRALAKGDGDINEFSQRMERAQKAMEKHGIQGFDSVQVQRNIQH
VRDSAAFLLGVEDINQQTAHWIHSKVTLTFKRLEPASLATNSEVDLTPSDSLNSKVLRQ
WFADNIVNPFLDTKTKNAVVAKTNALPGGGRGGRAQKPIDGRQVMLWFINMRRRSGW
TSFRKVYARKDTARLDRLVSALRRQGEAPPAGVTDGLASLMCAWNVPMVVERAKKSN
QSVAIPLSTVNPATADPTGAIRAKACRNAYSKMMDFISAGAKERVGDWLDEILEGVEEK
PKIVAQTGPVTSSNVKSERKIAVLPGRASKRQTSLMDHSANSSSASIVELPSSGSVVLSPR
KAARLPGSEIGPRLPSTVASESTLVPTNFWQASLASNPSLSSFITPSSNTSIPTASMNFTLS
MPLHGFDNGVPAQPRNASSDSSFSGITDITQSSQSTDSSATTQSSQASYAYFASGMGMG
DSSLVRSNCPSEASPLVGQSSSPAATRHGEVRARDDNTVESAPKRFRAGSRPRHTDGGA
PMNMTNLSMMPVDQAPFVAPTMNLNLSLSMNPQPGMGPFGAPQQQFHPGMSTPVMQ
MPNGPAAAAADLAAWYYQTLAQQHQQHQNEMSGAVLHSSTMAPGSAGGMRSFSDPQ
RFTDTHGPSVSSFNGPYMR
>Tc517_bW_1.2
MESFESATAQTLREISRHCDFLIQTFSATRRPSTPPTLDRPFADINLRDYISVKCSECSEAL
RRLRLSQYTSTELFGIYVHSVETFAAKASQMALAAYVPAGSSGTTSEIVVQEGLEVAVS
NSPAEDDDGDDPDDGTSSDDEEAAGAGDAQRYTPQTVAILERVFERTSSITHAEKNRIA
NAIGLTPRQVTVWFCNRRYRHSPSDSAPTPATKFESSKKRELDPHQAGDDTEEDVKSDG
ENVEELLLTTIQGGKKRLELDVSESAGEEEDDSDDDQDSGSFSSEDGAAETADARERPSD
VVALLDRALERNMNITQAEMNHIAKLINLTPQQVTVWFQKCRDQLKLSDGALTPAMQP
ESSKGRGLASQH
>Tco_bW_2
MESFESATAQTLREISRHCDFLIKTFSATHRPSTPPTLDRPFADINLRDYISVKCSECSEAL
RRLRLSQYMGTELFGIYVRSVETFAAKASQMALAAYVPAGSSGTTSEIVGALEAFLQRK
CDSFQELLLATVQEGLEVAVSNSPAEDDDGDDPDDGTSSDDEEAAGVGDAQRHTPHTV
AILERVFERTSSITHAEKNRIANAIGLTPRQVTVWFCNRRYRHNPSDAAPTPATKFESSKK
RVLAPHQAGDDTEEDVKSDGENVEELLLTTIQEGKKRLELDVSESAGVEDDDNDDDQD
SGTFSSDDDAAETADARERPSEVVALLDRALERNMNITQAEMNHIAKLINLTPQQVTVW
VSLIHAVCRMQPMTLLSSSISNIVPEMSRPSQAIRWSTHPSHAARV
>Tca_bW_1
MGTELFGIYVRSVETFAAKASQMALAAYVPAGSSGTTSEIVGALEAFLQRKCDSFQELL
LATVQEGLEVAVSNSPAEDDDGDDPDDGTSSDDEEAAGVGDAQRHTPHTVAILERVFE
RTSSITHAEKNRIANAIGLTPRQVTVWFCNRRYRHNPSDAAPTPATKFESSKKRVLAPHQ
AGDDTEEDVKSDGENVEELLLTTIQEGKKRLELDVSESAGVEDDDNDDDQDSGTFSSDD
DAAETADARERPSEVVALLDRALERNMNITQAEMNHIAKLINLTPQQVTVWVSLIHAVC

RMQPMTLLSSSISNIVPEMSRPSQAIRWSTHPSHAARV
>Tco_bW_1
MESLKSADSQYLLEIAQDCDFLIQRLSAQQGPLEPSNVDLPIAKIDIIGYGSAKFSEALQR
HRLSQSLSTKLLDIYVRAVENFAAEASRVASTAKIPAGPSGTTIGFANAVQAFAQRKCDD
LQETMLATLQEGKKRLEVEISESAAEDEDEDDDDASDDDSSSSSDDAADAGGARKHPP
QALRILELAFERTSNITHAEKNRIARAIGLTPRQVTVWFQNRRNRRKPESARTPVSQPESS
KKRKLATLQTADIDDEMEEVARVPTSLPEKRQRCSRQTSAASSSSISSVASFESNLGEWT
SHGAKYGRGPKPRNDSGDSSFTSVTSISSGSMDSSLADISQLSQESKGPPLRTEQPPMHDF
LFHPDGSIQPDWDDLELDLGQLEQSLAESFQMGASLVTSPSLAEAPVQPLTVSALRQIED
RTPTSGASFAEQQCLAAAHPSIAAFNSISQAENWSFLQTGVLGMPASQDPPTYSNDEWQ
SALLNLDCDVSDISSALSFVDEIRSNASQSGHSHQQDSASQLNAFLSQYGQDFGSSLSGH
SSGAESDDSKLHLDFDFDFSRLGFATSPVSTPSIASEDMRLNTIPEQADKQLC
>Tca_bW_2
MESLKSADSQYLLEIAQDCDFLIQRLSAQQGPLEPSNVDLPIAKIDIIGYGSAKFSEALQR
HRLSQSLSTKLLDIYVRAVENFAAEASRVASTAKIPAGPSGTTIGFANAVQAFAQRKCDD
LQETMLATLQEGKKRLEVEISESAAEDEDEDDDDASDDDSSSSSDDAADAGGARKHPP
QALRILELAFERTSNITHAEKNRIARAIGLTPRQVTVWFQNRRNRRKPESARTPVSQPESS
KKRKLATLQTADIDDEMEEVARVPTSLPEKRQRCSRQTSAASSSSISSVASFESNLGEWT
SHGAKYGRGPKPRNDSGDSSFTSVTSISSGSMDSSLADISQLSQESKGPPLRTEQPPMHDF
LFHPDGSIQPDWDDLELDLGQLEQSLAESFQMGASLVTSPSLAEAPVQPLTVSALRQIED
RTPTSGASFAEQQCLAAAHPSIAAFNSISQAENWSFLQTGVLGMPASQDPPTYSNDEWQ
SALLNLDCDVSDISSALSFVDEIRSNASQSGHSHQQDSASQLNAFLSQYGQDFGSSLSGH
SSGAESDDSKLHLDFDFDFSRLGFATSPVSTPSIASEDMRLNTIPEQADKQLC
>Tc517_bW_1.1
MPDVDLFTMTPIDISGYGTAEFSEALRQHRLSQSLSTELFDIYVRAVENFAAKASQVALM
AKIPAGHTVTTSDFIQTLQNFMQRKCDSLQKVMLATVQEGKKRLEVEVSESAAEDEDD
DDASDDDSSSSSDDAADAGGARKHPPQALRILELAFERTSNITHAEKNRIARAIGLTPRQ
VTVWFQNRRNRRKPESARTPASQPESSKKRKLATLQTADIDDEMEEVARVPSSLPEKRQ
RCSRQTSAASSSSISSVASFESNLGDWTSHGAKYGRAPKPRNDSGDSSFTSVTSISSGSMD
SSLADISQLSQESKGPPLRTEQPPMHDFLFHPDGSIQPDWDDLELDLGQLEQSLAESFQM
GASLVTSPSLAEAPVQPLTMSALRQIEDRTPTSGASFAEQQCLAAAHPSIAAFNSISQAEN
WSFLQTGVLDMPATQDPPTYSNDEWQSALLNLDCDVSDISSALSFVDEIRSNASQSGHS
HQQDSTSQLNAFLAQYGQDFGSSLSGHSSGAESDDSKLHLDFDFDFSRLGFATSPVSTPS
IASEDMRLNTIPEQADKQLC
>Twa_bW
MNSFDSDFSQWLREISRNCDDMTKSLLSNHPEPLRLKELDFQCAELEISGYGEVKMSHA
LQRHRISQSLSARLLSIYLRAVDKFSSETTRLVLKARIPAGPSVSQNDIVDTLQICMRRQC
DTLQDIVLAAVEDGKKRLEIDVSDSAADDDDGLDDDDSDSDDEAADLSGARKQHPPQA
LAILERAFERTNNITHAEKNRIAKAIGLTPRQVTVWFQNRRNRRKTSKPVATPASPSGSS
KKRKLASVNDEEMEDVTRTPSTLPEKRQRCTRSTSEASCISSSSSSCSVASFESEWTPQGI
KYGRTAKPRNESGDSSFTSITSVSSSSMESSVDSSQESKVEEPESDESLKAEEPLRAEQPA
VHDFLFHPDGSIQPDWDDLELDLGQLEQSLAESLAMGASLGSSPSPIDGTVQPLTASALQ
RMEDRTPTANTTFADQQRLATHPSIAVFTSISEAENWSFLESGVLGMPAEQDHSSYSND
DWQHALLNIDCDASDISSALSFMDEVLSKSESGQGHHQNESSQFEDYLSQLDKDSALSM

SSRSSTACTEDSTLHLDFDFDFSKLGFSTSPVGTPSTSSEAATFSTVPEAEDTPSC
>Tin_bW
MRWRCDSLQDALLAAVEDGKKRLEIDVSDSAADDDDESDEDGSDSDDEAVDPSGARK
QHPPQALAILERAFERTNNITHAEKNRIAKAIGLTPRQVTVWFQNRRNRRKISKQVATPA
SESGSSKKRKLASVNDEDMEDVARTPSTLPEKRQRCSRSASDASCISSSSSSCSVASFESE
WTPQGVKYGRTAKPRNESGDSSFTSITSISSSSMNSSVDSSQESKVEEPESYQSLKTAEPL
RAEQPAVHDFLFHPDGSIQPDWDDLELDLGQLEQSLAESLAMGASLGSSPSPIDGTVQPL
TASALQRMEDRTPTANTTFADQQRLAAHPSIAAFTSISQAENWSFLESGVLGMPADQGQ
SSYSSDDWQRALVNIDCDASDFSAALSFMDEVLSKSESGQGYHENDSSQFEDYLSQLDK
DSALSMSSRSSTACTEESKLHLDFDFDFSRLGFSTSPVGTPSTSSEATTFSTVPEAEDKLSC
>Tca_STE3 g9885
MKDNAYGPLGIVAIILAASVLPWHIKAKNTGVLILSAWLIVSVSCLTINAFIWWRNVDIR
CQIWCDISTKLIFGAYMGLPCSSICIIRQLEEIGSTRRVRITAKDRKHKLYFDLGVGVAIPL
LYMALHIVNQGHRFDVIESVGCFPTYYLTPVAIVLILVPPVLASSVALIYSFLALRWFFIR
RRQFQAVLQNSHSGLNRSRYLRLMAMAGTEALWSFPVNVTILVTKFTLLKLPTYPYISW
EDTHFNFGRADQFDSTFFDSDPDLKAYWRSNIALGRYGVLVGCFFLFVFFGTSQDALDF
YGGVLRQLTFFSPRRQAQHTTSPITSSQGDHWNIEICVSKEAGIETPTQSEMDRMEKFDP
SK
>Twa_STE3 Twa_g7605 (reversed) extraction
MKDNAYGPLGIVAVILAASVLPWHIRAKNSGVLILSSWLIISVACLTINAFIWWSDTEIRC
QIWCDISTKIAFGAFMGLPCSSICILRQLEEIGSTRRVRTTAKDRKHQLYFDLGVGVVIPVI
YMILHIVNQGHRFDIFESVGCFPTYYLTPLAIVLVSVPPVVASAVALIYSFLALRWFVIRR
RQFNTVLQNSHSGLNRSRYLRLMAMAGTEALWSFPINVIVLVSKYTLLKLPVYRYVSW
SDTHFDFGRIDRFPSTFWDASPNLKAYWVASLNLGRYGPLVGCLFLFFFFGTSRDALKW
YGALLRKICPMPSRNQAKRNIPGASQEDHWNIEVCVSKEAGLATPTHSEMDQLDKVDS
LN
>Tin_STE3_ Tin_g7227
MKDHAYGPLGIVAVILAASVLPWHIKAKNSGVLILCSWLIISVACLTINSFIWWNDTEIRC
QIWCDISTKITFGAFMGLPCSSVCILRQLEEIGSTRRVRSTAKDRKHQLYFDLGVGVGIPL
VYMILHIVNQGHRFDISESIGCFPTYYLTPVAIVLVLVPPVVASTVALIYSFLALRWFVIRR
RQFNAVLQNSHTGLNRNRYLRLMAMAGTEALWSFPINVTILVSKFTLQDKPIHPYLSWS
DTHFQFSRIDQYTSAFWDSAPGLKAYWKASLCLGRYGALVGCFFLFVFFGTGNDALKF
YGAVLRKILPIPSRTQATRKPTMSSSQDHWNIEVCVSKEAGLVTPTRSEMDQLEKVDSLS
>Tc517_STE3 g3435
MKDNAYGPLGIVAIILAASVLPWHIKAKNTGVLILSAWLIVSVSCLTINAFIWWRNVDIR
CQIWCDISTKLIFGAYMGLPCSSICIIRQLEEIGSTRRVRITAKDRKHKLYFDLGVGVAIPL
LYMALHIVNQGHRFDVIESVGCFPTYYITPVAIVLILVPPVVASAVALIYSFLALRWFIIRR
RQFNAVLQNSHSGLDRSRYLRLMTMASTEAIWSFPVNMTILVTKFTLIQEPINPYVSWQ
DTHFDFGRIGQFPAEFFDVSPDTKAYWRANMALGRYGALVGCFFLFVFFGTSQDALKF
YGTILRKITSIFPQFQAQRTMSPTTSSQADHWNIEVCVSKEAGVETPTQSEMDRMEKFEK
>Tco_STE3 g7596 translation
MKDNAYGPLGIVAIILAASVLPWHIKAKNTGVLILSAWLIVSVSCLTINAFIWWRNVDIR
CQIWCDISTKLIFGAYMGLPCSSICIIRQLEEIGSTRRVRITAKDRKHQLYFDLGVGVAIPV
LYMALHIVNQGHRFDIFESVGCYPTFYMTPVALVLILIPPVVASVVALVYSFLALRWFVI

RRRQFNAVLQNSHSGLNRSRYLRLMAMSSTEALWSFPVNVTVLASKFTLQHQPVFPYIS
WEDTHYNFSAIGQYPAAFWDSSPDLQGYWRASVSLGRYGVLVGCFFLFVFFGTSQDAL
KFYGAVLRKITSIFSRFQAQRTMSPTTSSQADHWNIEVCVSKEAGVETPTQSEMDRMEK
FEK
>Twa_mfa
MDSSLFNLPPSDEQVIQVSELVNMEHSASASGGCVIV
>Tin_mfa
MDSSLFNLPPSDEQVIQVGELVNQEHSNPVPNGGCVIV
>Tca_mfa
MDNSLFNLPASLFDVPHSDEQKVQGGELVNREHKLTPTGGCVIA
>Tc517_mfa
MDNSLFNLPSSLFDLPPSDEQKVKGVELVDMEHSKAPSGGCVIA
>Tco_mfa
MDNSLFNLPTSLFDLPPSNEQKVQVVALVDMEHSKNVPAGGCVIA

>Twa_g3288
ATGTTTGCCACCCCAATTCGAACATCCAGAACAAGTTTTTTCTCAGTCCCTTCCATCG
CTTCTCTAGACGCACCCGCCATAAACTCGACCTACACGACTGCCACAGGCTCTTCCA
CCTCTTCTTCCCCGTCCCTCGGCTACCCGTCGATGCAAGACAACGCAAGTGCAGGCA
TTTGGGCAAACCGTGGATCTGCTACTTCTTCAAACACTACTACTTCTGTTCACGAACG
CGAACGCACTAATACTTCCAGCGCTTACATCCCCCGTTCCATCCCCCCTGTCCCCGG
CGTCGACCTTTCGGGTACATCCTCTGCTTCGTCCCGATTCTTCCCACCCATCACTGCG
GGCAGTAACAACTCTTCTTCATCCATGTCGTCCTTCTACCCTGCCGATTTTGCTCACT
CCGCTCCCGGAAGCGCCAGCGGTTCGGGTTCGACCGGCAGTGCCTATCCAAACCCA
GCTCGCGCCGGTCCCGAAGGTTCACGGCCTTACACCATGGCAGGCTATCTCGGTCCC
TCCGCCAGTGGAGGCGATGGTCCCGCCGGTACAGGCTCAGGCAGTGCCTCCGGGCC
TTCCTCGTCAACCTATCATGGAGGCCCAGAGTACAGCTCTGGTCCGCACTCCTATGC
CACCGGCCCGCAAGGAAATACATCTGTCCTAACCGGCTACCGCCCAGCTACCAGTC
ACGAGACATTCAGTGGGGCTAATTACAGCTACCCTTCGTCGCGTTCTTCCTTCTCTAC
CTATGGGTCGGCTCCCGGCTCAGGTAGTTACAACTCTTCCGCTGTGCAATCCTCGTC
ATCGACTGCGACGCCTTATCACCACAGTGCCTACCAAACGCACCCAGCATATGGAG
CACCCACCGGTGCGGGTCCTTCTGGAATTCAGCCTGAGGGCTACCCGTCCTACCAGC
CCCCGCAGACAGCCAACTCGAGCTCGGACCCGTCCTCAACTTACAGCTCCGGTGTGC
CTGGATATGCAAACCCGACCTCGGGAAGCTACACGTATGGATCACAACCTGGGGTC
TATCCCAGCTCTCATATCGGACAAAACGCTGCCGGCGGACCTGTTTCTGCTGCGCCT
GGCGCATACAACGGTGGCTTCAACCATCCGGCTGGACCGGCCATCCATTACCTGAAC
CCCTTCGAAGTCAAGCATCGTCGCCGCACAACGAAGACTCAGTTCCGAGTACTGGA
GGGCACATTCAAGGAGAATCCCAAACCTAATGCAGCTGTGAGGAAGGCATTGGCGC
AGCAACTGGACATGCCCGGTCGAGCAGTGCAGATATGGTTCCAAAACCGACGAGCA
AAGGCAAAGGCGCAGGCGAAGAAGGAGGAGGCGACGACGAGAGAGCGAGCTGCG
ACATCTGAAGGGTCGACCAGCACTACCACCAGCAAGGCACAGCAACATACTCGATC
CACTACTGCACCCCAACCGGCGACAGGAGGCTCAAGTAGCACTGATAGCTACTCGA
CTAGATCCAGCGCGAGTCACGGTAATGGTGCGGGCGCGGGTGCGAGTTCTGGTGGT

GGTGGGAGTGGAGATTACCATCGCCACTCGCAGGCGCAGCAACCGCAAGGGTCGGC
CGGATATGGCTCGCAGAGCCAGGGTCACGGACACCATGGGCCTTCGCACCACGGTT
CGGGAGGCTCTGGAGGTTCCGGAGGATGGCATAACAGCCACATGTCTTCGTCGCGG
AACTGA
>Tin_g6649
ATGTTTGCCACCCCAATTCGAACATCCAGAACAAGTTTTTTCTCAGTCCCTTCCATCG
CTTCTCTAGACGCACCCGCCATAAACTCAACCTACACGACTGCCGCAGGCTCTTCAA
CCTCTTCTTCCCCGTCCCTCGGCTACCCGTCGATGCAAGACAACGCAAGTGCAGGCA
TTTGGGCAAACCGTGGATCTGCTACTTCTTCAAACACTACTACTTCTGTTCTCGAACG
CGAACGCACTAATACTTCCAGCGCTTACATTCCCCGTTCCATCCCCCCTGTCCCCGGC
GTCGACCTTTCGGGTACATCCTCTGCTTCGTCCCGATTCTTCCCACCCATCACTGCGG
GCAGTAACAACTCTTCTTCATCCATGTCGTCCTTCTACCCTGCCGATTTTGCTCACTC
CGCTCCCGGAAGCGCAAGCGGTTCGGGTTCGACGGGTAGTGCCTATCCAAACCCAG
CTCGCGCCGGTCCTGAGGGTTCACGGCCTTACACCATGGCAGGCTATCTCGGTCCCT
CCGCCAGTGGAGGCGATGGTCACACCGGTACAGGCTCAGGCAGTGCCTCCGGGCCT
TCCTCTTCAACCTATCATGGAGGCCCGGAGTACAGTTCTGGTCCGCACTCCTATGCC
ACCGGCCCACAAGGCAATACTTCTGTCCTAACCGGCTATCGCCCAGCTACCAGCCAC
GAGACATTCAATGGGGCTAATTACAGCTACCCTTCGTCGCGTTCCTCCTTCTCTACCT
ACGGATCGGCTCCCGGCTCAGGTAGTTACAACTCTTCCGCCGTGCAATCCTCGTCAT
CGACCGCGGCGCCTTATCACCACAGTGCCTACCAAACGCACCCAGCATACGGAGCA
CCCACCGGTGCGGGTCCTTCTGGAATCCAGCCTGAGGGCTACCCGTCCTACCAGCCC
CCGCAGACAGCCAACTCGAGCTCAGACCCGTCTTCGACTTACAGCTCCGGTGTGCCT
GGATATGCAAACCCGACCTCGGGAAGCTACACGTATGGATCACAACCTGGGGTCGA
TCCCAGCTCTCATCTCGGACAAAACGCTGCCGGCGGACCCGTCCCTGCTGCGCCTGG
CGCATACAACGGTGGCTTCAACCATCCGGCTGGACCGGCCATCCATTACCTGAACCC
TTTCGAAGTCAAGCATCGTCGCCGCACAACGAAGACTCAGTTCCGAGTACTGGAGG
GCACATTCAAGGAGAACCCTAAACCTAATGCGGCTGTGAGGAAGGCATTGGCGCAG
CAACTGGACATGCCCGGTCGAGCAGTGCAGATATGGTTCCAAAACCGACGAGCAAA
GGCAAAAGCACAGGCGAAGAAGGAGGAGGCGACGACGAGAGAGCGAGCTGCGAC
ATCCGAAGGGTCGACCAGCACTACCACCAGCAAGGCGCAGCAACATACTCGATCGA
CCACTGCACCCCAACCGGCGACAGGAGGCTCAAGTAGCACTGATAGCTACTCGACT
AGATCCAGCGCGAGTCACGGTAATGGTGCGGGCGCGGGTGCGAGTTCTGGTGGTGG
TGGGAGTGGAGATTACCATCGCCACTCGCAGGCGCAGCAACCGCAAGGGTCGGCCG
GATATGGCTCGCAGAGCCAGGGTCACGGACACCATGGGCCTTCGCACCACGGTTCG
GGAGGCTCTGGAGGTTCCGGAGGATGGCATAACAGCCACATGTCTTCGTCGCGGAA
CTGA
>Tco_g4688
ATGCAAGCCAGCACAGGTGCAAGCATTTGGGCAAATCGTGGATCTTCTTCTTCTTCC
TCCGACATTACTTCTTCTGCTCACGAACGGGATCGCAATCCCCCTAATCATTCTGCA
AACAGTTCGTACGTTCCCCGTTCCATTCCCTCCATCCCCGGCGTCGACCTTTCAGGTG
CATCTTCTGCTTCGTCCCGTCCGTTCCCACCCATCACCACCGGCAGCAGCAACACCT
CGTCATCTATGTCGTCCTTTTACCCCGCCGACTTCGCCCACTCGGCTCCCGGAAGCGC
GAGCGGGTCGACCAACAGTGCTTACCCTCCCCGAGCTGGTGCCCAGGAAGGATCAA
GACCGTACACCATGGCCAGTTATCTCGGACCATCGGGTGGTGGAGGTGAAGGTCCC

GCTGGAGCAGGATCTGGTAATACGTCCGGTCCTTCTTCGTCATCCTACCACGGTAGC
GGCTCCGAATACGTCTCCGGTGGACCATCCTACCAGTCTGCTGCCGGACCGGGAAAC
GGCTCGAGCGCAGCCGTGTTGGCCGGATACCGTCCAGCAACCAGCCACGAGACTTT
CAGTGGTGCGAATTACAGCTATCCTTCTTCACGGTCCTCATTCTCCGCCTACGGATCG
GCTCCGGTTACAGGTGCTTATGGCTCGTCCGGCGTGCCATCCTCATCATCGACTGCT
GCGCCTTATCATCATGGTGGCTACCCGCACCCTGCGGGATACGGCGCACCAGCAGG
GCCAGGAGGTTCCGGTCCGTCCGGCATCCCATCGGAAGGCTACCCTTCGTACCCTCC
GCAAACGTCAAACACCAGCTCGGACCCCTCAGCGACTTATAATTCAGGCGTTCCTGG
TTATGCTCATCCGAACTCGGCAGGCTATACATATGGCGGGTCACAACCTGGGTTTTA
CCCAAATCCTCATCTCGGGGTGAACGCTGCTGCTGCACCGGTACCAGGTGCACCTGG
TGCATACGGTGGCTTTAGCCACCCGGCTGGCCCGGCCATCCACTATCTGAATCCGTT
CGAGGTCAAGCACCGTCGTCGCACAACGAAGACTCAATTCCGCGTACTGGAGGGTA
CTTTTAAGGAGAATCCCAAGCCCAACGCTGCTGTCAGGAAGGCGTTGGCGCAACAG
TTGGACATGCCCGGCCGTGCAGTGCAGATTTGGTTCCAGAATCGTCGAGCCAAGGCC
AAGGCCCAGGCTAAGAAGGAAGAGCAAGGAACGAGAGAGCGAGCTGCGACATCGG
AAGGATCGACAAGCACCAGCAGCAAAACGCAGCAACAGCCGCAACAACATTCGCG
ATCTGTTACCGCACCGCAACCAATGGCAGGTGGAGGAAGTAGCACCGATAGCTACT
CAACCAGATCCTCAGCTACGAGCCACGGCAATGGTGGTGGAGGAAGCGGCGGCGGC
GGTAGCGGTGATTACCATCGGCATTCACAGGCGCAGCAGCAAGGGCCGGCTGGCTA
TAGCCAGGCAGGCCACCATGCCTCTTCGCATCATGGACCCAGTGGTTCAGGAGCGTG
GCATGCGAGCCACTCGTCTGCGTCACGGAATTGA
>Tca_g5325
ATGCAAGCCAGCACAGGTGCAAGCATTTGGGCAAATCGTGGATCTTCTTCTTCTTCC
TCCGACATTACTTCTTCTGCTCACGAACGGGATCGCAATCCCCCTAATCATTCTGCA
AACAGTTCGTACGTTCCCCGTTCCATTCCCTCCATCCCCGGCGTCGACCTTTCAGGTG
CATCTTCTGCTTCGTCCCGTCCGTTCCCACCCATCACCACCGGCAGCAGCAACACCT
CGTCATCTATGTCGTCCTTTTACCCCGCCGACTTCGCCCACTCGGCTCCCGGAAGCGC
GAGCGGGTCGACCAACAGTGCTTACCCTCCCCGAGCTGGTGCCCAGGAAGGATCAA
GACCGTACACCATGGCCAGTTATCTCGGACCATCGGGTGGTGGAGGTGAAGGTCCC
GCTGGAGCAGGATCTGGTAATACGTCCGGTCCTTCTTCGTCATCCTACCACGGTAGC
GGCTCCGAATACGTCTCCGGTGGACCATCCTACCAGTCTGCTGCCGGACCGGGAAAC
GGCTCGAGCGCAGCCGTGTTGGCCGGATACCGTCCAGCAACCAGCCACGAGACTTT
CAGTGGTGCGAATTACAGCTATCCTTCTTCACGGTCCTCATTCTCCGCCTACGGATCG
GCTCCGGTTACAGGTACTTATGGCTCGTCCGGCGTGCCATCCTCATCATCGACTGCT
GCGCCTTATCATCATGGTGGCTACCCGCACCCTGCGGGATACGGCGCACCAGCAGG
GCCAGGAGGTTCCGGTCCGTCCGGCATCCCATCGGAAGGCTACCCTTCGTACCCTCC
GCAAACGTCAAACACCAGCTCGGACCCCTCAGCGACTTATAATTCAGGCGTTCCTGG
TTATGCTCATCCGAACTCGGCAGGCTATACATATGGCGGGTCACAACCTGGGTTTTA
CCCAAATCCTCATCTCGGGGTGAACGCTGCTGCTGCGCCGGTACCAGGTGCACCTGG
TGCATACGGTGGCTTTAGCCACCCGGCTGGCCCGGCCATCCACTATCTGAATCCGTT
CGAGGTCAAGCACCGTCGTCGCACAACGAAGACTCAATTCCGCGTACTGGAGGGTA
CTTTTAAGGAGAATCCCAAGCCCAACGCTGCTGTCAGGAAGGCGTTGGCGCAACAG
TTGGACATGCCCGGCCGTGCAGTGCAGATTTGGTTCCAGAATCGTCGAGCCAAGGCC
AAGGCCCAGGCTAAGAAGGAAGAGCAAGGAACGAGAGAGCGAGCTGCGACATCGG

AAGGATCGACAAGCACCAGCAGCAAAACGCAGCAACAGCCGCAACAACATTCGCG
ATCTGTTACCGCACCGCAACCAATGGCAGGTGGAGGAAGTAGCACCGATAGCTACT
CAACCAGATCCTCAGCTACGAGCCACGGCAATGGTGGTGGAGGAAGCGGCGGCGGC
GGTAGCGGTGATTACCATCGGCATTCACAGGCGCAGCAGCAAGGGCCGGCTGGCTA
TAGCCAGGCAGGCCACCATGCCTCTTCGCATCATGGACCCAGTGGTTCAGGAGCGTG
GCATGCGAGCCACTCGTCTGCGTCACGGAATTGA
>Tc517_g394 32530_contig_
ATGCAAGCCAGCACAGGTGCAAGCATTTGGGCAAATCGTGGATCTTCTTCTTCTTCC
TCCGACATTACTTCTTCTGCTCACGAACGGGATCGCAATCCCCCTAATCATTCTGCA
AACAGTTCGTACGTTCCCCGTTCCATTCCCTCCATCCCCGGCGTCGACCTTTCAGGTG
CATCTTCTGCTTCGTCCCGTCCGTTCCCACCCATCACCACCGGCAGCAGCAACACCT
CGTCATCTATGTCGTCCTTTTACCCCGCCGACTTCGCCCACTCGGCTCCCGGAAGCGC
GAGCGGGTCGACCAACAGTGCTTACCCTCCCCGAGCTGGTGCCCAGGAAGGATCAA
GACCGTACACCATGGCCAGTTATCTCGGACCATCGGGTGGTGGAGGTGAAGGTCCC
GCTGGAGCAGGATCTGGTAATACGTCCGGTCCTTCTTCGTCATCCTACCACGGTAGC
GGCTCCGAATACGTCTCCGGTGGACCATCCTACCAGTCTGCTGCCGGACCGGGAAAC
GGCTCGAGCGCAGCCGTGTTGGCCGGATACCGTCCAGCAACCAGCCACGAGACTTT
CAGTGGTGCGAATTACAGCTATCCTTCTTCACGGTCCTCATTCTCCGCCTACGGATCG
GCTCCGGTTACAGGTACTTATGGCTCGTCCGGCGTGCCATCCTCATCATCGACTGCT
GCGCCTTATCATCATGGTGGCTACCCGCACCCTGCGGGATACGGCGCACCAGCAGG
GCCAGGAGGTTCCGGTCCGTCCGGCATCCCATCGGAAGGCTACCCTTCGTACCCTCC
GCAAACGTCAAACACCAGCTCGGACCCCTCAGCGACTTATAATTCAGGCGTTCCTGG
TTATGCTCATCCGAACTCGGCAGGCTATACATATGGCGGGTCACAACCTGGGTTTTA
CCCAAATCCTCATCTCGGGGTGAACGCTGCTGCTGCGCCGGTACCAGGTGCACCTGG
TGCATACGGTGGCTTTAGCCACCCGGCTGGCCCGGCCATCCACTATCTGAATCCGTT
CGAGGTCAAGCACCGTCGTCGCACAACGAAGACTCAATTCCGCGTACTGGAGGGTA
CTTTTAAGGAGAATCCCAAGCCCAACGCTGCTGTCAGGAAGGCGTTGGCGCAACAG
TTGGACATGCCCGGCCGTGCAGTGCAGATTTGGTTCCAGAATCGTCGAGCCAAGGCC
AAGGCCCAGGCTAAGAAGGAAGAGCAAGGAACGAGAGAGCGAGCTGCGACATCGG
AAGGATCGACAAGCACCAGCAGCAAAACGCAGCAACAGCCGCAACAACATTCGCG
ATCTGTTACCGCACCGCAACCAATGGCAGGTGGAGGAAGTAGCACCGATAGCTACT
CAACCAGATCCTCAGCTACGAGCCACGGCAATGGTGGTGGAGGAAGCGGCGGCGGC
GGTAGCGGTGATTACCATCGGCATTCACAGGCGCAGCAGCAAGGGCCGGCTGGCTA
TAGCCAGGCAGGCCACCATGCCTCTTCGCATCATGGACCCAGTGGTTCAGGAGCGTG
GCATGCGAGCCACTCGTCTGCGTCACGGAATTGA
>Twa_g197
ATGAGCAAGCATCCCGAACAAGTCAATTCTCCCCAAGGTACGACCACCTTCCCCCAC
GAGCGGAGGGAATCAAGGGGCAGCAACGGCTCTCTGAAGCGCGTTTACGAAGATTA
CGACAACGATGGGCACCAGCAAAGCACATTCAAGGGGGTGACGCTTCCTAGCTGTT
CTGACCTGGCACTCGGAGCCGAACAGTGGGTCGCCTCAGGCAGTTCGAGACATTCCT
CGCAGGGTCACTCGGCCCTGCTTCCCTCCATCGTCTCTGCCGCGGACCACTCGCCCC
CCTTCTACCAGTCATCGAATCGGCGCCCGAGCTCACACTTCCTCACCCGTTCGCCGT
CCGACCCTGAGAACTCGTACGCCGGCCACGCTCAACACGCCTCTTCAACCATCTCGA
GAATGGTTCGCACCAACACAGTCCCTGAATACGAGGAGGACGACAACGTTCCCAGC

AAGCGCAGACGGTCGTCCAGAGAGTCCATGAGAGACAGCATGGCCGACTCGTTTTC
GACTCGAACGGCCTCGAACTTCCCCGCTGCTGTTGTGAGGGATGAGTCGTCCTCCGC
ACCAGAGCGCTCGGCGTGGAACCACCGAACGTCAGATGGCTTTGACTGGAGGAGCA
GCTCCTCTCGATCATATCTTCGGGTTTCCCCGAACACCAGTCAGATGCAGCGCAGGT
TTTCTCGAGAGGATACGTCGAGATACAGTGAAGGTTTTCGACCGATGCTGCCGTCAA
ACACACTATTTCCCTCGGGCCTTCCGACGAAGCACAGCCCTCCCTCGGGCGCATTCG
CCGACGAGTCCCGACGGATCACGGGCAAGTACAGTGATCTGCGCACGACGAGCAAC
AGTCACAGTAACAACAATAGCTCGGGCGATAGTAGACGAGCATCACCCGACGCACC
CCAACCCTTCAATCGTTACGAGCCGCTCAGTCTGAGTGGATGTCCAGCTCAGAGGTA
TTCTGGCCGCACCTCCCCCGCAGCTGCCCAGGCGTATCGTTCCTCGGAAGCAGCCGA
TCGACAACAGCCTCGCTCCACCGGCGGCCACTCATCAGAGCATCATGCGCACATCTC
GCCTTCGACCTCGCGAGCCGGCAGCGACCGCGACTATGATCAGCGATCTCCTGGCTA
TCATGAGCAGCACCATTTCCCATCCTCATCCGCATCGGGTGCGACCAGTCCGCCATC
GTTGCCATCGCCGCATGGTGACCCGAGCTCCTTCTCACGCTACCGTTCGGGCTCGAT
GGACTCAGGTCGTCAACTGTCCTCACTCTTCGTCTCTCACTTGTCCGGCTTAGCTCAC
GATGAGGGCGGCAGACGGATTCGTCATGTTTACGATCCTCCTGCAGACGCTTCTGTC
AATCACGCATCTCATGAGCAATACCAGCAGCAGCAAGTGCGCCCTTCGACGAACCA
CTCGTCGAATGCGCACCGGTCCTGGTCGTATGTGAATGCCCTCCCGGCGGCTTCGGC
CAGCTATCACAACAATTCAGGGCCGACTTCGAACTCGGCCTTCCGCTCTCAAGAATC
GTCCAATCCCGTCATGTCGTCTGCAAGGATACTTACGCACCTCACTTATCCCGATCA
TGCCGCCCCACATGCCCAGGAGCCGCACGGTATGAGCAACAGCAGCAGCAAAGGCA
GCATTGGCACCAACGATGATAGCCAAGCAAATCAACTCGAACCAGACCTAGACTCC
GACAACGGCACGGTACCTCGGCGGGGCAACTTCCCTCGCCAGGTAACCGAGTACCT
TCGTCAGTGGCTCTTCACGCATCGGGATCATCCTTATCCCTCGGACCAGGAGAAGCG
CGACATGGCACAATATACCAACTTGGACTTAAGGCAGCTTAATAATTGGTTCATCAA
CGCTCGGAGGAGGGTGTTACGGAACGCAGAGGACGATGCGCAACAGCAGGGCGGC
GTCGCGACTGCCGAAGAGTAA
>Tco_g8961
ATGAGCAAGCATCCCGAGCAACCCGAACCCCAAGATGTCAACGTCGCCTACAATCA
GCAACAAAGGAGATATTCAGATACTACCGACGGAGCCCTCAAGCGCGGCTTCGACG
AAGTTCTGGACGATAGGCACCAACGGGCGACGTCTAAAGGAATCAAGCTCCCAAGC
CTTTCCGACCTAGCGCTCGGGGCCGGCCAGTGGGCAGCCTCAAGCCGGTCAAAGCG
ACCCACGTTAGGCTCTGCGGCACTTCTCCCATCCTTTGCCGGCACAAGGAGACAGTC
GTCCTCTCAGCACACAGCGGCCCAATTTCCCACTCGGTCGGCTTCCGATTCCAAGAA
CATGTATGCCGTGAACGCCAATCATGCCTCATCTGCCATCTCCAGGATGGTCCGGAA
TGATCCGGTGGCCGATTACGAGGAGAACGACGCCGGCCCTAGCAAGCGAAGGCGTT
CTCCTACGAGCTATTATGAAGATCACCCGATGGATTCATCCTCAGCTCGGACGGCCC
ACGATACTACTCCTGCTGTGCTGCACTCTGGGACACCCTCGGTCCTAGAACGCTCGG
GATGGAGCTACGGCTCTTCCGAAAGCTACGACTGGAGAGGTAATTCTTCACGACAA
ACTCTGCGAACATCGCCCAGTATCAATCAGATGAAGCGCAAGTTCTCTCAAGGGGA
CATGTCGAAGTCTAGCGAAGGCAATCGGTCAACGCTGGCTAGTGGTCTGCCGTCGAC
AAGTTCATCGTTTTCACGAATCCTCGCAATGGGTAGCCCACCCTTGGGAACGTTCCC
CGACGAGTCTCGACGGATCCCAAGCAGGTACAGCGATGTCCGCACGACGAGCGCCA
TCAACAGCAGCAGTACGAGCGGGAGCAGAAGAGCATCACCCGGCACGTCCCAGCCT

TTCAGCCGTTATGAGCCGCTCAGTCTGGGTGGGTGCCCTGGTCAGAGACACCCCAGT
CAGACCTCTCCTGTTCTCATTCAGTCACACCGGTCTTCGAACGAAGCGAATCAAAGC
TATCAACAGCCGCTTCGCGCCTTCAGTGGCCATTCTTCAGAAGCTCAGCAACAGCTT
TCGCCTTCTACCTCGCGGGACGGCAGCGACCTCGCATATCATCGGCACTCTTCCGAC
TACCGCGAAGAGCGCCAGTACTCCTCCTCATCGAAATCGAGCGCGACCAGTCCAGC
TTCGTTGGCTTCGCCCCGCGTTGATTGTGACTCGTTCTTGGTCAATCGCTCGGATTCG
ATTGATTCAGGGATTGTATCTCGGATGTCTGGGCTGGCGCACGACGAGGCGGAGAG
ACGAGGCTACCATGCTCACGAAGCGGTTACAGACACTCTTGCCTACCATCACCGCCG
CATGCACACTTCAACGGAGCCCTCGTCGACTGCGCTTGGGTCATGGCCGCCCCTAAC
CATGTTACCTCTGGCGTCGACCAGCTACTACCACACCTCGGGAGCAGCTCCGCAGTC
CACTCTCCACTCGCAGACTTCTCCAAACTTTGGCCCGACATCTGCAACTCCAATGCC
GCCGCTTTCTTCCCCCACAGCCCTCGTCCTTTCAACCGGTCATGCCGAAGTTGAGAG
CGATCATAACCAGATGTACAAAAACAGCAGCAGCAGCGGCAGCCACCACAGCACCA
ACACCTCTCCGCAAGCAAACCAGACGGATCCAGACTCGGATGGAGGACCAAGACGG
GGCAACTTCCCTCGTCCCGTTACCGAGTACCTCCGCCAATGGCTCTTCAACCACCGG
GACCATCCTTACCCTTCCGACCAGGAGAAGCGCGAGATGACCCAGATTACCAAATT
GGACCTGAGGCAGCTCAACAATTGGTTCATCAATGCGCGCCGCAGGGTGCTACGCC
ATGGGGACGCGCAACGAGAGCAACTGAGCGGACATCAGGGTGCCGTCGCAGCTGCC
GAAGAGTCATTGGATTGA
>Tca_g10368
ATGAGCAAGCATCCCGAGCAACCCGAACCCCAAGATGTCAACGTCGCCTACAATCA
GCAACAAAGGAGATATTCAGATACTACCGACGGAGCCCTCAAGCGCGGCTTCGACG
AAGTTCTGGACGATAGGCACCAACGGGCGACGTCTAAAGGAATCAAGCTCCCAAGC
CTTTCCGACCTAGCGCTCGGGGCCGGCCAGTGGGCAGCCTCAAGCCGGTCAAAGCG
ACCCACGTTAGGCTCTGCGGCACTTCTCCCATCCTTTGCCGGCACAAGGAGACAGTC
GTCCTCTCAGCACACAGCGGCCCAATTTCCCACTCGGTCGGCTTCCGATTCCAAGAA
CATGTATGCCGTGAACGCCAATCATGCCTCATCTGCCATCTCCAGGATGGTCCGGAA
TGATCCGGTGGCCGATTACGAGGAGAACGACGCCGGCCCTAGCAAGCGAAGGCGTT
CTCCTACGAGCTATTATGAAGATCACCCGATGGATTCATCCTCAGCTCGGACGGCCC
ACGATACTACTCCTGCTGTGCTGCACTCTGGGACACCCTCGGTCCTAGAACGCTCGG
GATGGAGCTACGGCTCTTCCGAAAGCTACGACTGGAGAGGTAATTCTTCACGACAA
ACTCTGCGAACATCGCCCAGTATCAATCAGATGAAGCGCAAGTTCTCTCAAGGGGA
CATGTCGAAGTCTAGCGAAGGCAATCGGTCAACGCTGGCTAGTGGTCTGCCGTCGAC
AAGTTCATCGTTTTCACGAATCCTCGCAATGGGTAGCCCACCCTTGGGAACGTTCCC
CGACGAGTCTCGACGGATCCCAAGCAGGTACAGCGATGTCCGCACGACGAGCGCCA
TCAACAGCAGCAGTACGAGCGGGAGCAGAAGAGCATCACCCGGCACGTCCCAGCCT
TTCAGCCGTTATGAGCCGCTCAGTCTGGGTGGGTGCCCTGGTCAGAGACACCCCAGT
CAGACCTCTCCTGTTCTCATTCAGTCACACCGGTCTTCGAACGAAGCGAATCAAAGC
TATCAACAGCCGCTTCGCGCCTTCAGTGGCCATTCTTCAGAAGCTCAGCAACAGCTT
TCGCCTTCTACCTCGCGGGACGGCAGCGACCTCGCATATCATCGGCACTCTTCCGAC
TACCGCGAAGAGCGCCAGTACTCCTCCTCATCGAAATCGAGCGCGACCAGTCCAGC
TTCGTTGGCTTCGCCCCGCGTTGATTGTGACTCGTTCTTGGTCAATCGCTCGGATTCG
ATTGATTCAGGGATTGTATCTCGGATGTCTGGGCTGGCGCACGACGAGGCGGAGAG
ACGAGGCTACCATGCTCACGAAGCGGTTACAGACACTCTTGCCTACCATCACCGCCG

CATGCACACTTCAACGGAGCCCTCGTCGACTGCGCTTGGGTCATGGCCGCCCCTAAC
CATGTTACCTCTGGCGTCGACCAGCTACTACCACACCTCGGGAGCAGCTCCGCAGTC
CACTCTCCACTCGCAGACTTCTCCAAACTTTGGCCCGACATCTGCAACTCCAATGCC
GCCGCTTTCTTCCCCCACAGCCCTCGTCCTTTCAACCGGTCATGCCGAAGTTGAGAG
CGATCATAACCAGATGTACAAAAACAGCAGCAGCAGCGGCAGCCACCACAGCACCA
ACACCTCTCCGCAAGCAAACCAGACGGATCCAGACTCGGATGGAGGACCAAGACGG
GGCAACTTCCCTCGTCCCGTTACCGAGTACCTCCGCCAATGGCTCTTCAACCACCGG
GACCATCCTTACCCTTCCGACCAGGAGAAGCGCGAGATGACCCAGATTACCAAATT
GGACCTGAGGCAGCTCAACAATTGGTTCATCAATGCGCGCCGCAGGGTGCTACGCC
ATGGGGACGCGCAACGAGAGCAACTGAGCGGACATCAGGGTGCCGTCGCAGCTGCC
GAAGAGTCATTGGATTGA
>Tin_g1181
ATGAGCAAGCATCCCGAACAAGTCAATTCTGCCCAAGGTACGACCACCTTCCCCCAC
GAGCGGAGGGAATCAAGGGGCAGCAACGGCTCTCTGAAGCGCGTTTACGAAGATTA
TGACAACGATGGGCACCAGCAAAGCACATCCAAGGGGGTGACGCTTCCTAGCTGTT
CTGACCTGGCACTCGGAGCCGAACAGTGGGTCGCCTCAGGCAGTTCGAGACCTTCCT
CGCAGGGTCACCCGGCCCTGCTTCCTTCTATCATCTCTGCTGCAGATCACTCGACCCC
CTTCTACCAGTCATCGTATCGGCGCCCGAGCTCACACTTCCTCACCCGTTCGCCGTCC
GAGCCTGAGAACTCGTACGCCGGCCACGCCCAGCACGCCTCTTCAACCATCTCGAG
AATGGTTCGCACCAACACAGTCCCTGAATACGAGGAGGACGACAACGTTCCCAGCA
AGCGCAGACGGTCGTCCAGAGAGTCCATGAGAGACAGCATGGCCGACTCGTTTTCG
ACTCGGACGGCCTTGAACTTCCCCGCTGCTGTTGTGAGGGATGAGTCGTCCTCCGCA
CCAGAGCGCTCGGCGTGGAACCACCGAACGTCAGATGGCTTTGACTGGAGGAGCAG
CTCCTCGCGATCATATCTTCGGGTCTCACCGAACACCAGTCAGATGCAGCGCAGGTT
TTCTCGAGAGGATACGTCGAGATACAGTGAAGGTTTTCGACCGACGCTGCCGTCAAC
GACACTATTTCCCTCGGGCCTTCCGACGAAGCACAGCCCTCCCTCGGGCGCATTCGC
CGACGAGTCTCGACGGATCACGGGCAAGTACAGTGATCTGCGCACGACGAGCAACA
GTAACAATAGCTCGGGCGATAGTAGACGAGCATCACCCGACGCACCCCAACCCTTC
AATCGTTACGAGCCGCTCAGTCTGAGTGGATGTCCAGCTCAGAGGTATTCTGGCCGC
ACCTCCCCCGCAGCTGCCCAGACGTATCGCTCCTCGGAAGCAGTCGATCGACAACA
GCCTCGCTCCACCGGCGGCCACTCGTCAGAGCATCATGCGCACATCTCGCCTTCGAC
CTCGCGAGCCGGCAGCGACCGCGACTATGATCAGCGGTCTCCTGGCTATCATGAGC
AGCACCATTTCCCATCCTCATCCGCATCGGGTGCGACCAGTCCGCCATCGTTGCCAT
CGCCGCATGTTGACCCGAGCTCCTTCTCACGCTACCGTTCGGGCTCGATGGACTCAG
GTCGTCAACTGTCCTCACTCTTCGTCTCTCACTTGTCCGGCTTAGCTCACGATGAGGG
CGGCAGACGGATTCGTCATGTTTACGATCCTCCTGCAGACGCTTCTGTCAATCACGC
ATCTCATGAGCAATACCAGCAGCAGCAAGTGCGCCCTTCGACGAACCACTCGTCGA
ATGCGCACCGGTCCTGGTCGTATGTGAATGCCCCCACGGCGGCTTCGGCCAGCTATC
ACAACAATTCAGGGCCGACTTCGAACTCGGCCTTCCGCTCTCAAGAATCGTCCAATC
CCGTCATGTCGTCTGCAAGGATACTTACGCACCTCACTTATCCCGATCATGCCGCCC
CACATGCCCAGGAGCCGCACGGTATGAGCAACAGCAGCAGCAAAGGCAGCATTGGC
ACCAACGATGATAGCCAAGCAAATCAACTCGAACCAGACCTAGACTCCGACAACGG
CACGGTACCTCGGCGGGGCAACTTCCCTCGCCAGGTAACCGAGTACCTTCGTCAGTG
GCTCTTCACGCATCGGGATCATCCTTACCCCTCGGACCAGGAGAAGCGCGACATGGC

ACAATATACCAACTTGGACTTAAGGCAGCTTAATAATTGGTTCATCAACGCTCGGAG
GAGGGTGTTACGGAACGCAGAGGACGATGCGCAACAGCAGGGCGGCGTCGCGACT
GCCGAAGAGTAA
>Tc517_g1873 _73415_contig
ATGAGCAAGCATCCCGAGCAACCCGAACCCCAAGATGTCAACGTCGCCTACAATCA
GCAACAAAGGAGATATTCAGATACTACCGACGGAGCCCTCAAGCGCGGCTTCGACG
AAGTTCTGGACGATAGGCACCAACGGGCGACGTCTAAAGGAATCAAGCTCCCAAGC
CTTTCCGACCTAGCGCTCGGGGCCGGCCAGTGGGCAGCCTCAAGCCGGTCAAAGCG
ACCCACGTTAGGCTCTGCGGCACTTCTCCCATCCTTTGCCGGCACAAGGAGACAGTC
GTCCTCTCAGCACACAGCGGCCCAATTTCCCACTCGGTCGGCTTCCGATTCCAAGAA
CATGTATGCCGTGAACGCCAATCATGCCTCATCTGCCATCTCCAGGATGGTCCGGAA
TGATCCGGTGGCCGATTACGAGGAGAACGACGCCGGCCCTAGCAAGCGAAGGCGTT
CTCCTACGAGCTATTATGAAGATCACCCGATGGATTCATCCTCAGCTCGGACGGCCC
ACGATACTACTCCTGCTGTGCTGCACTCTGGGACACCCTCGGTCCTAGAACGCTCGG
GATGGAGCTACGGCTCTTCCGAAAGCTACGACTGGAGAGGTAATTCTTCACGACAA
ACTCTGCGAACATCGCCCAGTATCAATCAGATGAAGCGCAAGTTCTCTCAAGGGGA
CATGTCGAAGTCTAGCGAAGGCAATCGGTCAACGCTGGCTAGTGGTCTGCCGTCGAC
AAGTTCATCGTTTTCACGAATCCTCGCAATGGGTAGCCCACCCTTGGGAACGTTCCC
CGACGAGTCTCGACGGATCCCAAGCAGGTACAGCGATGTCCGCACGACGAGCGCCA
TCAACAGCAGCAGTACGAGCGGGAGCAGAAGAGCATCACCCGGCACGTCCCAGCCT
TTCAGCCGTTATGAGCCGCTCAGTCTGGGTGGGTGCCCTGGTCAGAGACACCCCAGT
CAGACCTCTCCTGTTCTCATTCAGTCACACCGGTCTTCGAACGAAGCGAATCAAAGC
TATCAACAGCCGCTTCGCGCCTTCAGTGGCCATTCTTCAGAAGCTCAGCAACAGCTT
TCGCCTTCTACCTCGCGGGACGGCAGCGACCTCGCATATCATCGGCACTCTTCCGAC
TACCGCGAAGAGCGCCAGTACTCCTCCTCATCGAAATCGAGCGCGACCAGTCCAGC
TTCGTTGGCTTCGCCCCGCGTTGATTGTGACTCGTTCTTGGTCAATCGCTCGGATTCG
ATTGATTCAGGGATTGTATCTCGGATGTCTGGGCTGGCGCACGACGAGGCGGAGAG
ACGAGGCTACCATGCTCACGAAGCGGTTACAGACACTCTTGCCTACCATCACCGCCG
CATGCACACTTCAACGGAGCCCTCGTCGACTGCGCTTGGGTCATGGCCGCCCCTAAC
CATGTTACCTCTGGCGTCGACCAGCTACTACCACACCTCGGGAGCAGCTCCGCAGTC
CACTCTCCACTCGCAGACTTCTCCAAACTTTGGCCCGACATCTGCAACTCCAATGCC
GCCGCTTTCTTCCCCCACAGCCCTCGTCCTTTCAACCGGTCATGCCGAAGTTGAGAG
CGATCATAACCAGATGTACAAAAACAGCAGCAGCAGCGGCAGCCACCACAGCACCA
ACACCTCTCCGCAAGCAAACCAGACGGATCCAGACTCGGATGGAGGACCAAGACGG
GGCAACTTCCCTCGTCCCGTTACCGAGTACCTCCGCCAATGGCTCTTCAACCACCGG
GACCATCCTTACCCTTCCGACCAGGAGAAGCGCGAGATGACCCAGATTACCAAATT
GGACCTGAGGCAGCTCAACAATTGGTTCATCAATGCGCGCCGCAGGGTGCTACGCC
ATGGGGACGCGCAACGAGAGCAACTGAGCGGACATCAGGGTGCCGTCGCAGCTGCC
GAAGAGTCATTGGATTGA
>Tc517_bE _g2987_nt
ATGGATACTGTCTCTTTCCTTCAAGAAGCGGATACGCACCTCTGTCGTGCCCTCTCAA
AACCAAATGGACTTGAGGAATTCTCCAGATGGTCAACAGAGGTGCGAGAGGCTGTG
GAAAAGAATGGCATTCAAGGTCTCGACTCTCTGCATATTCGGAAAGTCGTGCAGCGT
GTTCGCGACTCTGCCGCCTTCTTCCTCGGCGTCGAGAAAATCAATCAGCGCGTTGGG

CACTGGATGCTCTCACAAGTCACCTTGACTCTCCGAGAGGAGAGAACCGCGGTGCT
GGCGACAAACGCCGAGATCGACATCACCCCATCCGACTCTCTCAACTCCAAGGTCCT
GCGCCAATGGTTCGCCGACAACATCGTCAACCCCTTTCTCGACACCAAGACAAAGA
ATGCCGTCGTCGCAAAAACGAATGCGTTGCCTGGTGGTGGACGAGGTGGCCGAGCC
CAAAAGCCGATCGATGGCAGGCAGGTCATGCTTTGGTTCATCAACATGCGTCGACG
ATCAGGCTGGACCTCGTTCAGGAAAGTATATGCCCGCAAGGACACAGCCAGGCTCG
ACCGTCTGATTAGCGCCCTCCGACGACAGGGTGAGGCGCCTCCCGCTGGCGTGCCG
GATGGACTCGCATCACTCATGTGCGCATGGAATGTGCCCATGGTCGTCGAACGTGCA
AAGAAATCGAACCAGTCAGTAGCAATCCCGCTCTCGACTGTGAACCCAGCAACGGC
CGATCCAACAGGCGCCATCCGTGCAAAGGCATGCCGAAATGCATACTCCAAGATGA
TGGACTTCATCTCGGCCGGCGCCAAAGAACGAGTCGGCGACTGGCTCGACGAGATC
CTCGAAGGCGTCGAGGAGAAGCCCAAGATCGTCGCACAGACAGGACCAGTGACTTC
TTTAAACGTCAAGTCGGAGCGCAAGATTGCTGTTCTGCCAGGTCGAGCATCCAAGCG
GCAAAACTCCTTGATGGATCACAGTGCGAACCCCTCAAGCGCATCGATTGTCGAGCT
GCCATCATCAGGGTCGATAGTCCTCTCCCCTCGCAAGGCAGCACGTCTGCCCGGCTC
AGAACTCGGTCCCCGCCTGCCATCGACCGTCGCCTCAGAGTCTACGCTGGTCCCGAC
CAACTTCTGGCAAGCCTCCTTGGCATCGAACCCAAGTTTGTCAAGTTTCATCACACC
CTCGGCCAGCACCTCACTTCCAACAGCATCGATAAACTTTACTCTGTCGATGCCCCT
TCATGGCTTCGACAACGGTGTGCCGGCTCAACCTCGCAACGCGTCCAGCGACTCGTC
GTTCAGCGGCATTACCGACATCACCCAATCGAGCCAGAGCACCGACAGCAGTGCGA
CCACGCAGTCGAGCCAGGCGTCGTATGCGTACTTCGCAAGCGGAATGGGGATGGGC
GACTCGTCTCTCGTTCGCTCCAACTGCCCTTCGGAGGCCTCCCCATTGGTGGGCCAA
TCCTCGTCTCCTGCAGCTGCCAGGCACGGAGAGGTCCGTGCCCGTGACGACAACAG
CGTCGAATCGGCACCCAAGCGATTCCGTGCAGGCAATCGACCTCGACACACCGACG
GCGGGGCGCCCATGAGCATGACCAACCTCTCGATGATGCCAGTAGATCAAGCACCT
TTTGCAGCACCGACGATGAACCTCAACCTCAGCCTGAGCATGAACCCTCAGCCCGGC
ATGGGCCCCTTCGGCGCACCTCAGCAACAGTTCCATCCAGGAATGTCCACTCCGGTG
ATGCAGATGCCGAACGTACCCGCAGCGGCAGCGGCAGATCTGGCGGCGTGGTACTA
TCAGACCTTGGCTCAGCAACATCAGCAACATCAGAATGAGATGCCGGTCACCGCGC
TTCACAACTCGACGATGCCTCCCGGCTCGGCTGGAGGGATGCGCTCATTCTCGGACC
CTCAGAGGTTTACGGACGCTCATGGGCCATCGGTCTCCTCGTTTAACGGACTTTACA
TGCGATGA
>Tc517_bW_1.1 g2986
ATGCCAGACGTCGACCTTTTTACAATGACACCGATCGACATCAGCGGGTACGGCAC
AGCGGAATTCTCCGAGGCTTTGCGACAACATCGACTTTCGCAGTCTCTGAGCACCGA
GCTTTTCGACATCTACGTTCGTGCTGTCGAAAACTTTGCAGCCAAAGCGTCGCAGGT
GGCTTTGATGGCGAAAATTCCTGCCGGCCATACTGTCACGACCAGCGATTTCATCCA
GACCCTACAAAACTTCATGCAGCGGAAATGTGATTCCTTGCAAAAGGTGATGCTGGC
GACGGTCCAGGAGGGCAAGAAGCGTCTCGAAGTCGAGGTCTCGGAATCAGCGGCAG
AAGACGAAGACGACGACGACGCTTCGGACGACGACAGCTCCTCTTCCTCCGACGAT
GCTGCCGACGCAGGCGGTGCACGCAAGCACCCTCCTCAAGCGCTTAGGATCCTCGA
ACTTGCATTCGAACGCACTAGCAACATCACGCATGCCGAGAAGAACCGCATCGCCA
GGGCCATCGGCCTTACGCCTCGACAGGTGACCGTCTGGTTCCAAAACCGTCGCAACC
GTCGAAAGCCCGAATCAGCGCGTACCCCAGCCTCTCAGCCTGAGTCCAGCAAGAAG

CGCAAGCTCGCCACTCTTCAAACTGCCGACATCGACGACGAGATGGAAGAGGTCGC
TCGCGTGCCAAGTAGCCTACCAGAGAAGCGCCAGCGATGCAGCCGCCAGACATCAG
CTGCGTCATCAAGCAGCATCAGCTCAGTTGCCTCTTTCGAGTCAAACCTGGGTGACT
GGACGTCCCACGGCGCAAAGTACGGTCGTGCCCCGAAGCCTCGCAACGACAGCGGC
GACAGCAGCTTCACCAGCGTCACTTCGATCTCATCAGGCTCGATGGACTCGTCATTG
GCCGACATCTCTCAACTCTCGCAAGAGAGCAAGGGACCTCCCTTGCGCACAGAGCA
GCCTCCCATGCACGACTTCCTCTTCCACCCTGATGGCTCCATTCAGCCCGACTGGGA
CGACCTCGAGCTCGACCTTGGACAGCTCGAGCAGAGCTTGGCCGAATCTTTCCAAAT
GGGAGCATCTCTTGTCACCAGCCCCTCCCTCGCCGAAGCACCGGTCCAGCCCCTGAC
GATGTCTGCGCTGCGACAAATCGAGGATCGTACCCCGACCAGCGGCGCAAGCTTTG
CAGAGCAACAGTGCTTGGCCGCTGCTCACCCTTCGATCGCCGCCTTCAACTCCATTT
CACAGGCCGAGAACTGGTCTTTCCTTCAGACGGGAGTCTTGGACATGCCCGCCACGC
AAGATCCTCCGACATACTCGAACGACGAATGGCAAAGCGCCCTTCTCAACCTCGATT
GCGACGTTTCGGACATCAGCTCCGCTCTCAGCTTCGTCGACGAGATTCGGAGCAACG
CTTCGCAGTCAGGACACAGCCACCAGCAAGACAGCACCTCGCAGCTCAACGCATTC
CTCGCTCAGTACGGTCAAGACTTCGGATCTTCGCTCTCTGGTCACAGCAGCGGCGCG
GAGTCTGATGACTCCAAACTTCATCTCGATTTCGACTTCGATTTCTCACGGCTCGGCT
TCGCGACCTCGCCTGTCAGCACTCCCAGCATCGCAAGCGAGGACATGAGGCTCAAC
ACGATCCCGGAGCAAGCAGATAAGCAACTGTGCTGA
>Tc517_bW_1.2 Tc517_g2988
ATGGAATCCTTCGAGTCGGCCACTGCCCAGACCTTGCGGGAGATATCTCGCCACTGC
GACTTCTTGATCCAGACATTTTCTGCGACACGCAGGCCCTCGACACCGCCAACTCTC
GACCGTCCATTCGCAGACATCAACCTCAGAGATTATATTTCAGTCAAATGCTCCGAG
TGCTCCGAGGCTCTGCGACGACTTCGACTTTCGCAGTATACGAGCACCGAGCTTTTC
GGCATATACGTACATTCTGTCGAAACCTTCGCAGCCAAAGCTTCGCAAATGGCGTTG
GCAGCGTATGTTCCTGCTGGATCTTCTGGGACGACGAGCGAAATTGTTGTCCAAGAG
GGACTCGAAGTGGCTGTCTCAAACTCACCAGCTGAAGACGACGACGGCGACGATCC
AGACGACGGGACCTCCTCGGACGACGAAGAAGCTGCCGGTGCAGGAGACGCCCAA
AGATACACTCCTCAAACTGTCGCCATTCTTGAACGTGTATTTGAACGCACCAGCAGC
ATCACACATGCCGAAAAGAACCGCATCGCTAATGCCATTGGCCTCACGCCTAGACA
GGTGACAGTCTGGTTTTGCAACCGTCGCTACCGTCACAGTCCATCAGATTCGGCACC
TACGCCAGCCACAAAGTTCGAATCTAGCAAGAAGCGCGAACTGGATCCTCATCAAG
CTGGCGATGACACCGAGGAGGATGTGAAAAGTGACGGCGAGAATGTCGAGGAACT
ACTACTGACTACGATTCAAGGGGGCAAAAAGCGGCTGGAACTTGATGTCTCGGAAT
CAGCAGGAGAAGAGGAAGACGACAGCGACGACGATCAGGACAGTGGAAGTTTCTC
TTCAGAAGACGGTGCTGCTGAAACAGCAGACGCCCGCGAGCGCCCTTCTGACGTAG
TCGCCCTTTTGGATCGCGCATTAGAACGCAACATGAACATCACGCAGGCCGAGATG
AACCACATCGCCAAGCTCATTAACCTCACGCCTCAACAGGTGACGGTCTGGTTCCAG
AAATGTCGCGACCAGCTCAAGCTATCCGATGGAGCACTCACCCCAGCCATGCAGCC
AGAGTCTAGCAAGGGACGTGGGCTCGCTTCTCAGCATTGA
>Tca_bE1 Tca_g1542
ATGGATGACATCGCTTTCTTCCTTCAAGCCGATACACAGGTCCTCGCTATTTCGAGTC
GAGAAGACGAAATCGTCGACTTCAATCGACAGTTGACGCAGGCGCAAGAGGTCGTG
GAAAATCTCAGCATCGATCAAGGACTCGGTGCGGTGCAGATCGAAATAAGCGTCCG

GTACGTCCGCGATTCCATCGCGTTCTTTGTCGGCGTCGAGGGAAGGTATCAGCAGAA
GCTGCAATGCATTCGCACGAAAACGGCCCTGGCTATCCGGGACTGGGTCCGCGCGG
AAGCGACCCTGGCTATCCGGGAAGAAACAGCGGCGGTTACAGCAACATCCATTCAG
GCCAACCCCACGCCGGCGACACGCTCCGAGGCCAACAACACTCCAGAAACAGGCTC
CCAGGCCAGCATCATCACTTCCGGGGTGTTCAACGCCCAGTTCATGCGTCAATGGTT
TGCCAAGAGTATCGTTCATCCCTATTTCCGCATCAAGATGATGCAGGCTGGTGTCGC
CATAGCAAATGCGTTTCCCAGCGGTGAACTCGGTGGCCGAGAACAGAGGCCGATCA
ATTATAAGCGTCTTCAGCTCTGGACCAACATCCTGCGTCATCAATCAGGCTGGACCT
CTTTCAAAAATATTTACGCCCGCAACGACAGAGTCTCGCTCGAGGCTTTGATCAGCG
TCCTTCGACGACAAGGTCAGGCGCCTCCCGCCGGCGAAATGGACGACATCGCAATG
CTCATGTCCGCATGGGATGTGCCCATGATGGTCGAAGGAGAAGGTGCCGACGATAT
CGATGACTATGACGATGACGAGGAGCAGCAGCAGCAAGTTCGAGCGAACATCGAGA
AAAGAAGCAGAGGGGAGGAGACGCTAGTCTCGAGGTCAAGCCAGTCAACAGCAAC
TCCACGGTCGAGCGTGAACCCGGCAGCAGCTGATCCGATAGGTACGATCCGCGCGA
AGGCATGCCGAAACGCGTACTCCAAGATGATGGACTCCATCTCAGCCGGCTCCAAA
ACGCGAGCCAACGACTGGCTCGACGAGATCTCCGAAACGGTTTTCGAGGGCGAAGA
GGAGAAATCCAAGATCGTCGCGCAAAAAGGGCCAGTAGCGCTTCAAACTGCGAATT
CGACTCCGTCTCAGTCGAATGCAAGCTCCTCAAGCGCATCGATTGACGAGCTGCCAT
CATCAGGGTCAATAGTCCTCTCCCCTCGCAAGGCAGCACGTCTGCCCGGCTCAGAAC
TCGGTCCCCGCCTGCCATCGACCGTCGCCTCAGAGTCTACGCTGGTCCCGACCAACT
TCTGGCAAGCCTCCTTGGCATCGAACCCAAGTTTGTCAAGTTTCATCACACCCTCGG
CCAACACCTCACTTCCAGCAGCATCGATGAACTTTACTCTGTCGATGCCCCTGCACG
GCTTCGACAATGGTGTGCCGGCCCAGCTTCGCAACTCGTCCATCGACTCGTCATTCA
GCGGCATCACCGACATCACCCAATCGAGCCAGAGCACCGACAGCAGTGCGACNTGA
>Tca_bE_2 WFGNMR; two motifs FGENESH_mRNA_PRED_Tca_g4472_Tr6610.2
ATGACAATGGATGACATCGCTTTCCTCCGCCAAGCAGATACAGAGATACTCCTTGCT
CTTTCGAATGGAGATGAAGGAATCATCGACTTCAATCGACGGTTTACGCAGGCGCTA
GAGGCTGTGCAAAATCTCAGCATCGATCAACGACCCAACTCGGTGCAGATCAAAAT
GATCATCCGGAACGTCCGCGATACTGTCGCGTATCTCCTCGAATTCGAGGACAAGTA
TCAGCAGAAGGTGCGCTGGGTTCACTCGAAAACGACTCTGACTATCCGAGACTGGG
TGCGCTCGAAAGCTATCCAGGCTATCCGAGAAGAAGGAGCAGCGGTTGCAGCTACA
GACATCCAGGCCAACCTCACCCCGACGACGGACTCTCAGACCAATATCACTCCGGC
GACAGACATCGAGGCCAATCCCACCTCGTTCAAGTTTTTCAACACTCAGATCTTGCG
TGATTGGTTTGTCGGCAATCTCGATAATCCCTTTCCCGACACCAAGACCAAGGATGA
CCTCGTCGCGGAAACAAATGCGTTGCCCGGTGACGGAGCTGGTGGCCGAGCATCAA
AGCCGATGAATACTATTCAGCTTCAGAACTGGTTCGGTAACATGCGTCGTCGATCAG
GCTGGCTCTCGTTCAAGAAAACCTATGCCCGCAACAACAGCGTCGCGTTCCAGGGTT
TGATTGACGCCCTTCGAAGACAGGGTCAGGCGCCTCCCGCAGGCGTCATGGACGAC
GTCGCCGAACTCATGTCCGCATGGAATGTGACTATGATGGTCGAAAAAGACGAGGT
CGATGACGACGACGAAGAAGGGCAGAACCAGCAAGCCCGTACAAAGACAGACAGC
AAAGGAAAGGAGGCATCCGCTTCGAGATCGAACCAGTCAACAGCAAACCCACGCTC
GAGCGTGAACCCAGCAACGGCCGATCCAACAGGTGCGATCCGCGCGAAGGCATGTC
GAGACGCATACTCAAAGATGATGGGCCATATCGCTGGCAAATACACAGAAGAAGGA
GCAGCAGTTGCAGCTACAGAAACCGAGGCCAATCCCACCTCGTCCAAGGTTTTCAA

CACCAAGATCTTGCGTGATTGGTTTGTCGATAATCTCGTTAATCCCTTTCTGACACCA
AGACCAAGGATGACCTCACTGCGGAAACGAATGCGTTGCCCGGTGACGGAGATGGT
GGCCGAGCATCAAAGCCGATGA
>Tca_bW_1
ATGGGCACCGAGCTTTTCGGCATATACGTCCGTTCTGTCGAAACCTTCGCAGCCAAA
GCTTCGCAAATGGCGTTGGCAGCGTATGTTCCTGCTGGATCTTCTGGGACGACTAGC
GAAATTGTTGGTGCGTTGGAAGCATTCCTGCAACGAAAATGCGATTCCTTTCAGGAA
CTACTATTGGCAACAGTCCAAGAGGGACTCGAAGTAGCTGTCTCAAACTCACCAGCT
GAAGACGACGACGGCGACGATCCAGACGACGGGACCTCCTCGGACGACGAAGAAG
CTGCCGGTGTAGGAGACGCCCAAAGACACACTCCTCATACTGTCGCCATTCTTGAAC
GTGTATTTGAACGCACCAGCAGCATCACACATGCCGAAAAGAACCGCATCGCTAAT
GCCATTGGCCTCACGCCCAGGCAGGTGACAGTCTGGTTTTGCAACCGTCGCTATCGT
CACAATCCATCAGATGCGGCACCTACTCCAGCCACAAAGTTCGAATCTAGCAAGAA
GCGCGTACTGGCTCCTCATCAAGCTGGCGATGACACCGAGGAGGATGTGAAAAGTG
ACGGCGAGAATGTCGAGGAACTACTACTGACTACGATTCAAGAGGGCAAAAAGCGG
CTGGAACTTGATGTCTCGGAATCAGCAGGAGTCGAGGACGACGACAACGACGACGA
TCAGGACAGTGGAACTTTCTCTTCAGACGACGATGCTGCTGAAACAGCAGACGCCC
GCGAGCGCCCTTCTGAAGTCGTCGCCCTTTTGGATCGCGCATTAGAACGCAACATGA
ACATCACGCAGGCCGAGATGAACCACATCGCCAAGCTCATTAACCTCACGCCTCAA
CAGGTGACGGTCTGGTTCCAGAAATGTCGCGACCATCTCAAGCTATCCGATGGAGCA
CGCATCCCAGCCACGCAGCCCGAGTCTAGCAAGGGACGTGGGCTCGCTTCTCAGAA
TTGA
>Tca_bW_2 Tca_g8104_Tr7861.1
ATGGGCACCGAGCTTTTCGGCATATACGTCCGTTCTGTCGAAACCTTCGCAGCCAAA
GCTTCGCAAATGGCGTTGGCAGCGTATGTTCCTGCTGGATCTTCTGGGACGACTAGC
GAAATTGTTGGTGCGTTGGAAGCATTCCTGCAACGAAAATGCGATTCCTTTCAGGAA
CTACTATTGGCAACAGTCCAAGAGGGACTCGAAGTAGCTGTCTCAAACTCACCAGCT
GAAGACGACGACGGCGACGATCCAGACGACGGGACCTCCTCGGACGACGAAGAAG
CTGCCGGTGTAGGAGACGCCCAAAGACACACTCCTCATACTGTCGCCATTCTTGAAC
GTGTATTTGAACGCACCAGCAGCATCACACATGCCGAAAAGAACCGCATCGCTAAT
GCCATTGGCCTCACGCCCAGGCAGGTGACAGTCTGGTTTTGCAACCGTCGCTATCGT
CACAATCCATCAGATGCGGCACCTACTCCAGCCACAAAGTTCGAATCTAGCAAGAA
GCGCGTACTGGCTCCTCATCAAGCTGGCGATGACACCGAGGAGGATGTGAAAAGTG
ACGGCGAGAATGTCGAGGAACTACTACTGACTACGATTCAAGAGGGCAAAAAGCGG
CTGGAACTTGATGTCTCGGAATCAGCAGGAGTCGAGGACGACGACAACGACGACGA
TCAGGACAGTGGAACTTTCTCTTCAGACGACGATGCTGCTGAAACAGCAGACGCCC
GCGAGCGCCCTTCTGAAGTCGTCGCCCTTTTGGATCGCGCATTAGAACGCAACATGA
ACATCACGCAGGCCGAGATGAACCACATCGCCAAGCTCATTAACCTCACGCCTCAA
CAGGTGACGGTCTGGGTGAGTCTCATCCACGCTGTCTGCAGAATGCAACCGATGACG
CTTCTTTCCTCTTCGATCTCCAATATAGTTCCAGAAATGTCGCGACCATCTCAAGCTA
TCCGATGGAGCACGCATCCCAGCCACGCAGCCCGAGTCTAG
>Tco_bE_1_ WFINMR; FGENESH:[mRNA]_17 2 exon (s) 70744 - 72694 1875 bp, chain - ; corresponds to Tr1433.1

ATGGAAAAACATGGCATTCAAGGATTCGACTCGGTACAGGTTCAGAGAAACATCCA
GCATGTCCGCGATTCTGCCGCCTTCCTTCTCGGCGTCGAGGACATCAATCAGCAGAC
TGCGCACTGGATACACTCGAAAGTGACCCTGACGTTCAAACGACTGGAACCGGCAA
GTTTGGCGACAAACTCCGAGGTCGACCTCACCCCATCCGACTCCCTCAACTCCAAGG
TCCTGCGCCAATGGTTCGCCGACAACATCGTCAACCCCTTTCTCGACACCAAGACAA
AGAATGCCGTCGTCGCGAAAACAAATGCGTTGCCCGGTGGTGGACGAGGTGGCCGA
GCCCAAAAGCCGATCGATGGCAGGCAGGTCATGCTCTGGTTCATCAACATGCGTCG
ACGATCAGGCTGGACCTCGTTCAGAAAAGTGTATGCCCGCAAGGACACAGCCAGGC
TCGACCGTCTGGTTAGCGCCCTCCGCCGACAGGGTGAGGCGCCTCCCGCTGGCGTCA
CGGACGGCCTCGCATCACTCATGTGCGCATGGAATGTGCCCATGGTCGTCGAGCGTG
CAAAGAAATCGAACCAGTCAGTAGCAATCCCGCTCTCGACTGTGAACCCAGCAACG
GCCGATCCAACAGGCGCCATCCGTGCAAAGGCATGCCGAAACGCATACTCCAAGAT
GATGGACTTCATCTCGGCCGGCGCCAAAGAACGAGTCGGCGACTGGCTCGACGAGA
TCCTCGAAGGCGTCGAGGAGAAGCCCAAGATCGTCGCGCAGACAGGACCAGTGACT
TCTTCAAACGTCAAGTCGGAGCGCAAGATTGCTGTTCTCCCAGGTCGAGCATCCAAG
CGGCAAACCTCTTTGATGGATCACAGTGCGAATTCCTCAAGCGCATCGATTGTCGAG
CTGCCATCATCAGGGTCGGTAGTCCTCTCCCCTCGCAAGGCAGCACGTCTGCCCGGC
TCAGAAATCGGTCCCCGCCTGCCATCGACCGTCGCCTCAGAGTCAACGCTGGTCCCG
ACCAACTTCTGGCAAGCCTCCTTGGCATCGAACCCAAGTTTGTCAAGCTTCATCACA
CCCTCGTCGAACACCTCAATTCCAACAGCATCGATGAACTTTACTCTGTCGATGCCC
CTTCACGGCTTCGACAACGGTGTGCCGGCTCAACCTCGCAACGCGTCCAGCGACTCG
TCGTTCAGCGGCATTACCGACATCACCCAATCGAGCCAGAGCACCGACAGCAGTGC
GACCACGCAGTCGAGCCAGGCGTCGTATGCGTACTTCGCAAGCGGAATGGGGATGG
GCGACTCGTCTCTCGTTCGCTCCAACTGCCCTTCGGAGGCCTCCCCGTTGGTGGGCC
AATCCTCGTCTCCTGCAGCTACCAGGCACGGGGAGGTCCGTGCCCGTGACGATAAC
ACCGTCGAATCGGCACCCAAGCGATTCCGCGCAGGCAGTCGACCTCGACACACCGA
CGGCGGGGCGCCCATGAACATGACCAACCTCTCGATGATGCCAGTCGATCAGGCAC
CTTTCGTAGCACCGACGATGAACCTCAACCTCAGCCTGAGCATGAACCCTCAGCCCG
GCATGGGCCCCTTCGGCGCACCTCAGCAGCAGTTCCATCCGGGGATGTCCACCCCAG
TGATGCAAATGCCGAATGGACCCGCAGCGGCAGCGGCAGATCTGGCGGCGTGGTAC
TATCAGACCTTGGCTCAACAACATCAGCAACATCAGAATGAGATGTCGGGAGCCGT
GCTTCACAGCTCGACGATGGCGCCCGGGTCGGCTGGAGGGATGCGCTCATTCTCGG
ATCCTCAGAGGTTTACGGACACTCACGGGCCATCGGTCTCCTCGTTTAACGGACCTT
ACATGCGGTGA
>Tco_bE_2
Tco_Tc_DAOM_236426_R01_4416_FGENESH:[mRNA]_1_3_exon_(s)_580_2588_1890_bp,_chain_+_Tco_g1542
ATGGATGACATCGCTTTCTTCCTTCAAGCCGATACACAGGTCCTCGCTATTTCGAGTC
GAGAAGACGAAATCGTCGACTTCAATCGACAGTTGACGCAGGCGCAAGAGGTCGTG
GAAAATCTCAGCATCGATCAAGGACTCGGTGCGGTGCAGATCGAAATAAGCGTCCG
GTACGTCCGCGATTCCATCGCGTTCTTTGTCGGCGTCGAGGGAAGGTATCAGCAGAA
GCTGCAATGCATTCGCACGAAAACGGCCCTGGCTATCCGGGACTGGGTCCGCGCGG
AAGCGACCCTGGCTATCCGGGAAGAAACAGCGGCGGTTACAGCAACATCCATTCAG
GCCAACCCCACGCCGGCGACACGCTCCGAGGCCAACAACACTCCAGAAACAGGCTC

CCAGGCCAGCATCATCACTTCCGGGGTGTTCAACGCCCAGTTCATGCGTCAATGGTT
TGCCAAGAGTATCGTTCATCCCTATTTCCGCATCAAGATGATGCAGGCTGGTGTCGC
CATAGCAAATGCGTTTCCCAGCGGTGAACTCGGTGGCCGAGAACAGAGGCCGATCA
ATTATAAGCGTCTCTGGACCAACATCCTGCGTCATCAATCAGGCTGGACCTCTTTCA
AAAATATTTACGCCCGCAACGACAGAGTCTCGCTCGAGGCTTTGATCAGCGTCCTTC
GACGACAAGGTCAGGCGCCTCCCGCCGGCGAAATGGACGACATCGCAATGCTCATG
TCCGCATGGGATGTGCCCATGATGGTCGAAGGAGAAGGTGCCGACGATATCGATGA
CTATGACGATGACGAGGAGCAGCAGCAGCAAGTTCGAGCGAACATCGAGAAAAGA
AGCAGAGGGGAGGAGACGCTAGTCTCGAGGTCAAGCCAGTCAACAGCAACTCCACG
GTCGAGCGTGAACCCGGCAGCAGCTGATCCGATAGGTACGATCCGCGCGAAGGCAT
GCCGAAACGCGTACTCCAAGATGATGGACTCCATCTCAGCCGGCTCCAAAACGCGA
GCCAACGACTGGCTCGACGAGATCTCCGAAACGGTTTTCGAGGGCGAAGAGGAGAA
ATCCAAGATCGTCGCGCAAAAAGGGCCAGTAGCGCTTCAAACTCGAGCATCCGAGC
GACAGGATCCATTGTTCGACCACAGTGCGAATTCGACTCCGTCTCAGTCGAATGCAA
GCTCCTCAAGCGCATCGATTGACGAGCTGCCATCATCAGGGTCAATAGTCCTCTCCC
CTCGCAAGGCAGCACGTCTGCCCGGCTCAGAACTCGGTCCCCGCCTGCCATCGACCG
TCGCCTCAGAGTCTACGCTGGTCCCGACCAACTTCTGGCAAGCCTCCTTGGCATCGA
ACCCAAGTTTGTCAAGTTTCATCACACCCTCGGCCAACACCTCACTTCCAGCAGCAT
CGATGAACTTTACTCTGTCGATGCCCCTGCACGGCTTCGACAATGGTGTGCCGGCCC
AGCTTCGCAACTCGTCCATCGACTCGTCATTCAGCGGCATCACCGACATCACCCAAT
CGAGCCAGAGCACCGACAGCAGTGCGACCACGCAGTCTAGCCAGGCTTCGTATGCT
TACTTCGCAAGCGGAATGGGGATGGGCGACTCGTCTCTCGTTCGCTCCAACTGCCCT
TCGGAGGCCTCCCCATCGGTGGGCCAATCCTGGTCTCCTGCAGCTGCCAGGCACGGA
GAGGTCCGTGCCCGTGACGATAACAGCGTCGAATCTGCACCCAAGCGATTCCGCGC
AGGCAATCGACCTCGACACACCGACGGCGGGGCGCCCATGAACATGACCAACCTCT
CGATGATGCCAGTCGATCAGGCACCTTGTACCGCGCGCACGATGAACCCTCAGCCC
GGCATGGGCGGACCTTCCACGCAGTGA
>Tco_bW_1 Tco_MAT-B_provisional_FGENESH:[mRNA]18_Tco_g7249
ATGGAATCCTTGAAGTCTGCCGACTCCCAGTATCTGCTGGAGATTGCTCAGGACTGC
GACTTCTTGATCCAAAGACTATCTGCTCAACAAGGGCCTCTGGAACCTTCAAACGTC
GACCTTCCAATCGCAAAAATCGACATCATCGGCTATGGCTCAGCCAAATTTTCCGAG
GCTCTGCAGCGACATCGACTGTCACAATCTCTAAGCACCAAGCTTCTCGACATCTAC
GTTCGTGCTGTAGAAAATTTCGCAGCCGAAGCGTCGCGAGTGGCTTCCACGGCGAA
GATTCCTGCAGGACCTTCTGGCACGACCATCGGATTCGCCAATGCGGTGCAAGCATT
CGCGCAACGAAAATGTGATGACTTACAAGAGACCATGCTGGCGACACTCCAGGAGG
GCAAGAAGCGTCTCGAAGTAGAGATCTCGGAATCAGCAGCAGAAGACGAAGACGA
AGACGACGACGATGCTTCAGACGACGACAGCTCCTCTTCCTCCGACGATGCTGCCGA
CGCAGGCGGTGCCCGCAAGCACCCTCCTCAAGCGCTCAGAATCCTCGAACTTGCATT
CGAGCGCACCAGCAACATCACGCATGCCGAAAAGAACCGCATCGCCAGGGCCATTG
GCCTTACGCCTCGACAGGTGACCGTCTGGTTCCAGAACCGTCGCAACCGTCGGAAGC
CCGAGTCAGCGCGTACCCCAGTCTCCCAGCCAGAATCTAGCAAGAAGCGCAAGCTC
GCCACTCTTCAAACTGCCGACATCGACGACGAGATGGAAGAGGTGGCCCGTGTGCC
GACTAGCCTGCCAGAGAAGCGCCAGCGATGCAGCCGCCAGACATCAGCCGCATCAT
CAAGTAGCATTAGCTCAGTTGCCTCCTTCGAGTCAAACCTGGGTGAATGGACCTCCC

ACGGCGCAAAGTACGGTCGTGGCCCGAAGCCTCGCAACGACAGCGGCGACAGCAGC
TTCACCAGCGTCACTTCGATCTCATCAGGCTCGATGGACTCGTCATTGGCCGACATC
TCTCAACTGTCGCAAGAGAGCAAGGGACCTCCCTTGCGCACAGAGCAGCCTCCCAT
GCACGACTTCCTCTTCCACCCTGATGGCTCCATTCAGCCCGACTGGGATGACCTCGA
GCTCGACCTTGGACAGCTTGAGCAGAGCTTGGCCGAATCTTTCCAAATGGGAGCATC
TCTCGTCACCAGCCCCTCCCTTGCCGAAGCACCGGTCCAGCCCCTGACGGTGTCTGC
GTTGCGACAGATCGAGGACCGTACCCCGACCAGCGGCGCAAGCTTTGCAGAGCAAC
AGTGCTTGGCCGCTGCTCACCCTTCGATCGCCGCCTTCAATTCCATCTCCCAGGCTGA
GAACTGGTCTTTCCTTCAGACGGGAGTCCTGGGTATGCCCGCGTCACAAGACCCTCC
GACATATTCGAACGACGAATGGCAAAGTGCCCTTCTCAACCTCGATTGCGACGTTTC
GGACATCAGCTCCGCTCTCAGCTTCGTCGACGAGATTCGGAGCAACGCTTCGCAGTC
AGGACACAGCCATCAGCAAGACAGCGCCTCACAGCTCAACGCATTCCTCTCTCAGT
ACGGTCAAGACTTCGGATCTTCGCTCTCCGGTCACAGCAGCGGCGCGGAGTCTGATG
ACTCCAAACTTCATCTCGATTTCGACTTCGATTTCTCACGGCTTGGCTTCGCGACCTC
GCCTGTCAGCACTCCCAGCATCGCAAGCGAGGACATGAGGCTCAACACGATCCCGG
AGCAAGCAGATAAGCAGCTGTGCTGA

>Tco_bW_2 Tco_g9842 modified

ATGGAATCCTTCGAGTCGGCCACTGCCCAGACCTTGCGGGAGATATCTCGCCACTGC
GACTTCTTGATCAAGACATTTTCTGCGACACACAGGCCCTCGACACCGCCAACTCTC
GACCGTCCATTCGCAGACATCAACCTCAGAGATTATATTTCAGTCAAATGCTCCGAG
TGCTCCGAGGCTCTGCGACGACTTCGACTTTCGCAGTATATGGGCACCGAGCTTTTC
GGCATATACGTCCGTTCTGTCGAAACCTTCGCAGCCAAAGCTTCGCAAATGGCGTTG
GCAGCGTATGTTCCTGCTGGATCTTCTGGGACGACTAGCGAAATTGTTGGTGCGTTG
GAAGCATTCCTGCAACGAAAATGCGATTCCTTTCAGGAACTACTATTGGCAACAGTC
CAAGAGGGACTCGAAGTAGCTGTCTCAAACTCACCAGCTGAAGACGACGACGGCGA
CGATCCAGACGACGGGACCTCCTCGGACGACGAAGAAGCTGCCGGTGTAGGAGACG
CCCAAAGACACACTCCTCATACTGTCGCCATTCTTGAACGTGTATTTGAACGCACCA
GCAGCATCACACATGCCGAAAAGAACCGCATCGCTAATGCCATTGGCCTCACGCCC
AGGCAGGTGACAGTCTGGTTTTGCAACCGTCGCTATCGTCACAATCCATCAGATGCG
GCACCTACTCCAGCCACAAAGTTCGAATCTAGCAAGAAGCGCGTACTGGCTCCTCAT
CAAGCTGGCGATGACACCGAGGAGGATGTGAAAAGTGACGGCGAGAATGTCGAGG
AACTACTACTGACTACGATTCAAGAGGGCAAAAAGCGGCTGGAACTTGATGTCTCG
GAATCAGCAGGAGTCGAGGACGACGACAACGACGACGATCAGGACAGTGGAACTTT
CTCTTCAGACGACGATGCTGCTGAAACAGCAGACGCCCGCGAGCGCCCTTCTGAAG
TCGTCGCCCTTTTGGATCGCGCATTAGAACGCAACATGAACATCACGCAGGCCGAG
ATGAACCACATCGCCAAGCTCATTAACCTCACGCCTCAACAGGTGACGGTCTGGCCA
CGCAGCCCGAGTCTAGCAAGGGACGTGGGCTCGCTTCTCAGAATTGACGATGGGGA
CGAGGAGATGGAAGGTGACAACGAGGACATCGAAGGTGCATGGACGCTCCACGGC
GCAAGGTACGGTCGTGCCCCGTACATCTGGGAATTCGGCATTGCAGCCTGGAAGCTC
GACGCTCCCCTCCTCGGTCTGGTCGTCTTGACATGGACGTAG

>Tin_bE g5612_nt

ATGGACGCGGTTTCGTTTCTTCGCCAGGCGGATACGCACCTCTTTCGAGTTATTTCAG
ACGGGAGAGGTATTCAAGATTATCTCAAACGGATTGAACAGGTGCGAGAGATCGTG
GACAAAAATGGCGTCCCGGGACCCGACGCCGTCCAGGTCTGGGCAAGCCTTCAGTG

TGTCCGTGACTCTGCAGCCTTCTTCCTCGGCATCGAGAGTATCTCTCGACAGATTGCG
GACTCGACAGCTTCAGAAATGTCCTCGGCTCTTCAATCCAAGGGCTCCTCCGCTTCG
GAAACGATCATTTCAGCCGATGTCACCCCTTCCGACTCTCTCAACGCTCAAGTCCTC
CGCCGATGGTTTGCCGACAACATCGTCGACCCCTCTCCCGACACCAAGACCAAGCAT
ACCCTCGTCGCACGGACAAATGCGTTACCCGGCGGAGGCGTCGGCGGTCGCGCCCA
AAAACCGATCGACTACAAGCAGGTCATGCTCTGGTTCATCAACATGCGTCGACGATC
AGGCTGGACCTCTTTCTACAAGACGTTCGCCCGGAAAGACAAAGCCAAGCTAGAAC
GTCTTGTCAACACCCTAAAACGACAAGGCCAAGTCCCACCTCCCGGCGTCAAGGAC
GACATCGCAGCACTCATGGTATCATGGACGGTCCCGATGGTCTCCGAAGAAGAAGA
CGATGAAGACGACGACGAAGAAGAGCAGCAGCAGAAAGGTCGTGGGAACAAAGGG
AGCAAGGCGAAGAAGGCACCAGCTTCAAAGTCTCGACGATCCAACAAGCCCATCAC
GATCCCACTCTCTAGCCAGAACCCGGCATTGGCCGACCCGACGGGAACGCTTCGTGC
AAAGGCATGTCGAAGTGCATACTCGAAGATGATGGACTTCATCTCGGCCGGCGCCA
AAGAGCGAGTCGGCGACTGGCTGGACGAGGTCCTTGAGGGTGTCGAGGAGAAGCCG
AAGATCGTCAAGCAGACGGAAGCAGTATCGTTTTCGAAGGTGTCTTCTTCATCCATG
TCGACAGGGCGCAAGATTGCAGCTCTACCCAGTCGGGCGTCGAAGCGACAAACATC
GTCACACTTTGCCTTCCCTGAGGCTAGCACCTCGAGCGCATCCATCGTTGAGCTCCC
CTCTTCGTCAGTCTCATCATTCGTTCCTCGCAAGTCGGCACGCCTCCCTGGCTCCGAA
CTCGGTCCTCGACTACCATCCACCCTAGCCTCGGACACACCTCTCGTCCCGACCAAC
TTCTGGCAATCATCCCTCTCATCGAACTCAAGTTCCGGAAGCTTCATCCCACCCTCCA
TCGGCAGCTCACTCCCAACAGCATCGATGAGTCTGACACTTTCGATGCCCTTCCACG
GCCTGAACAACGGTGTGCCGGCCCCACCTCGCAACGCTTCGGGACAATCTACGTTCA
GCAGCATCACTGATATCAGCCACTCAAGCCAGAGCACTGATAGTAGTGCAACGACA
CAATCCAGCCAGCCATCCTGCGCCTACTTCGCAAGTGGGGTAGGCATGGGAGACTC
CTCACTCGTTCGCTCCGCCTGCTCGTCGGAAGCTTCTCCCGTAATGGGCAGCGAGTC
GCCAACCCGTCAAGTTTCTTCTCCGGCTGGTATTAAGTCCAGGTCGACTTGCGATGA
CGGTATGCTCGAGTCTACTCCCAAGAGGTTCCGCCCGGGCAACCGACCTCGACACAC
CGATGGAGGAGCACCGACGAGTATCACCAATCTCCCGATGATGACGCTCCACCAGA
CGCCGTTCCCTGTCTCGACGATGAACCTCAATCTGAGCTTGAGTATGAACGGAGAGG
CTGGCATCATTCCGTTTGGGGCACCTCAGCAGACTTACTACCCTGGGATGTCTACGA
CGACGACGAATGTGCCCATGCCGCAGGTGCCTGCGATGGCGGCGGCTGACTTGGCG
GCGTGGTATTACCAGACTGTGGTGCAGCAAAACGGGCATCAGATGCCTGGGACTGC
TTTGTACAGCTCGACCATGCCTGGTACCTCCGCTATGGGGATGCGCTCGACTTCGGG
GCCTGAGAGGTTTATGGAATCTCAATCTCATGGTTCGTCCGGTCCTTCTTTCTACGGA
CCTGGCCATCCATGA
>Tin_bW Tin_g3384
ATGCGATGGCGTTGCGATTCTTTGCAGGACGCTTTGTTGGCAGCAGTCGAAGATGGC
AAGAAGCGTCTCGAAATCGACGTCTCGGATTCAGCAGCAGACGACGACGACGAATC
AGACGAGGATGGTTCCGACTCAGATGACGAGGCTGTCGACCCAAGCGGCGCCCGCA
AGCAGCACCCTCCTCAAGCTCTTGCCATTCTCGAACGAGCATTCGAGCGCACCAACA
ATATCACCCACGCCGAGAAGAACCGCATCGCCAAAGCAATTGGTCTCACGCCACGA
CAGGTCACCGTCTGGTTCCAGAACCGTCGCAACCGTCGAAAGATCTCCAAGCAAGT
AGCCACCCCAGCCTCTGAATCCGGCTCCAGCAAGAAGCGCAAGCTCGCCTCTGTCA
ACGACGAGGACATGGAAGACGTTGCTCGCACGCCATCCACCCTTCCGGAGAAGCGT

CAACGATGCAGCCGCTCAGCATCGGACGCATCTTGCATCTCTTCATCAAGTAGCAGC
TGCTCGGTCGCCTCCTTCGAGTCCGAGTGGACACCTCAGGGCGTCAAGTATGGTCGC
ACGGCCAAGCCTCGCAATGAGAGCGGCGACAGTAGCTTCACCAGCATCACTTCGAT
CTCTTCGAGCTCTATGAACTCATCGGTCGACTCCTCTCAGGAGAGCAAGGTCGAAGA
ACCCGAATCGTACCAATCCCTCAAAACTGCAGAGCCTCTCCGCGCAGAGCAGCCTG
CAGTGCACGACTTCCTCTTCCACCCTGATGGATCTATTCAGCCTGACTGGGACGACC
TCGAGCTCGACCTTGGACAGCTCGAGCAGAGCTTGGCCGAATCTCTCGCGATGGGA
GCTTCCCTTGGCAGTAGCCCTTCTCCTATCGATGGAACTGTTCAGCCTCTGACGGCTT
CCGCTCTTCAACGGATGGAGGATCGCACACCCACAGCCAACACCACATTTGCCGAC
CAGCAACGCCTGGCTGCTCATCCATCGATAGCCGCCTTCACCTCCATCTCTCAAGCT
GAGAACTGGTCCTTCCTCGAGTCAGGAGTCCTTGGCATGCCCGCGGATCAAGGTCAA
TCTTCATACTCAAGCGACGACTGGCAACGCGCACTTGTCAACATCGATTGCGATGCT
TCCGATTTCAGTGCCGCTCTCAGCTTCATGGACGAAGTTCTGAGCAAGTCGGAGTCA
GGACAAGGATATCACGAGAACGACTCCTCTCAGTTCGAGGACTACCTCTCTCAGCTC
GACAAGGACTCAGCTCTCTCCATGTCAAGTCGAAGCAGCACCGCGTGCACGGAGGA
GTCAAAACTTCACCTGGATTTCGACTTTGATTTCTCGAGGCTCGGGTTCTCGACATCG
CCTGTGGGCACTCCGAGCACATCGAGCGAGGCGACGACGTTCAGCACGGTCCCAGA
GGCAGAAGATAAGCTCTCCTGCTGA
>Twa_bE g988_nt
ATGGACGCGGTTTCGTTTCTTCGCCAAGCGGATACGCACCTCTTTCAAGTCCTTTCAG
ACGGGAGAGGTATCGAAGATTTTCGGAGACGAATTGAACAGGTACGAGAGCTGGTG
GAGAAAACGGGCGTCTCCGGACTCGACTCTTTCCAGATCTGGACAAACATTCAATGT
GTCCGTGATTCTGCGGCCTTCTTCCTCGGCATCGAGCGCATCTCTCGACAGATTGCG
GACCGGTCAGCTTCAGAAATGTCCTCGGCTCTTCAAAATCAAGCACTGTCCGCTCCG
GAAACGATCGTCCCAGCCGATGTCACCCCTTCCGACTCTCTTAACGCTCAAGTCCTC
CGCCGATGGTTTGCCGACAACATTGTCGACCCCTCTCCCGACACCAAGACCAAGCAC
ACCCTCGTCGCACGAACAAATGCGTTACCCGGTGGAGGCGTTGGCGGTCGTGCTCA
GAAACCTATCGACTACAAGCAGGTCATGCTCTGGTTCATCAACATGCGTCGACGATC
AGGCTGGACCTCCTTCTACAAAACGTTCGCTCGCAAGGACAAAGCCAAGCTCGAAC
GTCTTGTCAACACCCTTCGACGACAAGGTCAAGCCCCACCTCCCGGCGTTAAGGACG
ACATCGCAGCACTCATGGCAGCATGGACGGTTCCTATGGTCTCTGAAGAAGAAGAC
GATGAAGACGACGACGAAGAAGAGCAGCAGCCAAAAGGTCGTGGAAACAAAGGGA
GCAAGGCGAAGAAGGCACCAGCTTCAAAGTCTCGACGATCCAACAAGCCAATCACG
ATTCCACTCTCCAGCCAGAACCCGGCATTGGCCGACCCAACGGGAACGCTTCGTGCG
AAGGCATGTCGAAGTGCATACTCGAAGATGATGGACTTCATCTCGGCCGGCGCCAA
AGAACGGGTCGGTGACTGGCTCGACGAGGTCCTCGAGGGTGTCGAGGAGAAGCCGA
AGATTGTCAAGCAGACGGAAGCAGTATCGTTTTCGAAGGGGTCTTCTTCATCCATGT
CGACAGAGCGCAAGATCGCAGCTCTGCCCAGTCGGACGTCGAAGCGGCAAGCTTCG
TCACATTTTGCCTTCCCTGAGGCTAGCACCTCGGGCGCATCCATCGTTGAGCTCCCCT
CTTCGTCAGTCTCATCATTTGTTCCTCGCAAGTCGGCACGCCTACCCGGCTCGGAGCT
TGGTCCTCGCCTGCCGTCGACCCTAGCATCGGACACACCGCTCGTCCCGACCAACTT
CTGGCAATCATCCCTCTCATCGAACTCAAGTTCCGGAAGCTTCATCCACTCATCCAT
CGGCAGCTCACTTCCAACAGCATCGATGAGCCTGACACTTTCGATGCCCTTCCACGG
CCTCAACAACGGTGTGCCGGCCCCACCTCGCAACGCTTCGGGACAATCTACGTTCAG

CAGCATCACTGATATCAGTCACTCAGGTCAGAGCACTGATAGCAGTGTAACCACAC
AATCCAGCCAGCCATCCTGCGCATACTTCGCAAGTGGGGTAGGCATGGGAGACTCC
TCTCTCGTTCGCTCCGCCTGCTCGTCGGAAGCTTCTCCGATAGTGGGCAGCGAGTCG
CCAACCCGTCAAGTTTCTTCTCCGGCTGGTATGAAGTCCAGGTCGGCTTGCGATGAC
GGTATGCTCGAATCTGCTCCCAAGAGGTTCCGCCCGGGCAACCGACCTCGACACACC
GATGGAGGAGCACCGACGAGTATCACCAACCTCCCGATGATGACGCTCCACCAGAC
ACCGTTCCCGGTCTCGACGATGAACCTCAACCTCAGCTTGAGTATGAACGGACAGGC
TGGCATCATTCCATTTGGGGCGCCTCAGCAGAATTACTATCCCGGGATGTCGACGAC
GAATGTGCCCATGCCGCAGGTGCCTGCGATGGCGGCGGCGGACTTGGCGGCATGGT
ATTACCAGACTGTGGTTCAGCAAAACGGGCATCAGATGCCTGGGACTGCGTTATACA
ACTCGGCAATGCCCGGTACCTCCGCTGTTGGGATGCGCTCGACTTCGGGGCCCGAGA
GGTTTATGGAACATCATGGTTCGTCCGGTTCTTCTTTCTACGGACCTGGCGTTCCATG
A
>Twa_bW Twa_g989
ATGAACTCCTTCGACTCGGATTTTTCCCAATGGTTGCGGGAAATATCTCGCAACTGC
GACGATATGACCAAATCCTTGCTTTCGAATCATCCGGAGCCTCTCCGACTCAAGGAA
CTCGACTTTCAGTGTGCAGAACTCGAAATCAGCGGCTATGGCGAAGTCAAGATGTCC
CACGCTTTGCAACGTCACCGAATTTCGCAATCCCTGAGTGCCAGGCTCCTAAGCATC
TATCTTCGCGCTGTCGACAAATTCTCTTCCGAAACTACGCGGCTCGTTTTGAAGGCA
AGAATTCCTGCTGGCCCTTCAGTATCACAGAATGACATTGTGGACACGTTGCAGATC
TGCATGCGACGGCAATGTGATACTTTGCAGGACATTGTGTTGGCAGCAGTCGAAGAT
GGCAAGAAGCGCCTCGAAATCGACGTCTCGGATTCAGCAGCGGACGACGACGACGG
ATTGGACGATGACGATTCCGACTCAGATGACGAGGCTGCCGACCTAAGCGGCGCCC
GCAAGCAGCACCCTCCTCAAGCTCTCGCCATCCTGGAACGAGCATTCGAGCGCACC
AACAACATCACCCACGCCGAGAAGAACCGCATCGCCAAAGCGATTGGCCTCACGCC
TCGACAGGTCACCGTCTGGTTCCAGAACCGTCGCAACCGCCGAAAGACCTCCAAAC
CAGTAGCCACACCAGCCTCTCCATCTGGGTCCAGCAAGAAGCGCAAGCTCGCCTCTG
TCAACGACGAGGAGATGGAAGACGTTACTCGCACGCCATCCACCCTTCCGGAGAAG
CGCCAACGGTGCACCCGCTCAACATCGGAAGCATCCTGCATCTCTTCATCCAGTAGC
AGCTGCTCGGTCGCCTCCTTCGAGTCGGAGTGGACGCCTCAGGGCATCAAGTATGGT
CGCACGGCAAAGCCTCGCAACGAGAGCGGCGACAGTAGCTTCACCAGCATTACTTC
GGTATCGTCGAGCTCTATGGAGTCATCGGTCGACTCCTCCCAGGAGAGCAAGGTCGA
AGAACCCGAATCGGACGAGTCTCTCAAGGCTGAAGAGCCTCTCCGCGCAGAGCAAC
CTGCAGTGCACGACTTCCTCTTCCACCCTGATGGATCTATTCAGCCCGACTGGGACG
ACCTCGAGCTCGATCTCGGACAGCTCGAACAGAGCTTGGCTGAATCTCTCGCGATGG
GGGCCTCTCTTGGCAGTAGCCCTTCTCCTATCGATGGAACTGTTCAGCCTCTGACGG
CTTCCGCTCTTCAACGGATGGAGGATCGCACACCCACCGCCAACACCACCTTTGCCG
ACCAGCAACGCCTGGCTACTCATCCATCGATAGCCGTCTTTACCTCTATCTCTGAAG
CTGAGAACTGGTCCTTCCTTGAGTCAGGAGTCCTTGGCATGCCCGCGGAACAAGATC
ATTCATCATACTCAAACGACGACTGGCAGCACGCACTTCTCAACATCGATTGCGACG
CTTCCGACATCAGCTCCGCTCTCAGCTTCATGGACGAAGTTCTGAGCAAGTCGGAGT
CAGGACAAGGACATCACCAGAACGAATCCTCTCAGTTCGAGGATTATCTCTCTCAGC
TCGACAAGGATTCAGCTCTCTCCATGTCAAGTCGGAGCAGCACCGCTTGCACCGAGG
ACTCCACACTTCATCTGGATTTCGACTTTGATTTCTCAAAGCTCGGCTTCTCGACCTC

GCCTGTGGGCACTCCGAGCACATCAAGCGAGGCGGCGACGTTCAGCACGGTCCCAG
AGGCAGAAGATACGCCCTCCTGCTGA
>Tc517_STE3 _73883_contig_g3435
ATGAAAGACAACGCGTACGGACCTCTTGGCATCGTTGCGATCATCCTTGCCGCCTCG
GTGCTTCCCTGGCACATCAAAGCCAAGAATACAGGGGTCCTCATCCTCTCTGCGTGG
CTGATCGTCTCCGTCAGCTGCCTGACCATCAATGCCTTCATATGGTGGCGCAATGTG
GACATCAGATGCCAGATCTGGTGCGACATTAGCACGAAGCTCATTTTCGGTGCCTAC
ATGGGTCTCCCATGCTCGAGTATCTGCATCATTCGACAACTCGAAGAGATTGGATCC
ACTCGCAGGGTTCGGATCACTGCCAAAGATCGCAAACATAAGCTTTACTTCGACCTA
GGCGTTGGCGTTGCTATTCCGTTGCTCTACATGGCGCTACACATCGTCAATCAGGGG
CATCGTTTCGACGTCATCGAAAGCGTAGGCTGCTTCCCTACATACTATATCACGCCA
GTGGCGATCGTGCTCATCCTGGTTCCACCAGTGGTGGCATCCGCCGTCGCCTTGATC
TACAGTTTTCTAGCGCTGAGATGGTTCATCATCCGCCGCCGCCAATTCAACGCTGTG
CTCCAAAACTCCCACAGTGGTCTCGATCGCAGCCGGTACTTGCGACTCATGACGATG
GCAAGCACGGAAGCAATTTGGTCATTTCCGGTCAACATGACGATACTTGTGACCAAG
TTCACCCTCATCCAAGAGCCTATCAACCCATATGTCAGCTGGCAGGACACGCACTTC
GACTTCGGTAGGATCGGTCAGTTCCCCGCTGAGTTCTTCGACGTCAGTCCGGACACC
AAAGCCTACTGGAGAGCAAATATGGCTCTGGGGCGGTACGGGGCACTCGTCGGCTG
CTTCTTCCTCTTCGTCTTTTTCGGCACGAGTCAAGACGCTTTGAAGTTCTATGGTACA
ATCCTGCGCAAAATCACCTCCATTTTCCCTCAATTCCAGGCTCAACGCACCATGTCTC
CAACTACGTCGTCCCAAGCGGATCACTGGAACATCGAGGTCTGTGTTTCGAAGGAA
GCCGGTGTAGAGACTCCGACTCAGTCTGAGATGGACCGGATGGAGAAGTTCGAGAA
ATGA
>Tca_STE3 _g9885
ATGAAAGACAACGCGTACGGACCTCTTGGCATCGTTGCGATCATCCTTGCCGCCTCG
GTGCTTCCCTGGCACATCAAAGCCAAGAATACAGGGGTCCTCATCCTCTCTGCGTGG
CTGATCGTCTCCGTCAGCTGTCTGACCATCAATGCCTTCATATGGTGGCGCAATGTG
GACATCAGATGCCAGATCTGGTGCGACATTAGCACGAAACTCATTTTCGGTGCCTAC
ATGGGTCTCCCATGCTCGAGTATCTGCATCATTCGACAACTCGAAGAGATTGGATCC
ACTCGCAGGGTTCGGATCACTGCCAAAGATCGCAAACATAAGCTTTACTTCGACCTT
GGCGTTGGCGTTGCTATCCCGTTGCTCTACATGGCGCTACACATCGTCAATCAGGGG
CATCGTTTCGACGTCATCGAAAGCGTAGGCTGCTTCCCTACATACTACTTGACGCCA
GTGGCGATCGTCCTGATCTTGGTTCCGCCGGTTTTGGCCTCCTCCGTGGCCCTGATCT
ACAGCTTCTTGGCGCTGAGATGGTTCTTCATCCGCCGCCGCCAATTCCAGGCTGTGC
TTCAGAACTCCCATAGCGGCCTCAATCGAAGTCGGTACTTGCGACTTATGGCCATGG
CAGGGACCGAAGCGTTGTGGTCATTCCCCGTCAATGTGACGATACTCGTGACCAAGT
TCACCCTCCTCAAATTGCCCACTTACCCGTACATCAGCTGGGAGGACACACACTTCA
ATTTCGGGAGGGCCGACCAATTTGATTCCACGTTCTTCGATTCCGATCCTGACCTCA
AAGCCTACTGGCGCTCAAATATAGCTCTTGGGCGATACGGAGTACTTGTCGGCTGCT
TCTTCCTCTTCGTCTTCTTCGGCACGAGTCAAGACGCTTTGGATTTCTACGGAGGAGT
GCTGCGCCAGCTCACCTTTTTTTCCCCACGACGCCAAGCTCAACACACCACATCTCC
AATCACGTCTTCTCAAGGGGACCACTGGAATATCGAGATCTGCGTCTCAAAGGAAG
CCGGTATAGAGACTCCGACTCAGTCTGAGATGGACCGGATGGAGAAGTTCGACCCG

TCGAAATGA
>Tco_STE3
ATGAAAGACAACGCGTACGGACCTCTTGGCATCGTTGCGATCATCCTTGCCGCCTCG
GTGCTTCCCTGGCACATCAAAGCCAAGAATACAGGGGTCCTCATCCTCTCTGCGTGG
CTGATCGTCTCCGTCAGCTGCCTGACCATCAATGCCTTCATATGGTGGCGCAACGTG
GACATCAGATGCCAGATCTGGTGCGACATTAGTACGAAGCTCATTTTCGGTGCCTAC
ATGGGTCTCCCATGCTCCAGTATCTGCATCATTCGACAGCTCGAAGAGATTGGATCC
ACTCGCAGGGTCAGAATCACCGCCAAAGATCGAAAGCATCAGCTTTACTTCGACCTT
GGCGTTGGCGTTGCGATCCCAGTGCTCTACATGGCGCTACACATCGTCAACCAGGGA
CATCGGTTCGACATATTCGAGAGCGTGGGCTGCTACCCTACATTCTATATGACGCCA
GTGGCGCTTGTACTCATTTTGATTCCGCCAGTCGTGGCCTCCGTTGTGGCCCTCGTCT
ACAGTTTTCTAGCGCTGAGATGGTTCGTCATCCGCCGCCGCCAATTCAATGCCGTGC
TTCAAAACTCTCATAGCGGCCTCAACCGCAGCCGGTACTTGCGACTCATGGCCATGT
CAAGCACCGAAGCGCTTTGGTCGTTCCCAGTCAATGTGACGGTACTCGCGAGCAAGT
TCACCCTCCAACATCAGCCTGTCTTCCCATACATCAGCTGGGAGGACACGCACTACA
ATTTCTCGGCGATCGGACAATACCCTGCCGCTTTCTGGGATTCCAGTCCTGATCTCCA
AGGCTACTGGAGAGCATCTGTATCTCTCGGACGGTACGGGGTACTTGTCGGCTGTTT
CTTCCTCTTCGTCTTCTTCGGCACGAGTCAAGACGCCTTGAAGTTCTATGGGGCAGTC
CTGCGCAAGATCACCTCTATTTTCTCGCGATTCCAGGCTCAACGCACCATGTCTCCA
ACTACGTCGTCCCAAGCGGATCACTGGAACATCGAGGTCTGTGTTTCGAAGGAAGCC
GGTGTAGAGACTCCGACTCAGTCTGAGATGGACCGGATGGAGAAGTTCGAGAAATA
A
>Tin_STE3 Tin_g7227
ATGAAAGACCACGCATATGGTCCTCTGGGCATCGTTGCAGTCATCCTTGCCGCCTCG
GTACTCCCCTGGCACATCAAAGCAAAAAATTCAGGAGTCCTGATTCTCTGCTCGTGG
CTCATCATCTCCGTCGCATGTCTCACGATCAATTCCTTCATATGGTGGAACGACACG
GAGATCAGATGCCAGATCTGGTGCGACATCAGTACCAAAATTACCTTCGGTGCCTTC
ATGGGTCTGCCATGCTCGAGTGTCTGCATTCTGCGTCAACTCGAAGAGATTGGATCC
ACTCGCCGGGTTCGAAGCACGGCAAAGGATCGGAAACATCAACTCTACTTTGATCTT
GGCGTCGGCGTGGGTATTCCACTCGTCTACATGATCCTCCACATTGTCAATCAAGGA
CACCGGTTCGACATTTCCGAGAGCATCGGCTGCTTTCCGACATACTACTTGACGCCA
GTTGCCATTGTCCTGGTCCTGGTCCCACCAGTTGTGGCTTCCACTGTGGCTTTGATAT
ACAGCTTTTTGGCCCTGAGATGGTTCGTCATCCGTCGTCGCCAGTTCAATGCCGTGCT
TCAAAACTCTCACACCGGCCTCAACCGTAATCGTTATCTTCGACTTATGGCCATGGC
GGGTACGGAAGCGCTCTGGTCTTTCCCCATCAACGTGACGATTCTCGTGTCCAAATT
CACCCTTCAGGATAAGCCCATCCACCCGTACCTCAGCTGGAGCGATACACATTTCCA
ATTCAGTCGGATCGACCAATATACATCCGCTTTCTGGGATTCTGCTCCAGGACTCAA
GGCGTATTGGAAGGCATCTCTATGCCTAGGACGGTACGGAGCTCTTGTCGGCTGTTT
CTTCCTGTTCGTGTTCTTCGGCACGGGAAACGACGCTTTGAAGTTTTATGGCGCAGTC
TTGCGCAAAATTCTTCCCATTCCCTCGCGAACACAAGCAACACGGAAACCAACAAT
GTCGTCTTCACAGGATCATTGGAACATTGAAGTATGCGTTTCGAAGGAAGCAGGGCT
AGTGACACCCACTCGCTCAGAGATGGACCAATTGGAGAAGGTGGATTCTCTGAGCT
GA
>Twa_STE3 Twa_g7605 (reversed) extraction

TCAGTTCAATGAGTCGACCTTGTCCAGCTGATCCATCTCGGAGTGAGTGGGCGTCGC
TAGCCCTGCCTCCTTCGAGACGCAGACCTCGATATTCCAGTGATCCTCCTGGGATGC
GCCTGGAATGTTTCGTTTTGCTTGGTTTCGCGAAGGCATTGGGCAGATCTTGCGCAG
GAGAGCGCCATACCACTTGAGAGCATCTCGACTCGTGCCAAAGAAGAAGAACAGGA
ACAAGCAGCCGACGAGAGGCCCATACCTCCCAAGATTGAGACTCGCAACCCAATAG
GCCTTGAGGTTAGGACTGGCATCCCAGAATGTAGACGGGAACCTGTCGATCCTACCG
AAGTCGAAGTGGGTGTCGCTCCAGCTGACGTATCGGTAGACTGGCAGTTTCAAGAG
GGTGTACTTCGACACGAGTACGATCACGTTGATTGGGAACGACCAAAGAGCTTCGG
TTCCTGCCATCGCCATGAGTCGAAGATACCGACTTCGGTTGAGGCCGGAGTGAGAGT
TTTGAAGCACAGTATTGAATTGGCGACGACGGATGACAAACCATCTCAACGCCAAG
AAGCTGTATATCAGAGCCACTGCAGAGGCCACAACTGGCGGGACGGATACCAGGAC
GATCGCTAACGGCGTCAGGTAGTATGTCGGAAAGCAGCCGACGCTTTCAAAAATGT
CGAACCGATGTCCTTGATTGACGATGTGCAAGATCATGTAGATCACGGGAATCACCA
CGCCAACGCCGAGATCAAAGTATAGCTGGTGTTTGCGATCTTTGGCGGTGGTTCGGA
CCCGGCGGGTGGATCCGATCTCTTCGAGTTGACGCAGAATGCAGATACTCGAACATG
GGAGACCCATGAAGGCACCGAAGGCGATTTTGGTACTGATGTCGCACCAGATCTGG
CATCTGATCTCCGTGTCGCTCCACCAAATGAAGGCATTGATCGTAAGACATGCGACG
GAGATGATGAGCCACGAGGAGAGAATTAAGACTCCTGAATTTTTGGCTCTGATATGC
CAGGGGAGCACCGAGGCGGCAAGGATGACTGCAACGATCCCAAGAGGGCCGTATG
CGTTGTCTTTCAT
>Tc517_MFA 73883_contig_91 [15711 - 15580] (REVERSE SENSE)
ATGGACAACAGCCTTTTCAACCTCCCTTCCTCCCTTTTCGACCTTCCTCCCTCCGACG
AGCAAAAGGTCAAGGGCGTCGAGCTCGTCGATATGGAGCACAGCAAAGCACCATCC
GGCGGTTGTGTCATCGCT
>Tc_MFA DAOM_236426_R01_818_33
ATGGACAACAGCCTTTTCAACCTTCCTACCTCCCTTTTCGACCTTCCCCCCTCCAACG
AGCAAAAGGTCCAGGTCGTTGCGCTCGTCGATATGGAGCACTCCAAAAACGTTCCA
GCCGGTGGTTGTGTCATTGCC
>Tca_MFA DAOM_238032_R01_3120_11
ATGGACAACAGCCTTTTCAACCTTCCTGCCTCCCTTTTCGACGTTCCTCACTCCGACG
AACAAAAGGTCCAGGGCGGCGAGCTCGTCAACAGGGAACACAAACTGACTCCGACC
GGCGGCTGTGTTATCGCT
>Ti_MFA DAOM_236416_R01_3250_112
ATGGACAGCAGCCTTTTCAACCTACCCCCCTCCGACGAGCAAGTGATTCAGGTCGGC
GAGCTCGTCAACCAGGAGCACAGCAATCCCGTTCCTAACGGTGGCTGTGTTATCGTT
>Tw_MFA DAOM_236422_R01_1251_396
ATGGACAGCAGCCTTTTCAACCTACCCCCCTCGGACGAGCAAGTGATTCAGGTCAGC
GAGCTCGTCAACATGGAACACAGTGCATCAGCGAGTGGCGGCTGTGTTATCGTC