UNIVERSITY OF CALIFORNIA
RIVERSIDE

A Study of Comparative Analysis of Transposable Elements in filamentous fungi

A Thesis submitted in partial satisfaction
of the requirements for the degree of

Master of Science

in

Genetics, Genomics and Bioinformatics

by

Yi Zhou

March 2012

Thesis Committee:

Dr. Jason Stajich, Chairperson
Dr. Katherine Brokovich
Dr. Thomas Girke

The Thesis of Yi Zhou is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Acknowledgement

Firstly, I would like to thank my supervisor Dr. Jason E. Stajich for his kindness and care all through my three years of study and research here. He has taught me how to carry out a scientific research step by step, from raising interesting questions, designing a scheme of a project, searching for the right tools and approaches and learning to use them to find the answer to all the questions. Moreover, I have learnt programming in Perl, bioperl and shell from him. And he is also such a kind and considerable tutor who always gives me care and help when I meet problems in life.

Next, I wish to give my appreciation to Dr. Katherine Brokovich and Dr. Thomas Girke for being my teachers, my guidance committee members, qualifying committee members and also my thesis committee members. They not only taught me useful and important knowledge and skills in class but also gave me precious opinion and support all through my study and research.

I am also grateful to Yaowu Yuan, a postdoc in Dr. Susan R. Wessler's lab. He taught me how to carry out the manual process of identifying and annotating DNA transposable elements by using the conserved DDE/D domains for different DNA superfamilies. And also I want to thank all my lab members, especially Divya Sain, Yizhou Wang, John Abramyan, and Anastasia Gioti. Instead of being just colleagues, they are like my big sisters and brothers who made me feel at home here.

I couldn't be here without the support of my family especially my mother CaiPing Chen and my father Shaohua Zhou and I hope to give them a big hug when I go back home.

Finally I would like to thank Tingting Ma and Nicola Hsu and I'll treasure our friendship here forever.

Thanks,

Yi Zhou

ABSTRACT OF THE THESIS

A Study of Comparative Analysis of Transposable Elements in filamentous fungi

by

Yi Zhou

Master of Science, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, March 2012
Dr. Jason Stajich, Chairperson

Transposable elements (TE) are genetic elements, which can move within the genome. TE are widely distributed in almost all organisms, both prokaryotes and eukrayotes Nearly 45% of human genomes are constituted of transposable elements. There are two major classes of transposable elements: class I and class II. Class I elements, also called retrotransposons, use a so-called “copy and paste” mechanism to replicate them and insert into new positions via an RNA intermediate. Class II elements, also called DNA transposons, do not use an RNA intermediate but a “cut and paste” mechanism to move within genomes. In some filamentous fungi, there is a genome defense mechanism called Repeat-induced Point mutation (RIP) which causes C: G to T: A mutations to transposons regions and thus repress their transposition. My research involved characterization and annotation of transposable elements in seven filamentous fungi and RIP analysis in these species. I have developed an integrated pipeline for TE identification and annotation and

found evidence of RIP using the RIP indices in 5 filamentous ascomycete fungi (*Neurospora crassa, Neurospora tetrasperma, Neurospora discreta, Sporotrichum thermophile*, and *Thielavia terrestris*), but no evidence of RIP in *Chaetomium globosum* and *Sordaria macrosporus*. I found that *Gyspy* and *Copia* LTRs were the most abundant TEs. My results presented two paradoxes: 1) *S.macrospora* and *C.globosum* have low percentage of interspersed repeats but lack evidence for RIP; 2) *S.thermophile* and *T.terrestris* show evidence for RIP but also have many repeats. Moreover, I have discovered several factors related to RIP mechanism, such as the length of transposons, the type of transposons, etc.

Table of Contents

List of Figures

List of Tables

Chapter I

Introduction

Transposable Elements

Transposable elements are sequences of DNA that can move or transpose themselves to new positions within the genome of a single cell. Transposable elements were firstly discovered in maize by Barbara McClintock (McClintock et al. 1948). Now they are known to widely exist in almost all eukaryotic genomes. Transposable elements occupy at least 45% of the human genome and 78% in the maize genome. Transposons are assigned to one of two classes according to their mechanism of transposition, which can be described as either "copy and paste" (class I) or "cut and paste" (class II).

Class I transposons.

Class I elements, also known as retrotransposons, all transpose via an RNA intermediate, which is transcribed from a genomic copy, then reverse-transcribed into DNA by a TE-encoded reverse trancriptase (RT). Retrotransposons are divided into five orders on the basis of their mechanistic features, organization and RT domain phylogeny: LTR retrotransposons, *DIRS-like* elements, *Penelope-like* elements, LINEs (long interspersed nuclear elements) and SINEs (short interspersed nuclear elements) (Wicker et al. 2007). By contrast, Class II transposons do not involve an RNA intermediate. Their transposition is catalyzed by transposase enzymes. The transposase makes a staggered cut at the target site producing sticky ends, excises the transposon DNA and ligates it into the target site. Not all DNA transposons transpose through a cut-and-paste mechanism. Some transposons have been observed to replicate themselves to a new target site.

Both classes of transposable elements have autonomous members, which contain intact ORF encoding for reverse transcriptase or transposase and are capable of self-transposition, and non-autonomous members, which cannot transpose by themselves but can use exogenous enzymes from autonomous members within the same TE family or on occasion across some families.

LINEs occupy almost 17% of human genome (Lander et al. 2001) and SINES constitute 12% of the human genome. And approximately 0.27% of all human disease mutations are attributable to retrotransposable elements (Callinan et al. 2006). The only active autonomous transposable elements in human are said to be LINEs-1 (or L1s) (Kazazian et al. 1998). L1 transposons implemented a transposition mechanism called reverse transcription (TPRT)(Luan et al. 1993). The target site for L1 endonucleases to cleave is a short consensus sequence 5'-TTTT/AA-3' (Feng et al. 1996), therefore the endonucleases can integrate at a great many of sites in the genome. Moreover, L1s can facilitate the formation of pseudogenes that are homologous sequences of protein-coding genes by transposing other sequence, eg. *Alu* retrotransposons and copies of cellular RNAs (Esnault et al. 2000). It is known that for each human gene there are 1~10 (up to 100 in certain cases) pseudogenes (Brosius et al. 1999). The most abundant transposable elements in human is *Alu* elements. And it resulted from a major burst of *Alu* retransposition which occurred about 50~60 million years ago. However, after that burst, the frequency decreased sharply to one new retrotransposition at every 20~125 new births (Shen et al. 1991).

About 8% of human genome is composed of LTR-containing elements(Bannert et

al. 2004). LTR elements usually encode two open reading frames: one named *gag* and the other named *pol* (Figure 1.1). The latter is a polyprotein region including a reverse transcriptase doman that produces a cDNA copy from the RNA intermediate, a DDE integrases domain that functions to insert the new cDNA into the genome and distantly related to *Mariner* DNA DDE transposase (Capy et al. 1997), one aspartic protease domain responsible for cleaving the polyprotein and a RNase H domain that functions to separate the DNA-RNA hybrid. Additional domains can be contained in LTR elements, e.g. chromodomains (Kordis et al. 2005). Based on the ordering of the polypeptide domains in *pol* region, LTR retrotranspons are classified into five superfamilies: *Ty1/Copia, Ty3/Gypsy, Bel/Pao*, retroviruses and ERV. In filamentous fungi, only *Ty1/Copia and Ty3/Gypsy* LTR elements have been detected. Plant scientists have found that by comparing the sequence difference of the long terminal repeats of a single element and the element where it is inserted, it is possible to date the LTR-containing element insertions. For example, Bennetzen and his scientific team reported that the sequence difference between the two LTRs of a certain element is almost always less than the two LTRs of the element which it is inserted in maize. Moreover, they suggested that all of the LTR insertions happened about the last 5 million years after the divergence of sorghum and maize (SanMiguel et al. 1998).

Endogenous retroviruses (ERVs) exist in all vertebrate genomes and are known to represent genomic evidence of ancient germ-cell retroviral infections (Gogvadze et al. 2009). ERVs occupied about 1% of human DNA (Sverdlov et al. 2000). And most of ERVs have lost their ability of transcription and transposition due to extensive mutations

and deletions. The most common form of ERVs in the genome are called solo LTRs, which are derived from homologues recombination between two LTRs of an intact element.

Another group of LTR-containing elements are called tyrosine-recombinase encoding retrotransposons, which encode a tyrosine recombinase instead of integrase (Poulter et al. 2005). DIRS is the first element discovered in this group in slime mold, *Dictyostelium discoideum* and was later found in fungi, plants and animals (Cappello et al. 1985).

Penelope-like elements (PLEs) are a novel type of retroelements different from both LTR-containing and non-LTR retrotransposons (Evgen'ev et al. 2005). They were discovered first in *Drosophila virilis* and later in other eukaryotic genomes. From previous studies, PLEs have their own internal promoter and one ORF encoding for endonuclease and reverse transcriptase distinct from both LTR and non-LTR retroelements.

Class I transposable elements facilitate structural changes to the genome by formation of new retroelements. For example, SVA elements are a composite element including four parts: hexamer repeats (CCCTCT)n, Alu, 15~23 tandemly repeated sequences(VNTR), and SINE-R (SVA = SINE=R +VNTR +Alu) (Wang et al. 2005). And likely formed through integration of several elements into the same genomic locus (Shen et al. 1994), SVA insertions can cause diseases in human. Secondly, events of illegitimate homologous recombination occur in retrotransponsons due to the high copy number and sequence similarity of retroelements (RE). Homologous recombination is an

efficient way to contribute to adaption in evolution since the new combinations represent genetic variation in offspring. Thus recombination between class I transposable elements, or called ectopic recombination, causes deleterious, advantageous or null genomic rearrangement (Gogvadze et al. 2009). REs can also serve as alternative promoters for a host gene transcription, which can either alter the tissue-specificity of its expression or influence the level of transcription of the host gene. Moreover, it was shown that 7~10% of transcription factor binding sites experimentally characterized were derived from repetitive sequences such as simple sequence repeats and transposable elements (Polavarapu et al. 2008). In humans, some REs are the only identified promoter for certain human genes, for example, *Alu*s and antisense *L1* were reported to play the role of the only known promoter for HYAL-4 gene in human (Lagemaat et al. 2003). Additional contributions of REs to the structural or functional diversification to the genome include: as transcriptional enhancers for cellular genes; providers of novel splice sites for host genes; sources of new polyadenylation signals; transcriptional silencers; antisense regulators of the host gene transcription; insulator elements which distinguish blocks of active and transcriptionally silent chromatin; regulators of translation (Gogvadze et al. 2009). Since REs may bring deleterious effects to the structure and function of genome, plants, fungi and animals use different strategies to protect themselves from the proliferation of transposons including RNAi-related mechanism, DNA methylation within REs, histone modifying enzymes, chromatin remodeling enzymes(Gogvadze et al. 2009) and Repeat-induced point mutation (RIP).

The influence of transposable elements.

TEs are estimated to occupy 47% of the yellow-fever mosquito genome, *Aedes aepypti*(Nene et al. 2007), and 39% of the rice genome, *Oryza sativa*. Transposable elements bring mobility to genomes and contribute to the variability of genomes. And the expansion of TEs can lead to overall increase of genome size. When McClintock discovered transposable elements, she also uncovered several ways that TEs can alter the genetic information: by inserting into or around genes and thus generating new alleles; by restructuring the genome through element-mediated chromosomal rearrangement; by imposing their epigenetic marks on flanking DNA(Wessler et al. 2006).

There are also studies which hypothesize that by inserting adjacent to the promoter region of host genes, TEs can up- or down- regulate the expression of the host gene via so-called "transcript infection"(Lankenau et al. 2008). TEs inserted into new positions may disrupt gene expression (Wright et al. 2003).

Transposable elements can cause damage to the genome in the following aspects: 1) a transposons that is inserted into a functional gene will most likely disable the gene; 2) after a transposon leaves a gene, the resulting gap may not be correctly repaired; 3) multiple copies of the same sequence would hinder precise chromosome pairing during meiosis and mitosis; 4) some transposons have their own promoters and these promoters would cause aberrant expression of the linked gene and even diseases, such as hemophilia A and B, Duchenne muscular dystrophy, predisposition to cancer.

The kingdom of Fungi.

Fungi form a unique kingdom of organisms, equivalent to the plant and animal kingdoms. There are estimated to be at least 1.5 million different species of fungi,

including microorganisms such as yeast, mushrooms and molds. Fungi have a combination of features which make them important model organisms for genetics research: 1) They are easy to grow in laboratory conditions and they complete the life cycle in a short time; 2) Most fungi are haploid so they are easy to mutate and to select for mutants; 3) Fungi have a sexual stage and all products of meiosis can be retrieved in the haploid sexual spores, etc.

Fungal species in my study.

The Sordariomycetes are a class of fungi in the subdivision Pezizomycotina (Ascomycota). In my study, I chose 7 Sordariomycetes genomes, including *Neurospora crassa, Neurospora discreta, Neurospora tetrasperma, Chaetomium globosum, Sordaria macrosporus, Sporotrichum thermophile* and *Thielavia terrestris*. Among them, *Chaetomium globosum* is a pathogenic fungus, while *Sporotrichum thermophile* and *Thielavia terrestris* are two thermophilic fungi which can grow at or above 50°C.

Among these species, *N. crassa* is an important model organism used world-wide in research for studying epigenetics and gene silencing, as well as many aspects of biochemistry and cell biology. *C. globosum* can be agents of skin and nail infections in humans by contaminating decaying plant material, seed and other cellulosic substrates. It is also known to produce mycotoxins. Cases of deep fatal infection have been reported in immunocompromised patients. So as a contaminant in the indoor environment, *C. globosum* is important to human health.

Thermophilic fungi have a minimum growth temperature at or above 20°C and a maximum growth temperature extending up to 60°C to 62°C (Maheshwari et al. 2000).

Since high temperatures help solubilize some components of lingo-cellulosic feedstock and decrease the viscosity of slurries of biomass, the thermophilic fungi can be excellent sources of thermophilic enzymes which promote the development of advanced technologies for the biomass derived fuels and chemicals sector and many other industries. Furthermore, *T. terrestris* plays an important role in the global carbon cycle by returning carbon in biomass polysaccharides to the atmosphere. By comparing the thermophilic fungus with the model organism *N. crassa* and the pathogenic fungus *C. globosum*, I can learn more about the evolutionary history of these species and the features of thermophilic and pathogenic fungi.

Transposable elements in filamentous fungi.

While transposons have long been known in bacteria, plants, and animals, it was not until 1989 that the first molecular analysis of a transposon from a filamentous fungus was reported- the *Tad* element from *N. crassa*. Since then, the scientific community has witnessed an enormous increase in the number of cloned and sequenced fungal transposable elements (Kempken et al. 1998). To date, all kinds of transposable elements have been characterized and identified in fungi: Class I elements (Retrotransposons, LINEs or SINEs); Class II elements (importantly in biotechnology *Tc1/Mariner*-family, *Fot1/Pogo* family, *hAT* family).

Applications of transposable elements in fungi.

Most importantly, transposable elements can be developed as useful tools in biotechnology in multiple perspectives, for example, gene tagging, strain identification, diagnostic and population analysis of fungal strains (Kempken et al. 1999). 1) *T.inflatum*

strain ATCC34921 has been used in industry to produce cyclosporine and contains an active *hAT* transposons called *Restless* (Kempken and Kuck et al. 1996), which has relation with the well-known maize TE *Activator* (Kunze et al. 1996) and thus has the potential to be used as a molecular tool. 2) It was believed that methods like transposon mutagenesis might bring benefits to sexually deficient fungal species by genetic manipulation. 3) In addition to the most well-know experiment methods like restriction fragment length polymorphism (RFLP) and randomly applied polymorphic DNA, transposable elements may be a new source of tools for pharmaceutical industry to identify specific pathogens in plant pathology and characterize industrially useful production strains (Kempken et al. 1999). 4) There is a kind PCR named rep-PCR (repetitive element based PCR)(George et al. 1998), which used transposons-specific primers and requires abundant transposons, e.g., in the case of Pot2 about 100 copies are present in the genome (Kempken et al. 1999). 5) Taken distinctive properties of different transposons, developing transposons-aided gene tagging (Figure 1.2) is an important goal. The major problems lie in two aspects: sufficient expression of transposase and functionally efficient selection system (Kempken et al. 1999).

Repeat-induced Point Mutation (RIP) mechanism.

Different organisms have developed their own genome defense mechanisms against mobile elements such as transponsons. Bacteria have a high rate of gene deletion, which can restrict the activity of transposable elements. In animals and plants, there are RNA interference mechanisms (RNAi), like siRNAs, microRNAs and piRNAs. In the kingdom of fungi, three main genome defense mechanisms have been reported: Repeat-

induced point mutation (RIP), quelling and methylation (Muszewska et al. 2011). Recently another new mechanism named sex-induced silencing has been described in fungi (Wang et al. 2010).

Repeat-induced point mutation (RIP) was first discovered as a genome defense mechanism against transposable elements in *Neurospora crassa* (Selker et al., 1987) which causes C:G to A:T mutations to duplicated regions of DNA during sexual cycle. The process in *N. crassa* is illustrated by Figure 1.3. RIP is triggered by both tandem and ectopic duplication (Selker et al. 1990). In previous study, RIP was observed to identify tandem duplications longer than ~400bp or ~1kb for unlinked duplications with sequence similarity greater than ~80% (Galagan et al. 2004). If there are two duplicated sequences, either none or both of them would be RIP'ed and it will never happen that only a single copy is affected (Selker and Garrett 1988). Even in multiple sequences, duplicates are RIP'ed in pairs (Fincham et al. 1989). In *Neurospora*, mutations that are induced by RIP are reported to occur preferentially in CpA dinucleotides, and less frequently in other dinucleotide contexts. In other fungi, the dinucleotide preference may differ, e.g. A/TpCpA/T triplets are the most common RIP substrates in *M. oryzae* (Ikeda et al. 2002). RIP-mutated sequences are frequent targets for DNA methylation in vegetative cells, but the relation between RIP and DNA methylation is still unknown. Because of RIP, repetitive elements in *N. crassa* are sensitive to point mutation and inactivation during meiosis. On the other hand, RIP also accelerates the rate of evolution by generating new transposons. For example, three novel retrotransposons and their degenerate relatives of

the *Tad* LINE-like element probably caused by RIP have been characterized in chromosome VII of *N. crassa* (Cambareri et al. 1998).

RIP remains a mysterious process because it occurs in specialized microscopic, dikaryotic ascogenous tissue, making it difficult to observe. The only molecular component that has been discovered to be involved in RIP is the gene named *rid* (RIP Defective). The *rid* gene was identified with the other DNA methyltransferase (DMT) homologue, *defective in methylation* 2 (DIM-2) in the genome sequence and these two predicted proteins contained a number of conserved motifs identified in all known DMTs(Goll et al. 2005). However, only RID is essential for RIP (Freitag et al. 2002) while DMT-2 is essential for many kinds of DNA methylation in vegetative tissue but not involved in RIP (Kouzminova and Selker 2001). It is still unknown if RID has deaminase and/ or DNA methyltranferase activity. Scientists found that methylated sequences slightly mutated by RIP do not serve as *de novo* methylation signals, which suggests that DNA methylation was perhaps established by RID during sexual cycle and maintained by DIM-2 during vegetative growth (Singer et al. 1995). Moreover, it was found that the severity of RIP showed an inverse correlative relation to the cellular level of AdoMet, the methyl donor (Rosa and Mautino 2004). And by inducing a mutation that caused slowed development during sexual cycle when RIP happens in *P.anserina* (closely related to *Neurospora*), scientists found that extended time in the microscopic specialized dikaryotic ascogenous tissue causes the efficiency of RIP mechanism to increase (Bouhouche et al. 2004).

A process similar to RIP called methylation induced premeiotically (MIP) has been identified in *Ascobolus immerses* and *Coprinus cinereus*. It is not caused by point mutations, which is a characteristic of RIP, but instead depends on DNA methylation to silence duplicated sequences (Rossignol and Faugeron et al. 1994, Freedman and Pukkila et al.1993).

The efficiency of RIP varies from species to species. There is considerable variation for RIP efficiency even among wild-collected *N. crassa* strains (Bhat and Kasbekar et al. 2001). It was found that RIP was the most severe in *N. crassa* than other studied fungi. For example, in the two fungal species with RIP defense mechanism *M. grisea* (Kachroo et al. 1994, Kito et al. 2003) and *F. oxysporum* (Chalvet et al. 2003, Daviere and Daboussi et al. 2001), complete or nearly identical transposable elements and "active" transposons have been found.

Despite RIP being such a powerful "defender" in some filamentous fungi like *N. crassa,* the following genomic elements can evade RIP due to their intrinsic properties: 1) 5S rRNA genes which have 75 copies in the genome but are not arranged in tandem; 2) H2A, H2B and H3 histone genes which contain introns and introns may interrupt regions of high homology; 3) rDNA in tandem repeats producing large ribosomal rRNAs and it was believed that the nuclear organization helped them escape from RIP.

Approaches to transposon identification and annotation.

In a recent review (Lerat et al. 2010), approaches to identify TEs are classified into mainly three groups: structure-based, homology-based and phylogenetic-based techniques. Each technique has its own strengths and limitations. RepeatMasker

representing library-based technique can give a broad picture of distribution of different types of TEs at large-scale auto genome analysis but it is dependent on the quality of the well-annotated TE library. As we know, each genome may have some transposons with a low similarity to the sequences of library and these species-specific elements would not be missed by these homology-based methods. LTR-harvest is a structure-based program for identification of LTR-containing elements and can identify LTR retroelements according to the structure features but may not perform very well in special cases like solo-LTR elements, nested or truncated LTR-containing elements, etc. The theory of phylogenetic-based method is that each superfamily has its own “signature string” of the DDE/D catalytic domain but as we know, only 11 of the identified 19 superfamilies of DNA transposons were found to contain this domain.

Questions of interest.

The number and kinds of transposable elements can vary greatly from species to species. For example, retrotransposons occupy more than 80% in wheat, 26% in rice, 54% in sorghum, 21.4% in the *Brachypodium* genome, but only 8% in the human genome. So in this thesis, I asked the question "How do different classes of transposable elements distribute in the seven filamentous fungi and which kinds of transposable elements are the dominant type?" After I determined the distribution of TEs in each genome, it turned out that LTR-containing elements were the dominant type in these seven genomes. And the research team at the University of Texas at Arlington carried out an analysis and found that isolated transposons in the parasitic triatomine bug were 98% identical to transposons found in opossums and squirrel monkeys that are the hosts of the

bug and they made a hypothesis that there might be a "horizontal DNA transfer" between some parasites and vertebrate hosts. By constructing a phylogenetic tree on LTR elements in primates (gorilla, gibbon and chimpanzee), scientists deduced that two LTR elements were proliferated during the last 2 to 5 million years from the integration of the original LTR elements (Huh et al. 2003). Therefore, I was interested in looking at the evolution history for LTR-containing transposable elements in these seven closely related fungi species by doing phylogenetic analysis.

Up to now, there were not very well-known software or programs especially for DNA transposon annotation. I used a novel method based on the conserved DDE/D domain in the transposase of different DNA transposon superfamilies to identify Class II transposable elements. Previous studies revealed that in *N .crassa* only a LINE-like transposable element called Tad was found to be still active in this genome (Cambareri et al. 1994). Thus I'm going to design methods to help identify "active" or "potentially active" transposable elements in these seven genomes.

Repeat-induced Point mutation (RIP) mechanism is a very important defense mechanism against transposable elements and has influence on the other kinds of DNA elements as well. Therefore, I wanted to investigate if there are RIP patterns in these seven genomes by using an approach called RIP indices. In addition, previous studies show that RIP mechanism require a minimum length of transposable elements to be at least ~400bp. My hypothesis was that RIP mechanism might have a preference on affecting longer transposable elements and in my research I wrote a Perl script to test this hypothesis.

Figure 1.1 A schematic representation of LTR retrotransposons in fungal genomes. LTR = Long Terminal Repeat, AP = aspartic protease, RT = reverse transcriptase, RH = RNase H, INT = DDE integrase.

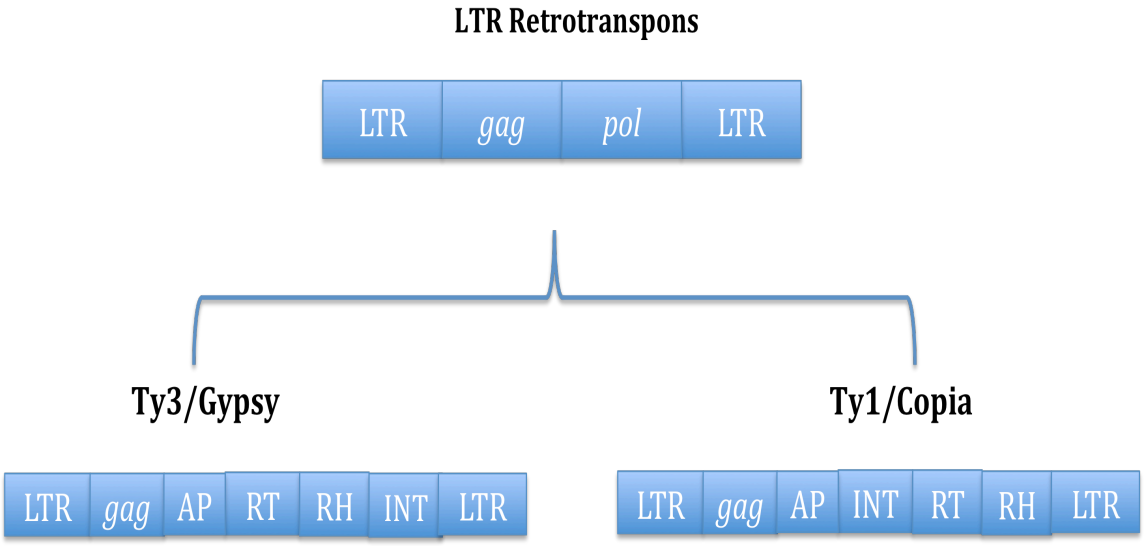


Figure 1.2 Gene tagging with transposable elements.
This method is based on the fact that transposons are able to transpose themselves into the ORFs of other genes. After insertion, the target gene is disrupted and thus creates a mutant phenotype. With the help of the known sequence for the transposons, PCR methods can identify the flanking gene of interest.

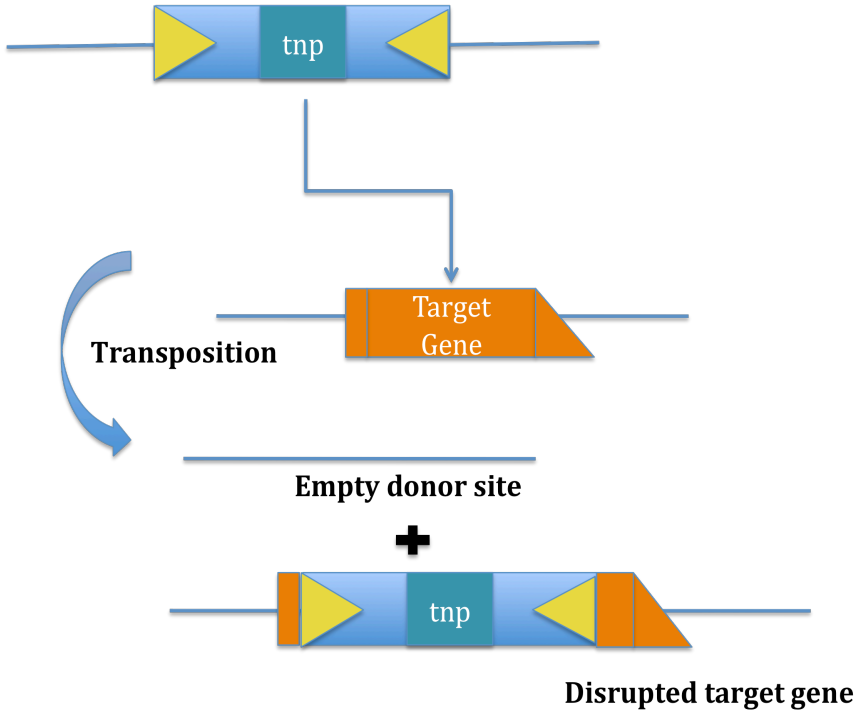
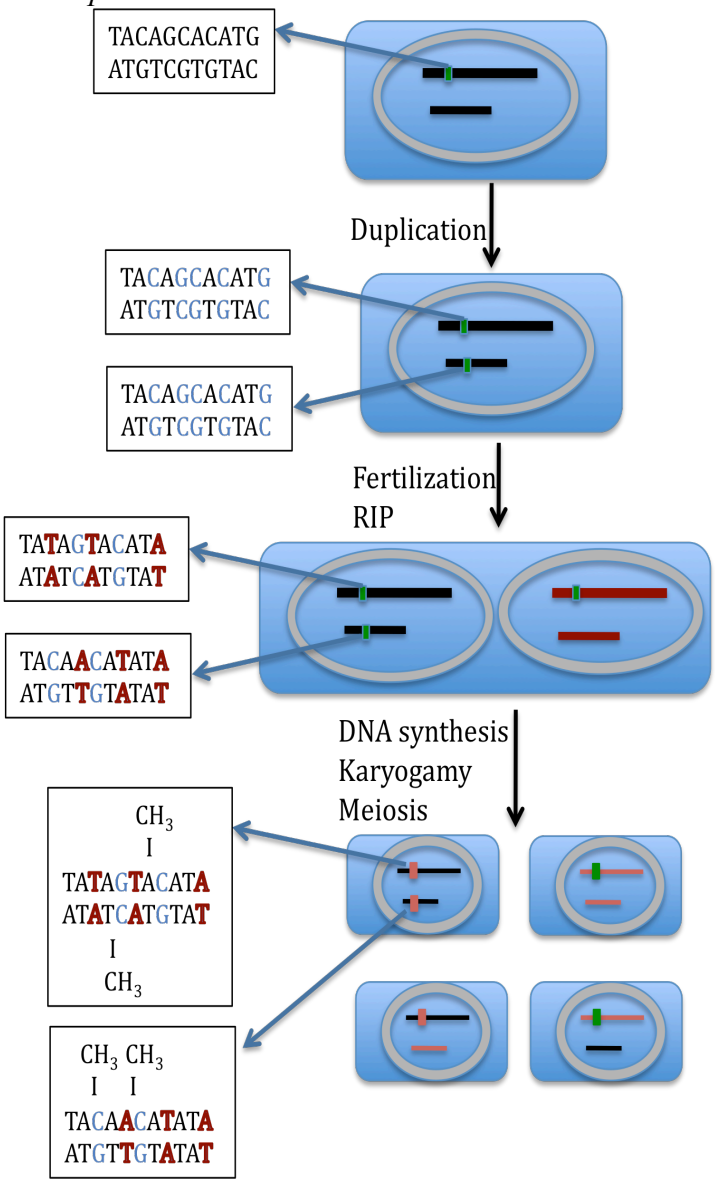


Figure 1.3 RIP Process in *N.crassa*

This graph is based on Galagan and Selker's paper on RIP in 2004**.** RIP happens after fertilization but before DNA synthesis and karyogamy meiosis. (RIP-Mutated C:G pairs are shown in red letters; unmutated C:G pairs are shown in blue). RIP-mutated sequences are frequent targets for DNA methylation and this may cause transcriptional silencing in *Neurospora.*



Chapter II

An overall characterization of transposable elements in seven fungal genomes using RepeatMasker and RepeatModeller

Introduction

Filamentous fungi are characterized by long, often multinucleate hyphae that grow by tip extension (Roper et al. 2011). And scientists are becoming increasingly interested in them. It is said that there are more than 15,000 publications in just the last five years. Well-studied fields include basic forms of fungal growth, development and proliferation, genetic information, cell biology, gene and gene network, epigenetics and so on. Since *N. crassa* is easy to grow under laboratory conditions and has a haploid life stage which facilitates genetic analysis, it becomes an excellent model system in many perspectives, like cell polarity, cell fusion, circadian rhythms, etc. There are more than 30 laboratories on the world carrying out research work on this model organism. Compared to other aspects, there are only a limited number of studies on repetitive elements such as transposable elements in *N.crassa* and other recent sequenced fungal organisms. Therefore, my goal was to carry out an analysis of *de novo* transposon identifiction aided with homology-based TE identification in *N.crassa* and the other six genomes to get a broad picture of the distribution of major superfamilies of transposable elements.

Materials and Methods

The kinds and percentage of transposable elements from the same superfamily can vary greatly from species to species, or even from individual to individual. Just by a homology search against the most commonly used TE database-Repbase, it is not

sufficient to identify most transposable elements, especially species-specific elements. Therefore, I carried out an analysis integrating homology-based approaches represented by RepeatMasker (http://www.repeatmasker.org) and *de novo* based approaches represented by RepeatModeler (http://www.repeatmasker.org).

RepeatMasker is a program which searches genome sequences for seven kinds of genomic interspersed repeats: simple repeat, tandem repeats mostly found in telomeres or centromeres of chromosomes 100~200 base pairs, segmental repeats and other interspersed repeats like DNA transposons, retrovirus retrotransposons, non-retrovirus retrotransposons, and processed pseudogenes, SINES, and RNA genes. It is said that up to 50% of human genomic DNA sequence had be masked by RepeatMasker. This program does not only generate a detailed annotation of the genomic interspersed repeats present in the query sequence but also a modified version of the query sequence where all the already annotated interspersed repeats have been masked. RepeatMasker screens a database called Repbase Update (RU) for repetitive elements, which is a comprehensive database of repetitive DNAconsensus sequences from different eukaryotic species provided by the Genetic Information Research Institute (giri).

The genomes of the seven filamentous genomes were downloaded from the Joint Genome Institute (JGI) and Broad Institute and the strains and versions of the genomes I used are illustrated in Table 2.1. The *N. crassa* genome is 43MB, organized in 7 chromosomes and encodes about 10,000 protein-coding genes. The genome was sequenced by deep whole-genome shotgun (WGS) and paired-end sequencing from various clone types. The data provided an average of > 20-fold sequencing coverage and

>98-fold physical coverage of the genome and an N50 length of 1.56Mb (Galagan et al. 2003). *N. discreta* FGSC 8579 *mat A* strain has been sequenced by Sanger sequencing and assembled by Arachne assemble, using paired end sequencing reads at a coverage of ~8.59X. The current draft release, version 1.0, includes 176 main genome scaffolds totaling 37.3Mb and contains a total of 9948 gene models annotated and predicted by JGI annotation pipeline. *N. tetrasperma* FGSC 2508 *mat A* strain has been sequenced by 454 and Sanger sequencing and assembled by the Newbler assembler aided by JGI gapResolution software. The newest draft release, version 2, assembled the genome into 81 genome scaffolds totaling 39.1Mb and a total of 10,380 genes were structurally and functionally annotated (Ellison et al. 2011). *C. globosum* was sequenced by Sanger sequencing and assembled into 37 main scaffolds totaling 34.34Mb. A total of 11,124 gene models have been annotated in *C.globosum. Sordaria macrospora* was sequenced by a combination of Illumina/Solexa and Roche/454 sequencing and assembled into a 40Mb draft version with the Velvet assembler (Nowrousian et al. 2010). *Sporotrichum thermophile* has been sequenced by Sanger sequencing and assembled by Arachne assembler. The whole genome contains 7 main scaffolds totaling 38.74 Mb. *Thielavia terrestris* plays an important role in the global carbon cycle. *Thielavia terrestris* NRRL 8126 finished genome sequence assembly v2.0 was assembled by Archne assembler and into 6 main scaffolds totaling 36.91 Mb (Berka et al. 2011).

Seven fungal genomes were firstly input into the software of RepeatModeler to obtain repetitive elements and construct a repeat library using a module called BuildXDFDatabase. There were several rounds of computation in RepeatModeler and in

order to run it more efficiently, I havewritten a bash script named “run_RM_all.sh” which realize parallel computation in seven genomes at the same time. For constructing a better TE library with more families of elements, these seven repeat libraries were compiled together into a big library called “Sordaria superlibrary”. The program of RepeatMasker scanned each fungal genome for interspersed repeats, which had homologues either in the well-curated TE library-Repbase Update or in the “Sordaria superlibrary”. As RM also includes simple repeats like TATATA… in each genome during the identification process and these kinds of simple direct repeats are not transposons, I asked RM to filter them by changing parameters. In RM results, it gave the sequences of the masked transposable elements in FASTA format and generated a table showing the total amount in basepairs for interspersed repeats and the length and genome content for major classes of repeats, eg. SINEs (ALUs, MIRs), LINEs (LINE1, LINE2, L3/CR1,etc), LTR elements (ERVL, ERVL-MaLRs, ERV_ClassI, ERV_Class II,etc), DNA elements (hAT-Charlie, TcMar-Tigger, etc) and unclassified repeats. I wrote two Perl script called “count_length.pl” and “count_number.pl” (http://code.google.com/p/stajichlab/ source/browse/#svn/trunk/transposon) to calculate the total length of transposable elements for five major kinds: SINEs, LINEs, LTR elements, DNA elements and unclassified elements and then divided by the genome length for each fungal genome to obtain the proportion of TEs in the genome. For better illustration, I used R commands to generate a bargraph based on these results with each major kind of TEs represented by a color. In addition, repetitive elements identified in

the seven fungal genomes were joint together and UCLUST (http: //drive5.com/usearch/ usearch3.0.html) was used to classify these interspersed repeats into different families.

Results

For the first step, RepeatModeler was used to build up a species-specific TE library for each fungi. Secondly, the seven TE libraries were integrated into a big TE library named "Sordaria superlibrary". With aid of this super-library, each of the seven studied genome was screened for repeats by RepeatMasker program. The genome content in percentage each TE superfamily contributes to is shown in Table 2.2. Here I observed two interesting facts: 1) My RIP index analysis showed *C .glosobum* and *S. macrosporus* had a RIP pattern but Table 2.1 showed the percentage for transposable elements in these two genomes were 6.44% and 1.87% respectively. This percentage was not as high as what we expected since they don't have a RIP defense mechanism to protect themselves from expansion and proliferation of TEs; 2) My RIP index analysis showed the two thermophilic fungi *S. thermophile* and *T. terrestris* may not have RIP mechanism. However, the genome content of transposable elements is up to 19.8% and 18.74% respectively and this amount is considerably higher than it in other fungal genomes with RIP defense mechanism such as 7.74% in *N. crassa,* 6.72% in *N. discreta,* and 5% in *N.tetrasperma.*

For better illustration, I converted the table into a barplot (Figure 2.1) using R language. From these results, it is obvious that class I elements outnumber class II elements in all of the seven genomes and LTR retrotransposons occupy the greatest fraction of genome content among all TE superfamilies, especially in the two

thermophilic fungi-*S. thermophilie* and *T. terrestris*. Moreover, each of the fungus genome contains a considerate amount of repeats, which cannot be classified into any of the known superfamily and labeled as "unclassified" in RepeatMasker.

Conclusion

By using a homology-based method integrated with a *de novo* TE-finding method, I successfully characterized the transposable elements in the seven genomes and counted the total length of them and obtained the genome proportions they occupied. From the RM results, I observed that Class I elements were much more than Class II elements in all these seven fungal genomes. And LTR transposons were the dominant type in the two thermophilic fungi (*S. thermophilie* and *T.terrestris*). Moreover, I observed two paradoxes: 1) *C.globosum* and *S.macrosporus* might not have RIP defense mechanism and the percentage of TEs in the genomes are not as high as expected; 2) Computational methods showed that *S.thermophile* and *T.terrestris* have positive evidence for RIP mechanism but they have abundant transposable elements.

Figure 2.1 Distribution of TE superfamilies in seven fungal genomes (%). Ncra = *N.crassa,* Ntet = *N.tetrasperma,* Ndis = *N.discreta,* Smac = *S.macrosporus,* Cglo = *C.globosum,* Sthe = *S.thermophile,* Tter = *T.terrestris.* Different colors represent different kinds of transposable elements, like SINE in red, LINE in light green, LTR elements in green, DNA transposons in dark blue and unclassified elements in purple. From the table, Class I elements and unclassified elements occupy a much greater proportion than Class II elements in these fungi and LTR elements are the dominant type of transposons in two thermophilic fungi-*S.thermophile* and *T.terrestris.*

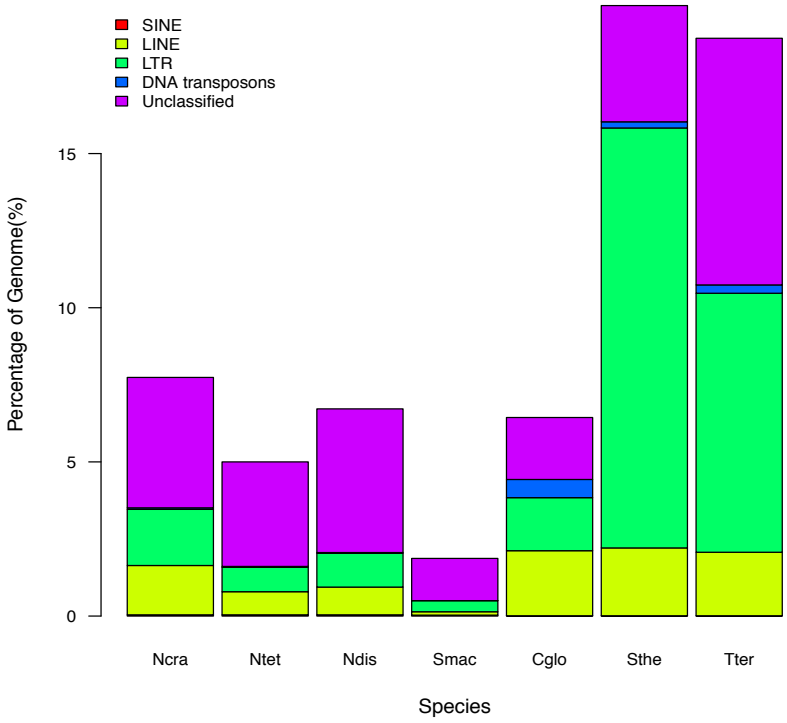


Table 2.1 Information of strains and versions for the seven fungi

Species	Strain	Version	Link
N.crassa	FGSC 2489	10	http://www.broadinstitute.org/annotation/genome/neurospora/AssemblyStats.html
N.tetrasperma	FGSC 2508	1	http://genome.jgi-psf.org/Neute_matA2/Neute_matA2.info.html
N.discreta	FGSC 8579	1	http://genome.jgi-psf.org/Neudi1/Neudi1.info.html
S.macrosporus	HELL	3	http://www.ncbi.nlm.nih.gov/pubmed/20386741
C.globosum	-	1	http://www.broadinstitute.org/annotation/genome/chaetomium_globosum.2/Assembly.html
S.thermophile	ATCC 42464	1	http://genome.jgipsf.org/Spoth1/Spoth1.info.html
T.terrestris	-	1	http://genome.jgipsf.org/Thite1/Thite1.info.htm

Table 2.2 Genome content of different TE superfamilies by RepeatMasker (%)

Species	SINE	LINE	LTR	DNA Elements	Unclassified	Total
N. crassa	0.04%	1.6%	1.82%	0.05%	4.23%	7.74%
N. tetrasperma	0.04%	0.75%	0.79%	0.03%	3.39%	5%
N. discreta	0.04%	0.9%	1.1%	0.02%	4.66%	6.72%
S. macrosporus	0.03%	0.11%	0.36%	0%	1.37%	1.87%
C. globosum	0.01%	2.11%	1.72%	0.59%	2.01%	6.44%
S. thermophile	0.01%	2.2%	13.62%	0.2%	3.77%	19.8%
T. terrestris	0.01%	2.06%	8.4%	0.27%	8%	18.74%

Chapter III

LTR element identification and evolution in seven closely related filamentous fungi

Introduction

One of the best studied type of transposable elements are transposable elements with long terminal direct repeats (LTR TEs) and a typical LTR element structure is a polyprotein gene flanking by two long terminal direct repeats (LTRs). Due to the different order of the domains in the *pol* ORF, LTR retrotransposons are classified into five superfamilies: *Ty1/Copia, Ty3/Gypsy, Bel/Pao,* ERV (endogenous retroviruses) and retroviruses. Previous studies show that two types of transposons are highly represented in plant genomes, which are long terminal repeat (LTR) retrotransposons and miniature inverted-repeat transposable element (MITEs) and can account for up to 50~80% of the genome. And scientists are more and more interested in them since both these two kinds of transposons are found to be associated to genes, suggesting that their activity has influence on the evolution of plant genes (Casacuberta et al. 2003).

The number and content of LTR retroelements in analyzed fungi differs differently. Some genomes have abundant LTR TEs (up to 8000 elements) while others have only a few of them (< 50 elements). I characterized about 1,310 transposable elements with a typical LTR TE structure in the seven fungal genomes. By searching in the literature library, I found there still lacked an analysis of LTR TEs in fungi across the whole kingdom yet. The published literature is mostly focused on single genomes, like in *N. crassa* or *M. oryzae.* Muszewska has done a large-scale search for LTR retrotransposons in 59 fungal genome sequences (Muszewska et al. 2011). And the

studies show that the transposon proliferation in fungi usually involves an increase in both the copy number of individual elements and in the number element types. The majority of the highest-copy LTR transposons in these 59 genomes are from *Ty3/Gypsy*. A phylogenetic analysis of these LTR elements showed that TE expansions appeared independently of each other at different taxonomical levels and in distant genomes. I would also take a look at the LTR transposons in my seven studied fungi and used a phylogenetic tree to deduce the evolutionary relation of them.

Materials and Methods

From the annotation results in study 1, Class I transposons especially LTR elements occupy a dominant role among all these TE superfamilies across the seven genomes. Though RepeatMasker gave a broad view of the genome content of LTR elements in each genome, it did not gave a good performance in identifying long LTR-containing elements with an intact structure and prediction of the number of identified LTR-containing elements. The reason is that RepeatMasker program fragments the LTR retroelement model in Repbase Update into multiple domains like *gag,* INT*,* AP, RT, RH and this results in a many-to-one relationship with Repbase full-length entries (described in the website of RepeatMasker). By looking into the annotated library of RepeatMasker, I found that one full-length LTR element could be divided into several short-lengthed fragments. Therefore, only by looking at RM results, it was not sufficient to obtain complete transposable elements with an intact protein-coding region and boundaries. I used a structure-based TE annotation approach for characterization of LTR transposons called LTR-harvest.

LTR-harvest is a structure-based approach for identifying and annotating LTR-containing transposable elements programmed in C. The theory base for LTR-harvest is that a typical LTR retrotransposon is composed of an internal region containing several open reading frames, two long terminal repeats which are nearly identical, flanked by target site duplications of usually 4~6bp. The internal region usually involves ORFs such as integrase, protease, reverse transcriptase, *gag* gene encoding for structural proteins of virus-like particle. It is also observed that in some rare cases an *env-like* gene indispensible for retroviruses life cycle is present in LTR retrotransposons. In addition, there are some other structure features which can be taken into account, e.g. the primer binding site acting as the starting point for reverse transcription and a purine rich sequence at 3' end of the internal region named the poly purine tract. According to these structure features, software designers implemented several parameters into the model: length constraint, distance constraint, similarity constraint, target site duplications, LTR motifs and so on. And these parameters are flexible and users can define them with previous knowledge about the features of the genome. In LTR-harvest results, it gave the sequences of identified LTR transposable elements in FASTA and gff3 format.

There are three advantages of LTR-harvest: 1) it can run fast on large sequence data sets in FASTA format, for example, it takes only 8 minutes on a Linux PC with 4 GB of memory to process the largest human chromosome; 2) Compared to other LTR transposon-finding software like LTR_STRUC and LTR_Seq, LTR-harvest has comparable high sensitivity but much better specificity with specific parameter setting on *Drosophila* test data; 3) Flexible parameter settings make it convenient for uses to

incorporate biological features like TSD length, motifs, LTR distance and length. However, as we know, mutations during evolution bring incomplete and degenerate LTR sequences and disruption of internal open reading frames. Insertions or deletions of other kinds of transposons may lead to nested or truncated LTR retrotransposons as well. One kind of LTR retrotransposons are very common as a result of homology recombination between the two LTRs, called solo LTRs. Therefore, these LTR retrotransposons wouldn't include the previous structure features that a canonical one should have and may be missed by the previous structure model. And a further homology searches with full-length LTR retrotransposns will help detect these special cases (Ellinghaus et al. 2008).

After LTR-harvest program identified LTR TEs in each fungal genome, I carried out a TBLASTN search using two known reverse transcriptase (RT) domains (*Copia* and *Gypsy*) from one of the fungal genome- *C.globosum* from PFAM and obtained the corresponding region for RT domain in these genomes under E-value of 10^{-3}. Dr. Jason Stajich wrote a Perl script called "extract_tblastn.pl" (http://code.google.com/p/stajichlab/source/browse /#svn/trunk/transposon) to extract the RT domain DNA sequence from the genome according to the blast results. These RT domain sequences from 7 species were joint together and I carried out a multiple sequence alignment using the program of MUSCLE (http:// www.drive5.com/muscle/) and used a perl script developed by our lab called "bp_sreformat.pl" to transform the ".fasaln" format into nexus format for building phylogenetic tree by MrBayes Program.

MrBayes is a program based on the Bayesian estimation of phylogeny (Huelsenbeck et al. 2001). Bayesian inference of phylogeny is based upon a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observation data. The conditioning is accomplished using Bayes' theorem. MrBayes uses a simulation technique called Markov chain Monte Carlo (or MCMC) to approximate the posterior probabilities of trees. MrBayes uses the NEXUS file format for input and has abundant evolutionary models including 4 by 4, doublet, and codon models for nucleotide data and many standard rate matrices for amino acid data. It has parallel version for high-throughput computing. I wrote a bash script called "sordaries_domians_mrbayes.sh" and run it on the data set.

Results

My previous results showed LTR elements were the dominant TE type in seven closely related fungi. In order to understand the evolution history of LTR elements, a comparative analysis was carried out using LTR-harvest. LTR-harvest is efficient software at identifying full length LTR retrotransposons at large sequence sets (Ellinghaus et al., 2008). It is based on the known structure features of LTR retrotransposons, like the presence of LTR pairs, the 4~6-bp TSDs, the primer binding site (PBS) and polypurine tract (PPT), as well as length, distance and sequence motif, etc. The prediction result by LTR-harvest is illustrated in table 3.1. Next, a *copia*- and *gypsy*-specific reference sequence were constructed by downloading the most conserved region of the RT (reverse transcriptase) domain from Pfam family PF07727 and PF00078, respectively. Finally, all RT domains from the LTR element candidates were extracted

and used for building a phylogenetic tree with the program of MrBayes. The whole process is illustrated in Figure 3.1. And the generated phylogenetic tree of the RT domains is shown in Figure 3.2.

From the tree, I observed three things: 1) *C. globosum, S. thermophile,* and *T. terrestris* have much more LTR retrotransposons than others and some of the families of LTR retrotranspons is largely expanded; 2) There are cases where the RT domain from *gypsy* elements and *copia* elements are clustered together which may suggest that the RT of *gypsy* and *copia* may have a common ancestor; 3) Some RT domains from different species cluster together which may suggest that there may be horizontal DNA transfer between these species or *gypsy* and *copia* elements.

Conclusion

By using LTR-harvest program, about 1,300 transposable elements with long terminal repeats (LTR TEs) were identified and characterized from the seven fungal genomes. These reverse transcriptase (RT) domain sequences of LTR retroelements were then used to construct a phylogenetic tree to reveal the evolutionary relationship of these elements by MrBayes software. From the tree, it was obvious to see that in most cases, the expansion of LTR elements were independent of each other but there are cases where the RT domains from different species (*N. crassa, N. tetrasperma, S. thermophile* and *T. terrestris*) were grouped together, suggesting that these LTR elements might come from the same transposon family in the ancestor species of the two mesophilic fungi and the two thermophilic fungi millions of years ago. Moreover, in some cases, RT domains from *Ty1/copia* and *Ty3/gypsy* superfamilies were grouped together, suggesting these two

superfamilies of LTR retroelements may be also derived from the same transposon family long ago.

Figure 3.1 The process of constructing a phylogenetic tree of RT domains of all identified LTR elements. Firstly, load genomes into LTR-harvest program to obtain sequences of identified LTR candidates; secondly, use BLASTX to extract RT domains and finally construct a phylogenetic tree usying MrBayes.

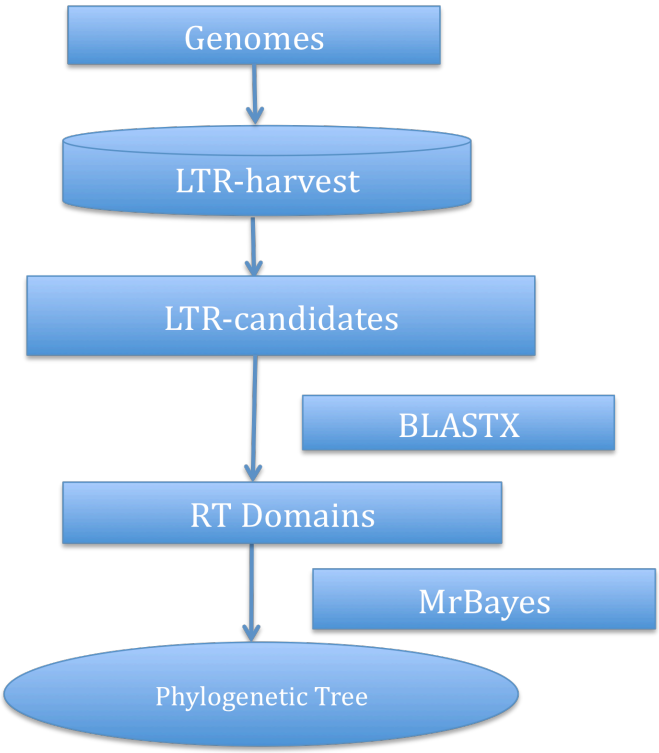


Figure 3.2 The phylogenetic tree based on RT domains from indentified LTR elements. From the tree, some RT domains of *Copia* and *Gypsy* LTRs from different species have common ancestors. And for *C.globosum*, *S.thermophile* and *T.terresstris,* they have big families with members really close to each other, which shows the possibility of recent activities.

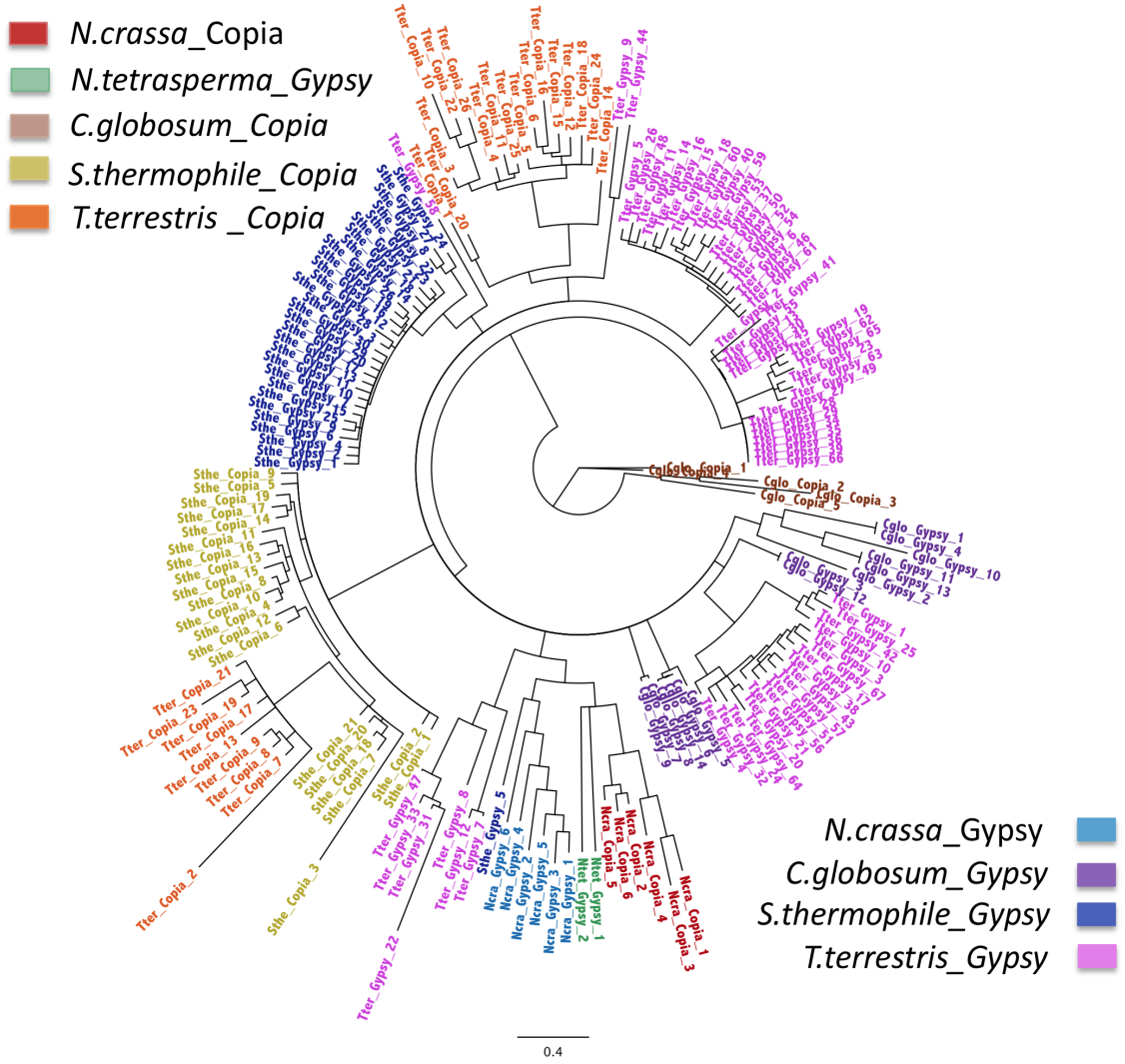


Table 3.1 The distribution of LTR elements in seven fungi by LTR-harvest

Species	Numbers of LTR elements	Total amount of LTR elements by bps (Genome content by percentage)
N.crassa	94	715222(1.82%)
N.discreta	92	412459(1.10%)
N.tetrasperma	64	298770(0.79%)
C.globosum	47	590132(1.72%)
S.macrosporus	25	143251(0.36%)
S.thermophile	644	5277372(13.62%)
T.terrestris	387	3100048(8.40%)

Chapter VI

Characterization and identification of DNA transposons in seven fungal genomes

Introduction

Class II transposable elements, or called "cut-and-paste" transposable elements, uses a transposase instead of a reverse transcriptase via transposition and do not generate an RNA intermediate in the process. Thus they are also called DNA transposons. The first transposable element discovered in maize by Barbara McClintock in the 1940s (*Ac/Ds*), P element of *Drosophila,* and *Tc1* element of *Caenorhabditis elegans* all belongs to DNA transposons. In structure, there is a pair of terminal inverted repeat (TIRs) flanks the transposase sequence. During transposition of Class II elements, the transposase cuts out the original copy from the donor site and insert it elsewhere in the genome. It causes a double-strand break at the donor site and this break can be restored by host repair mechanism and in this way the copy number increases. And the insertion of the new genome locus results in a target site duplication (TSD) with a length of 2~10bp.

Class II elements can be grouped into large superfamilies mainly based on the sequence similarity of the transposase, eg. *Tc1/Mariner*, *PiggyBac*, *PIF/Harbinger, hAT*, etc. Sequences of transposable elements with an E-value less than 0.01 in PSI-BLAST or BLASTP searches are clustered into a same superfamily (Jurka et al. 2005). Up to date, there are 19 superfamilies of DNA transposons, namely *CACTA (En/Spm), hAT, Mutator (MuDR), Mirage, Merlin, Tc1/mariner, P, PiggyBac, PIF/Harbinger, Transib, Novosib, Rehavkus, ISL2EU, Kolobok, Sola, Zator, Chapev, Academ and Ginger.*

The most active DNA transposon element identified to-date in plants or animals is *mPing,* which is a rice *Tourist-like* MITE derived from the autonomous *Ping* element. Some rice strains accumulate ~40 new *mPing* insertions per plant per generation (Hancock et al. 2010). In my study I used a phylogenetic-based approach to identify DNA transposons to see if there were any "active" DNA transposons.

It was found that protein domains containing an acidic amino acid triad (DDE or DDD) were present in the transposase sequences of 11 superfamilies of "cut-and-paste" transposons. This protein domain catalyzes the transposition reaction and forms a characteristic RNase H-like fold composed of α-helices and β-strands. From the 3D structure of the DDE/D triad for members of *hAT* and *Tc1/mariner,* this catalytic domain forms a catalytic pocket containing two divalent metal ions which assist in the reactions during DNA cleavage. Previous studies show that each DNA transposon superfamily pocesses a superfamily-specific "signature string" consisting of multiple conserved amino acid residues and motifs within the DDE/D domain (Yaowu et al. 2011). This is the theory base for phylogenetic-based TE annotation approach.

Due to their transposition mechanism, DNA transposons plays a very important role in gene tagging, mutagenesis, molecular biology, biotechnology, gene therapy and so on. The Sleeping Beauty transposons system (SBTS) is one of the most well-known example. SBTS is a non-viral vector system using a synthetic DNA transposon that was constructed to introduce precisely defined DNA sequences into chromosomes of vertebrate animals especially into human beings for the goal of replacing a defective gene, introducing new traits and discovering new genes and their functions. Four patent

applications regarding the use of SBTS for introducing new DNA into the chromosomes of a cell have been filed. It is believed that this SBTS system offers a leading and long-lasting way to insert genes into chromosomes of cells without using a virus and a very important solution in gene therapy. In my study, my goal is to identify DNA transposons and check if there are "active" elements or high-copy transposon families like MITEs in rice.

Materials and Methods

I have built up an integrated pipeline for TE annotation, which can be divided into the following parts:

1) Annotation of Class I elements:

I chose the program of LTR-harvest in my pipeline for LTR element prediction and RepeatMasker for non-LTR elements like SINEs, LINEs, etc.

2) Annotation of Class II elements

Based on the high conservancy in DDE/D domain for each superfamily of DNA transposons, I used a large collection of representative DDE domains for different DNA transposon superfamily as a query and searched against the studied genome to identify DNA transposable elements. Since each superfamily of DNA transposons have their own signature for Terminal Inverted Repeats and Target Site Duplications, this information could be used to identify the boundaries for each element and obtain the full-length element sequence.

2) Annotation of Non-autonomous elements

Both class I and class II elements have their own non-autonomous members which are derived from autonomous elements but lose the ability to synthesis reverse transcriptase or transposase through mutation. To find these incomplete and relatively short elements, I carried out a Blast search with the well-annotated transposons sequence in the previous steps against genome and used the program of MITE-Hunter, which is designed to find non-autonomous DNA transposable elements.

3) Discovery of active or potentially active transposable elements

In the process of annotation, if one element has several almost identical or highly similar copies, the pipeline then checks whether there is a stop codon in the coding region and if the coding region for reverse transcriptase or transposase had a significant similarity to known reverse transcriptase or transposase. If there is no stop codon and the coding region is still intact, it was probably an "active" or "potentially active" element.

Results

Compared with RNA retroelements, DNA transposable elements only occupied a small proportion in the seven fungal genomes. There are three superfamilies in my annotated TE library collecting DNA transposable elements, which are *Tc1/Mariner, PIF/Harbinger* and *MULE*. And *Tc1/Mariner* is the dominant type among the three superfamilies (shown in Table 4.1). There were 66 *Tc1/Mariners* and 28 *MULEs* in the pathogenetic fungi *C. globosum,* which is significantly higher than in other fungi. And there was only one *PIF/Harbinger* element in *N. crassa, N. discreta, N. tetrasperma,* and *C. globosum.* There were 29 and 10 *Tc1/Mariners* in the two thermophilic fungi as well.

In the process of annotating DNA transposons, I identified a small *Tc1/Mariner* family with three autonomous members and five non-autonomous members in *C. globosum,* of which two elements were 100% identical sharing 80% sequence similarity with the rest one. I also predicted the transposase genes in the three autonomous elements using FGENESH (http://linux1.softberry.com/berry.phtml). And there are no stop codons in the ORFs and sequence encoding for transposase in *C. globosum* shared around 46% amino acid sequence similarity with the well-known transposase sequence in another fungus *Penicillium marneffei.* All of these results provided positive evidence that this small "*Tc1/Mariner*" family might be "active" or have the potential to be reactive again.

Conclusion

Unlike RNA transposons, DNA transposons only occupied a small proportion in the genome content of these seven fungi and there were not so many copies in them. In *C. globosum,* the copy number of two transposon superfamily- *Tc1/Mariner* and *MULE* were significantly higher than in other fungal genomes, suggesting there may be a "burst" of expansion and transposition in these elements in this fungus during its independent evolution. Similar results were observed in the two thermophilic fungi (*S. thermophile* and *T. terrestris*). Future work includes constructing a phylogenetic tree among *Tc1/Mariner* elements across seven species to observe the relation between them in evolution.

Table 4.1 Number of DNA transposable elements identified in 7 fungi

Species	*Tc1/Mariner*	*PIF/Harbinger*	*MULE*	*hAT*	*PiggyBac*
N.crassa	14	1	7	0	0
N.discreta	14	1	0	0	0
N.tetrasperma	8	1	3	0	0
C.globosum	66	1	28	0	0
S.macrosporus	2	0	0	0	0
S.thermophile	29	0	0	0	0
T.terrestris	10	0	1	0	0

Chapter V

Repeat-Induced Point Mutation (RIP) defense mechanism in seven fungal genomes

Introduction

Selfish elements like transposable elements are ubiquitous in all eukaryotes and different biological organisms have developed different levels of genome defense mechanism including both transcriptional gene silencing and post-transcriptional gene silencing mechanisms to protect the genome from being disrupted by mobile elements. And Repeat-induced Point mutation mechanism is one of the defense mechanisms which induces mutation directly in the genomic sequence and is found only in fungi. RIP was firstly discovered by Selker in *N. crassa* strains containing experimentally induced duplications during sexual cycle (Selker et al. 1987). RIP induces C:G to T:A mutations into duplicated sequences and was considered as a defense against the proliferation and expansion of transposable elements. Studies found that RIP-mutated sequences are frequent targets for methylation, which results in transcription silencing in *Neurospora* but the relationship between RIP and DNA methylation is still uncertain. Just like a double-edged "sword", RIP also has an evolutionary cost, since it affects multiple gene families, as well as transposable elements. For example, in *N. crassa*, there are only 6 pairs of genes among the putative 10,082 protein coding genes sharing >80% amino acid similarity.

RIP has been demonstrated and validated experimentally in *N. crassa, P. anserina* (Graïa et al., 2001; Bouhouche et al., 2004; Arnaise et al., 2008), *M. oryzae* (Ikeda et al. 2002), *Leptosphaeria maculans* (Idnurm and Howlett et al. 2003) and *Nectria*

haematococca (Coleman et al. 2009). Effects of RIP were not observed in experiments of *Aspergillus nidulans* carrying transformation-induced duplications at sexual cycle, suggesting this fungus does not have the RIP defense mechanism (Lee et al. 2008). And there was also no experimental evidence to support that RIP exists in *Sordaria macrospora* (Le Chevanton et al. 1989; Walz and Kuck et al. 1995). In *M. oryzae,* although RIP effects were not obviously as frequent and severe as in *N. crassa,* examination of transposon sequences strongly suggested that RIP did occurred (Nielsen et al. 2001; Clutterbuck, 2004; Galagan et al. 2005; Clutterbuck et al. 2008).

In this study, I sought to identify evidence for RIP in the seven fungal genomes and how frequently it occurred in each of them. Since the mechanism for RIP has been poorly understood, I was interested in the relation between RIP and the length of TEs and would test the hypothesis that RIP preferred to act on longer transposons than short ones.

Materials and Methods

Genomes from seven ascomycetes were obtained from Joint Genome Institute (http://www.jgi.doe.gov/) and the genome information has been shown in Table 2.1.

Overlapping windows of the genome were characterized as subjected to RIP or not based on the indices calculated from di-nucleotide frequencies. Two RIP indicies were calculated for each window: the "RIP product index" (TpA/ApT) and the "RIP substrate index" (CpA + TpG/ ApC + GpT) (Margolin et al. 1998; Selker et al. 2003). A window was considered RIPed if the composite RIP index (CRI), which is the substrate index subtracted from the product index, had a value greater than 0. These values were computed with a Perl script written by Dr. Jason E. Stajich

(http://code.google.com/p/stajichlab/source/browse/trunk/RIP/calculate_RIP_index_windows.pl) by scanning the whole genome in windows of ~1kb with an overlap of 200bp between windows. The results were summarized with a R script into barplots.

In Chapter II, III and IV, I obtained an annotated TE library for each of the 7 genomes collecting both Class I elements and Class II elements, of which most are autonomous elements. Using the Perl script I wrote, "RIP_index_length.pl" ((http://code.google.com/p/stajichlab/source/browse/trunk/RIP/), I calculated the three RIP indices (substrate index, product index and composite RIP index) for each element and and the total basepairs of element to test if there is a relation between the length of TE and RIP.

I also used MITE-Hunter software developed by Yujun Han who was a postdoc in Dr. Susan R. Wessler's lab, a structure-based program pipeline that can identify transposablements with TIP or TSD, to search for miniature inverted-repeat transposable elements (MITEs) in these fungi (Yujun Han et al. 2010) and used these short transposon sequences as test data in my analysis of RIP against transposons of different lengths.

Results

Based on RIP indices, I found evidence of RIP in 5 filamentous ascomycete fungi shown in Figure 5.1 (*N. crassa, N. tetrasperma, N. discreta, S. thermophile,* and *T. terrestris*), but no evidence of RIP in *C. globosum* and *S. macrosporus* shown in Figure 5.2.

Both class I and class II elements have their own non-autonomous members which are derived from autonomous elements but lose the ability to synthesis reverse

transcriptase or transposase through mutation. To find these incomplete and relatively short elements, I Blasted the annotated autonomous transposons in the previous steps against genome and used MITE-Hunter, which is designed to find non-autonomous DNA transposable elements. The total number of Miniature Inverted-repeat Transposable Elements (MITEs) I identified is shown in Table 5.1

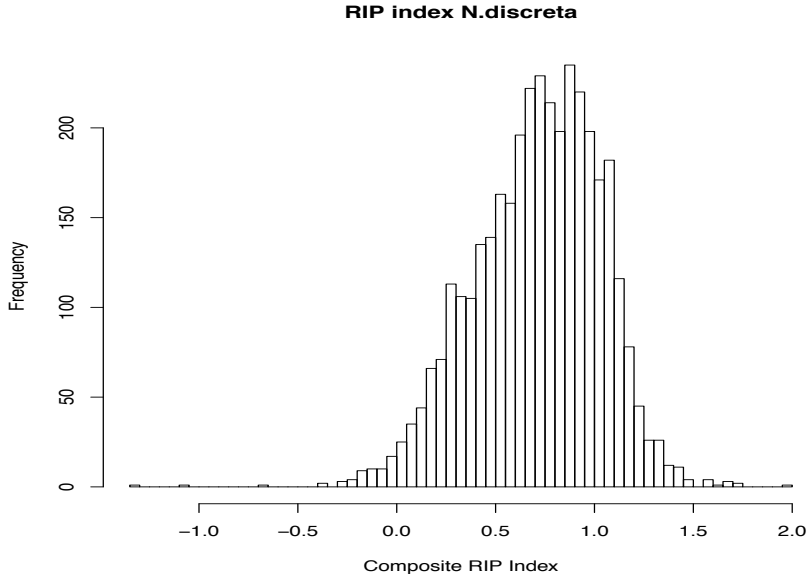
Previous studies in *N. crassa* found that RIP requires ≥ 80% identity over a length of ≥ 400 bp for tandem repeats (Galagan et al. 2004). My hypothesis is that shorter elements may be more likely to escape RIP than longer elements. From the current annotation results, I carried out an analysis of the relationship between the RIP index and the element length. The results show that almost all transposons longer than 1,500 bp showed evidence for RIP and some TEs in *N. crassa* shorter than 400bp still showed evidence for RIP as shown in Figure 5.3.

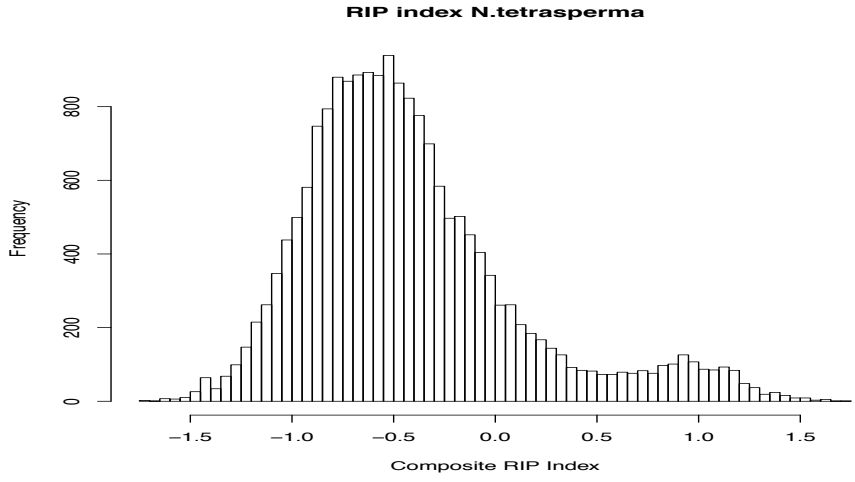
Conclusion

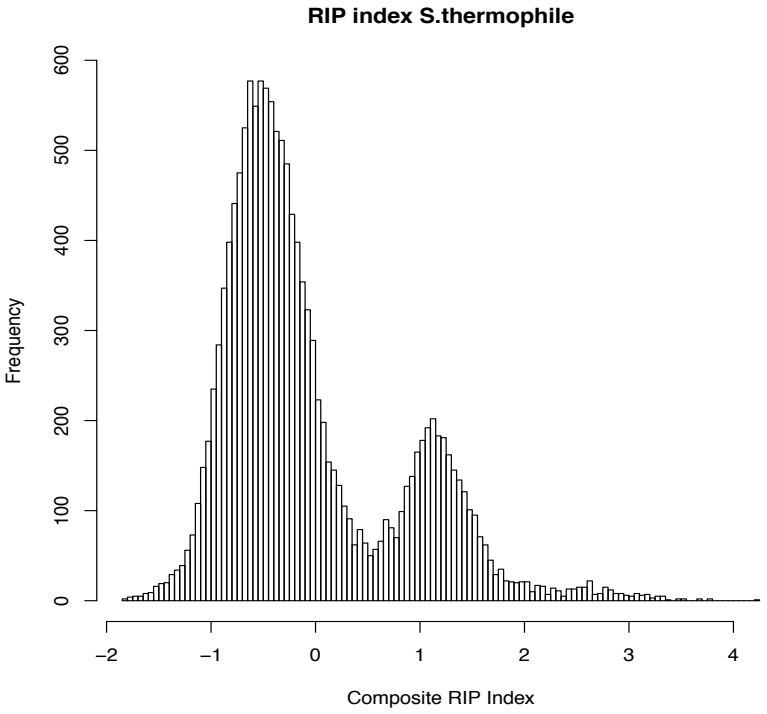
I found evidence of RIP using the RIP indices in 5 of the seven fungal genomes including the two thermophilic fungi and no typical RIP patterns in the other two, suggesting that they may lack the molecular pathway for RIP or RIP may be very mild and have little influence on the genome. Moreover, I used the identified collection of MITEs (Miniature Inverted Transposable Elements, which are usually 200~500bp long) from seven genomes obtained from MITE-Hunter to represent short transposable element and the manually annotated library of full-length DNA transposons to represent long TEs to test the relationship between the length of TE and RIP mechanism. Though there were some short ones still being affected in *N. crassa,* short transposons were more likely to

escape from RIP defense. And in all the five fungal genomes found to have RIP, my annotated DNA transposons longer than 1,500bp always showed signatures of RIP, suggesting that longer transposons were more likely to be detected by the genome defense. Future work is necessary to test this hypothesis in a broader fungal transposon library with a collection of more transposable elements of different lengths from more families and more species to achieve higher statistical significance.

Figure 5.1 Species showing RIP pattern.
A composite RIP index (**CRI**) can be determined by subtracting the substrate index from the product index; thus, a positive CRI value implies that the DNA has been subjected to RIP. These graphs show a CRI computed for non-overlapping window of 1kb. A large fraction of genomes of these fungi have a positive CRI indicating that they have been RIPed.






RIP index S.thermophile
Frequency
0
100
200
300
400
500
600
−2
−1
0
1
2
3
4
Composite RIP Index

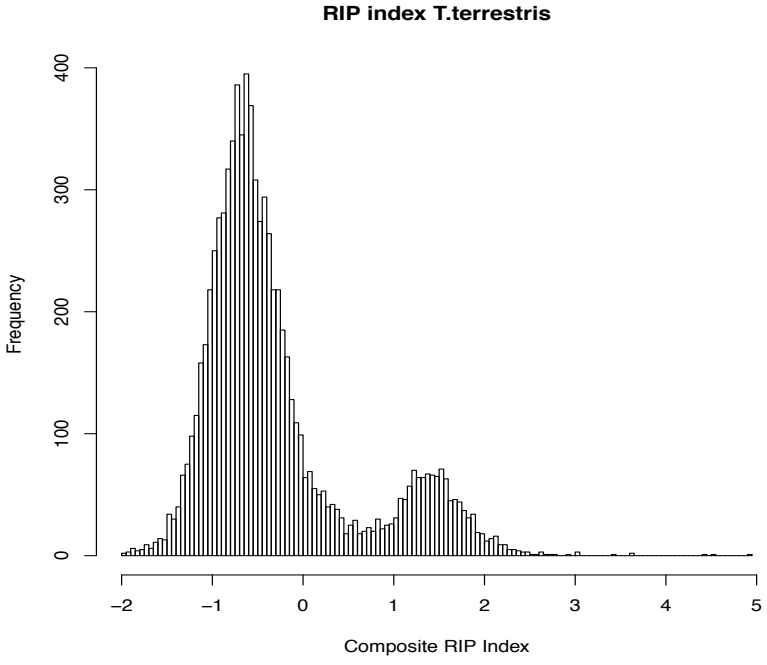
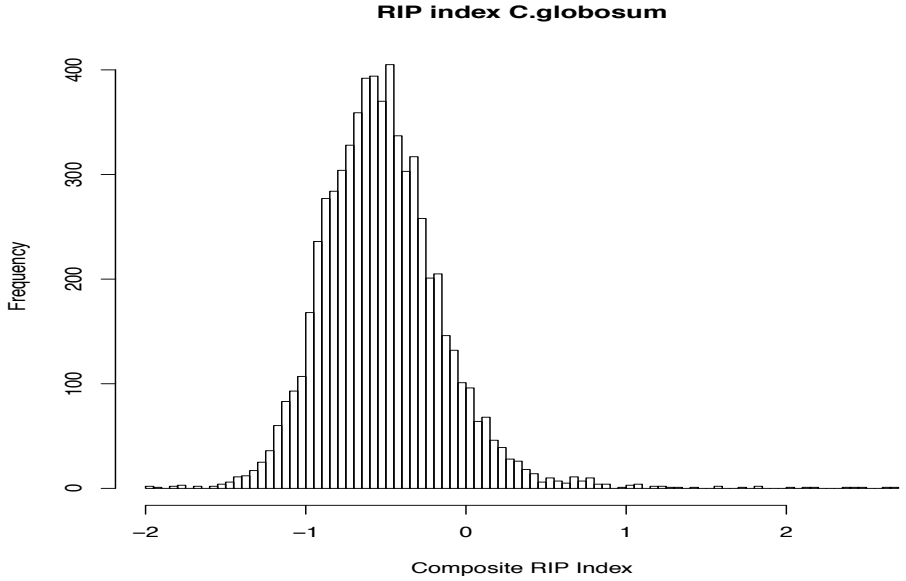

RIP index T.terrestris
Frequency
0
100
200
300
400
−2
−1
0
1
2
3
4
5
Composite RIP Index

Figure 5.2 Species not showing RIP pattern. Compared to the other genomes, *C.globosum* and *S.macrosporus* do not show an abundance of sequences with positive Composite RIP Index (CRI) suggesting these species may not have an intact RIP defense mechanism or have relatively weak RIP mechanism.



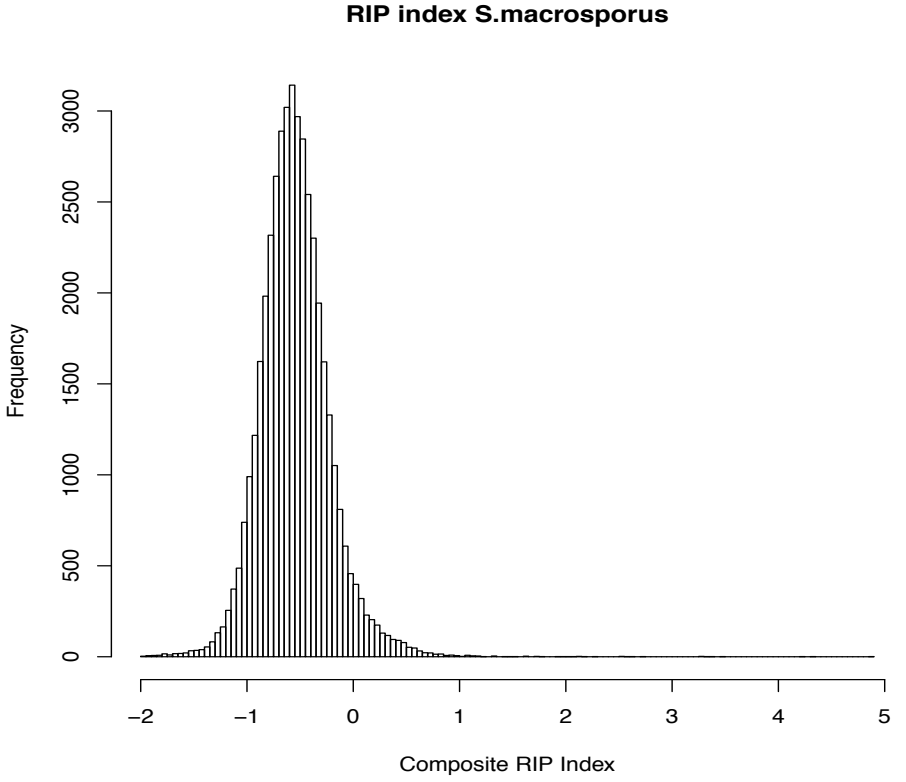


Figure 5.3 A plot showing the Composite RIP index (CRI) of TEs of different lengths. Short (length under 500bp) but still RIP-mutated transposable element in *N.crassa* are shown in red dots. Sequences with positive Composite RIP index (CRI) are evidence for RIP.

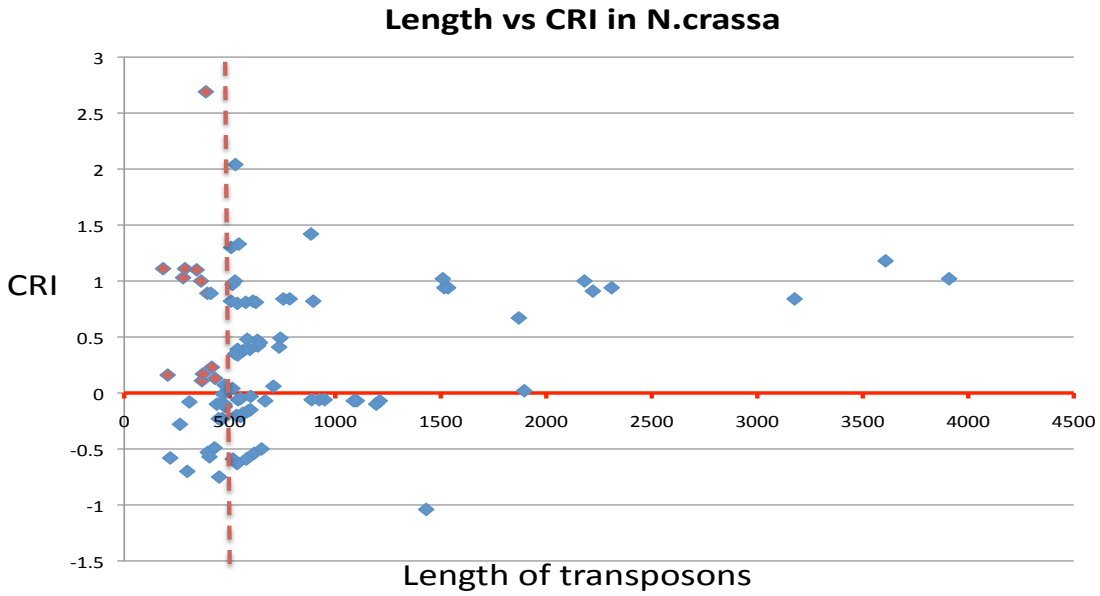


Table 5.1 Total numbers of MITEs by MITE-Hunter in seven fungal genomes.

Species	Number of MITEs
N. crassa	5195
N. discreta	4943
N. tetrasperma	5013
C. globosum	4566
S. macroporus	5879
S. thermophile	5161
T. terrestris	4918

Chapter VI

Summary of Work

In order to obtain a broad view of the distribution of different types of transposable elements and their genome percentage, I carried out an analysis based on RepeatMasker and RepeatModeler. My results showed that RNA transposons are much more frequent in the seven Sordariomycete genomes than DNA transposons. LINEs and LTR-containing transposons are the two most abundant types found in all species. The distribution patterns for different superfamilies of transposons varied among organisms with *S.thermophile* having the most transposons and *S.macrosporus* having the fewest transposons. Each fungal organism contained a considerable proportion of unclassified elements which remain to be curated and investigated. In the two thermophilic fungal genomes (*S. thermophile* and *T. terrestris),* the proportion of LTR-containing elements in these genomes was significantly higher (up to 19.8% and 18.74% respectively) than the other five genomes, suggesting that LTR-containing elements might have a recent transpositional 'burst' in these two genomes and there might be some "active" LTR transposable elements in them.

As LTR transposable elements were the dominant type of elements in the seven genomes, I used a program that is targeted to find these elements based on its structural features called LTR-harvest. The identified elements were classified into two major superfamiles: *Ty1/Copia* and *Ty3/Gypsy*, based on their domain organization. Moreover, I utilized Mrbayes to build a phylogenetic tree based on the reverse transcriptase domain sequences on these elements and found that LTR transposons have diversified and

expanded in copy number independently in each species lineage in most cases. There are situations that LTR transposons from different fungi were clustered together, suggesting that there might be "horizontal transfer" between these species or these elements were derived from the same transposons family in the common ancestor of these seven species. Groups of elements that cluster in the tree with short branches and have high sequence similarity were frequently found in *C.globosum, S.thermophile* and *T.terrestris,* providing evidence for recent transposition activity of LTR transposons.

As different DNA superfamilies have their own "signature strings" for conserved sequences of the DDE/D catalytic domain, a phylogentic-based approach was used to characterize the DNA transposons in these fungi. In *C.globosum*, the number of *Tc1/Mariners* and *MULEs* was significantly higher than these types of elements in other fungi, suggesting these superfamilies might have experienced a transposition "burst" and there may be "active" DNA transposons in this fungus. However they do not have completely full length sequences so may have been active in the past but may lack the ability to transpose in the present. Additional experimental research to explore whether these are active through transposition assays would confirm these observations.

Repeat-induced Point Mutation (RIP) is a very important fungus-specific defense mechanism against repetitive elements especially transposons and was found in 3 *Neurospora species* (*N. crassa, N. discreta, N. tetrasperma)* and 2 thermophilic fungi (*S.thermophile* and *T.terrestris*) according to a RIP-index analysis. However, the RIP pattern in the latter two genomes is not as pronounced the thermophilic fungi as in the *Neurospora*, suggesting that RIP may not be as efficient or severe in the two thermophilic

fungi. By using a Perl script to calculate the RIP indices and element length, I observed that RIP is likely more effective against transposable elements longer than 1,500bp but that it is still seen affect some very short ones, in sizes that are less than what has been empirically shown for RIP.

Discussion and Future Work

In my study, I observed two paradoxes. One is that that *C. globosum* and *S. macrosporus* may not have RIP defense, but their genomes remain relatively uncolonized by transposable elements. There are several reasons to explain this. One might be that these two fungi were sequenced by whole-genome methods and repeat regions such as transposable elements might not be easily assembled, resulting in missing parts of transposons in annotation. Improving the assembly quality by resequencing using Sanger sequencing and physical mapping may be necessary. However, Southern analyses did not reveal any obvious hidden repetitive sequences that were unassembled. Another explanation may be that another defense mechanism exists that controls the expansion and transposition of mobile elements without the use of the mutational RIP process. One of the possibilities is DNA methylation, which results in transcriptional silencing. Techniques like Bisulfite sequencing can be used to detect if transposon regions show patterns of methylation.

The other paradox is that *S. thermophile* and *T. terrestris* showed strong evidence for RIPed sequence in a large fraction of the genome, but their genomes are also rich in transposable elements, including both Class I and Class II types. There are two explanations. One is that the enzyme activity of the reverse trancriptase or transposase in

thermophilic fungi may be higher under a higher temperature and reduce the effectiveness of RIP. Future work could follow up on these observations in three steps. Firstly, it is necessary to carry out a multiple sequence alignment among the sequences of reverse transcriptase or transposase between mesophilic fungi and thermophilic fungi. Secondly, 3D structures of the transposase can be predicted and compared between two kinds of fungi by using protein-folding prediction software. Finally, biochemical experiments can be designed to test if these transposases show different enzyme activity under different temperatures. The other explanation is that transposable elements insert around some functional genes related to the adaption to high temperature of these fungi and have an influence on gene expression. Therefore, it is necessary to see the insertion locus of these transposons in thermophilic fungi using bioinformatics analysis. Additional experiment work, which measured transposase activity, could help assess the capabilities of these enzymes. As RIP requires a sexual cycle, it may be important to explore the frequency of sexual reproduction in these fungi through population genetics to better understand whether there is ample opportunity to silence these transposons. It may be that high temperature stresses induce transposition but without a frequent outcrossing these fungi may not have opportunity to invoke genome defense to mutate the duplicated elements. It would also be important to compare the makeup of duplicated genes from transposons to see if RIP is specifically limiting duplication frequency of all regions of the genome as in *N. crassa* or only the highly identical transposable elements.

As my previous RepeatMasker results show that the seven fungal genomes all carry a considerable amount of unclassified elements, there are several possibilities. They

may be transposons that are missed by *de novo* identification approaches, multiple copies genes or ESTs derived from transposons (types of false positive of the program), degenerated transposons that have been mutated by RIP and cannot be classified into any of the well-curated transposon superfamily or lineage-specific transposons of novel types. Future study includes similarity search to the well-anotated genes in GenBank and futher searches against all well-annotated TE libraries like Repbase that have been updated from detailed literature searching. A deeper detection for canonical transposable element properties such as target site duplications and terminal inverted repeats would help classify the superfamilies of the elements, which are likely to be DNA transposons. There is also the possibility that more exotic elements like Helitrons are part of the group of unclassified elements and would require additional curation to identify these as such.

Up to now, my pipeline of TE identification and annotation is composed of several manual steps. In future, it can be combined into an automatic process by several Perl scripts by a few lines of linux commands in future work. Moreover, with a good library of transposable elements, each superfamily can be classified into different small transposable element families with sequence clustering software like UCLUST or TribeMCL.

Previous studies in *N. crassa* found that RIP requires ≥ 80% identity over a length of ≥ 400 bp for tandem repeats. From Fig.9, I found that some TEs in *N. crassa* shorter than 400bp still showed evidence for RIP. However, almost all transposons longer than 1,500bp showed evidence for RIP. Since my current version of annotated TE libraries are not complete and the short repetitive elements were mostly MITEs (Miniature Inverted

Transposable Elements) derived from DNA autonomous transposable elements, it is necessary to integrate different approaches to identify more TEs and also characterize the non-autonomous members of Class I transposable elements and also the truncated derivatives from intact and full-length elements. With more transposons from all kinds of superfamilies and also special kinds of transposons like non-autonomous elements and truncated element, future researchers can run the Perl script "RIP_index_length.pl" again in the five fungal genomes with RIP mechanism (*N. crassa, N. tetrasperma, N. discreta, S. thermophile* and *T. terrestris*) and carried out a similar analysis to see if there are more evidence to support my hypothesis that transposable elements longer than 1,500bp would have little chance to escape from RIP defense mechanism.

A greater understanding of how genomes defend themselves against invading elements can have important uses for development of stable strains for industrial applications. By being able to eliminate identify, and then inactivate, all active transposable elements in an industrial strain, this will reduce the amount of genetic variability that will occur among generations. Increased understanding of how to modulate genome defense in fungi can also be a boon for laboratories attempting to create hypervariable strains that can be used to explore new biosynthetic production potential of strains or to create strains that can have a higher production of thermophilic enzymes or industrially important products.

Reference:

International Rice Genome Sequencing Project (2005) The map-based sequence of the rice genome. Nature 436:793-800

Arnaise S, Zickler D, Bourdais A, Dequard-Chablat M, Debuchy R (2008) Mutations in mating-type genes greatly decrease repeat-induced point mutation process in the fungus *Podospora anserina*. Fungal Genet Biol 45:207-220

Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 101 Suppl 2:14572-14579

Berka RM, Grigoriev IV, Otillar R, Salamov A, Grimwood J, Reid I, Ishmael N, John T, Darmond C, Moisan MC, Henrissat B, Coutinho PM, Lombard V, Natvig DO, Lindquist E, Schmutz J, Lucas S, Harris P, Powlowski J, Bellemare A, Taylor D, Butler G, de Vries RP, Allijn IE, van den Brink J, Ushinsky S, Storms R, Powell AJ, Paulsen IT, Elbourne LD, Baker SE, Magnuson J, Laboissiere S, Clutterbuck AJ, Martinez D, Wogulis M, de Leon AL, Rey MW, Tsang (2011) A Comparative genomic analysis of the thermophilic biomass-degrading fungi Myceliophthora thermophila and Thielavia terrestris. Nat Biotechnol 29:922-927

Bhat A, Kasbekar DP (2001) Escape from repeat-induced point mutation of a gene-sized duplication in *Neurospora crassa* crosses that are heterozygous for a larger chromosome segment duplication. Genetics 157:1581-1590

Bouhouche K, Zickler D, Debuchy R, Arnaise S (2004) Altering a gene involved in nuclear distribution increases the repeat-induced point mutation process in the fungus *Podospora anserina*. Genetics 167:151-159

Brosius J (1999) RNAs from all categories generate retrosequences that may be exapted as novel genes or regulatory elements. Gene 238:115-134

Goll MG, Bestor TH (2005) Eukaryotic cytosine methyltransferases. Annu Rev Biochem 74:481-514

Cambareri EB, Aisner R, Carbon J (1998) Structure of the chromosome VII centromere region in Neurospora crassa: degenerate transposons and simple repeats. Mol Cell Biol 18:5465-5477

Callinan PA, Batzer MA (2006) Retrotransposable elements and human disease. Genome Dyn 1:104-115

Cappello J, Handelsman K, Lodish HF (1985) Sequence of Dictyostelium DIRS-1: an apparent retrotransposon with inverted terminal repeats and an internal circle junction sequence. Cell 43:105-115

Capy P, Langin T, Higuet D, Maurer P, Bazin C (1997) Do the integrases of LTR-retrotransposons and class II element transposases have a common ancestor? Genetica 100:63-72

Casacuberta JM, Santiago N (2003) Plant LTR-retrotransposons and MITEs: control of transposition and impact on the evolution of plant genes and genomes. Gene 311:1-11

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

Clutterbuck AJ (2004) MATE transposable elements in *Aspergillus nidulans*: evidence of repeat-induced point mutation. Fungal Genet Biol 41:308-316

Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, Schmutz J, Taga M, White GJ, Zhou S, Schwartz DC, Freitag M, Ma LJ, Danchin EG, Henrissat B, Coutinho PM, Nelson DR, Straney D, Napoli CA, Barker BM, Gribskov M, Rep M, Kroken S, Molnar I, Rensing C, Kennell JC, Zamora J, Farman ML, Selker EU, Salamov A, Shapiro H, Pangilinan J, Lindquist E, Lamers C, Grigoriev IV, Geiser DM, Covert SF, Temporini E, Vanetten HD (2009) The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet 5:e1000618

Daviere JM, Langin T, Daboussi MJ (2001) Potential role of transposable elements in the rapid reorganization of the *Fusarium oxysporum* genome. Fungal Genet Biol 34:177-192

Sverdlov ED (2000) Retroviruses and primate evolution. Bioessays 22:161-171

Ellinghaus D, Kurtz S, Willhoeft U (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18

Ellison CE, Stajich JE, Jacobson DJ, Natvig DO, Lapidus A, Foster B, Aerts A, Riley R, Lindquist EA, Grigoriev IV, Taylor JW (2011) Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus Neurospora tetrasperma. Genetics 189:55-69

Esnault C, Maestre J, Heidmann T (2000) Human LINE retrotransposons generate processed pseudogenes. Nat Genet 24:363-367

Evgen'ev MB, Arkhipova IR (2005) Penelope-like elements--a new class of retroelements: distribution, function and possible evolutionary significance. Cytogenet Genome Res 110:510-521

Feng Q, Moran JV, Kazazian HH, Jr., Boeke JD (1996) Human L1 retrotransposon encodes a conserved endonuclease required for retrotransposition. Cell 87:905-916

Fincham JR, Connerton IF, Notarianni E, Harrington K (1989) Premeiotic disruption of duplicated and triplicated copies of the *Neurospora crassa* am (glutamate dehydrogenase) gene. Curr Genet 15:327-334

Freedman T, Pukkila PJ (1993) De novo methylation of repeated sequences in *Coprinus cinereus*. Genetics 135:357-366

Freitag M, Williams RL, Kothe GO, Selker EU (2002) A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. Proc Natl Acad Sci U S A 99:8802-8807

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859-868

Galagan JE, Selker EU (2004) RIP: the evolutionary cost of genome defense. Trends Genet 20:417-423

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

George ML, Nelson RJ, Zeigler RS, Leung H (1998) Rapid Population Analysis of *Magnaporthe grisea* by Using rep-PCR and Endogenous Repetitive DNA Sequences. Phytopathology 88:223-229

Gogvadze E, Buzdin A (2009) Retroelements and their impact on genome evolution and functioning. Cell Mol Life Sci 66:3727-3742

Han Y, Wessler SR MITE-Hunter (2010) a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res 38:e199

Hancock CN, Zhang F, Wessler SR (2010) Transposition of the Tourist-MITE mPing in yeast: an assay that retains key features of catalysis by the class 2 PIF/Harbinger superfamily. Mob DNA 1:5

Hane JK, Oliver RP (2010) In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. BMC Genomics 11:655

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755

Idnurm A, Howlett BJ (2003) Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the *Dothideomycete Leptosphaeria maculans*. Fungal Genet Biol 39:31-37

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

Jurka J, Kapitonov VV, Kohany O, Jurka MV (2007) Repetitive sequences in complex genomes: structure and evolution. Annu Rev Genomics Hum Genet 8:241-259

Kachroo P, Leong SA, Chattoo BB (1994) Pot2, an inverted repeat transposon from the rice blast fungus *Magnaporthe grisea*. Mol Gen Genet 245:339-348

Kazazian HH, Jr., Moran JV (1998) The impact of L1 retrotransposons on the human genome. Nat Genet 19:19-24

Kempken F (1999) Fungal transposons: from mobile elements towards molecular tools. Appl Microbiol Biotechnol 52: 756-760

Kempken F, Kuck U (1996) restless, an active Ac-like transposon from the fungus Tolypocladium inflatum: structure, expression, and alternative RNA splicing. Mol Cell Biol 16:6563-6572

Kempken F, Kuck U (1998) Transposons in filamentous fungi--facts and perspectives. Bioessays 20:652-659

Kennedy RC, Unger MF, Christley S, Collins FH, Madey GR (2011) An automated homology-based approach for identifying transposable elements. BMC Bioinformatics 12:130

Kordis D (2005) A genomic perspective on the chromodomain-containing retrotransposons: Chromoviruses. Gene 347:161-173

Kouzminova E, Selker EU (2001) dim-2 encodes a DNA methyltransferase responsible for all known cytosine methylation in Neurospora. EMBO J 20:4309-4323

Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W, Funke R, Gage D, Harris K, Heaford A, Howland J, Kann L, Lehoczky J, LeVine R, McEwan P, McKernan K, Meldrim J, Mesirov JP, Miranda C, Morris W, Naylor J, Raymond C, Rosetti M, Santos R, Sheridan A, Sougnez C, Stange-Thomann N, Stojanovic N, Subramanian A, Wyman D, Rogers J, Sulston J, Ainscough R, Beck S, Bentley D, Burton J, Clee C, Carter N, Coulson A, Deadman R, Deloukas P, Dunham A, Dunham I, Durbin R, French L, Grafham D, Gregory S, Hubbard T, Humphray S, Hunt A, Jones M, Lloyd C, McMurray A, Matthews L, Mercer S, Milne S, Mullikin JC, Mungall A, Plumb R, Ross M, Shownkeen R, Sims S, Waterston RH, Wilson RK, Hillier LW, McPherson JD, Marra MA, Mardis ER, Fulton LA, Chinwalla AT, Pepin KH, Gish WR, Chissoe SL, Wendl MC, Delehaunty KD, Miner TL, Delehaunty A, Kramer JB, Cook LL, Fulton RS, Johnson DL, Minx PJ, Clifton SW, Hawkins T, Branscomb E, Predki P, Richardson P, Wenning S, Slezak T, Doggett N, Cheng JF, Olsen A, Lucas S, Elkin C, Uberbacher E, Frazier M, Gibbs RA, Muzny DM, Scherer SE, Bouck JB, Sodergren EJ, Worley KC, Rives CM, Gorrell JH, Metzker ML, Naylor SL, Kucherlapati RS, Nelson DL, Weinstock GM, Sakaki Y, Fujiyama A, Hattori M, Yada T, Toyoda A, Itoh T, Kawagoe C, Watanabe H, Totoki Y, Taylor T, Weissenbach J, Heilig R, Saurin W, Artiguenave F, Brottier P, Bruls T, Pelletier E, Robert C, Wincker P, Smith DR, Doucette-Stamm L, Rubenfield M, Weinstock K, Lee HM, Dubois J, Rosenthal A, Platzer M, Nyakatura G, Taudien S, Rump A, Yang H, Yu J, Wang J, Huang G, Gu J, Hood L, Rowen L, Madan A, Qin S, Davis RW, Federspiel NA, Abola AP, Proctor MJ, Myers RM, Schmutz J, Dickson M, Grimwood J, Cox DR, Olson MV, Kaul R, Shimizu N, Kawasaki K, Minoshima S, Evans GA, Athanasiou M, Schultz R, Roe BA, Chen F, Pan H, Ramser J, Lehrach H, Reinhardt R, McCombie WR, de la Bastide M, Dedhia N, Blocker H, Hornischer K, Nordsiek G, Agarwala R, Aravind L, Bailey JA, Bateman A, Batzoglou S, Birney E, Bork P, Brown DG, Burge CB, Cerutti L, Chen HC, Church D, Clamp M, Copley RR, Doerks T, Eddy SR, Eichler EE, Furey TS, Galagan J, Gilbert JG, Harmon C, Hayashizaki Y, Haussler D, Hermjakob H, Hokamp K, Jang W, Johnson LS, Jones TA, Kasif S, Kaspryzk A, Kennedy S, Kent WJ, Kitts P, Koonin EV, Korf I, Kulp D, Lancet D, Lowe TM, McLysaght A, Mikkelsen T, Moran JV, Mulder N, Pollara VJ, Ponting CP, Schuler G, Schultz J, Slater G, Smit AF, Stupka E, Szustakowski J, Thierry-Mieg D, Thierry-Mieg J, Wagner L, Wallis J, Wheeler R, Williams A, Wolf YI, Wolfe KH, Yang SP, Yeh RF, Collins F, Guyer MS, Peterson J, Felsenfeld A, Wetterstrand KA, Patrinos A, Morgan MJ, de Jong P, Catanese JJ, Osoegawa K, Shizuya H, Choi S, Chen YJ (2001) Initial sequencing and analysis of the human genome. Nature 409:860-921

Lankenau D-HV, Jean-Nicolas (2009) Transposons and the Dynamic Genome

Lerat E (2010) Identifying repeats and transposable elements in sequenced genomes: how to find your way through the dense forest of programs. Heredity (Edinb) 104:520-533

Lewis ZA, Honda S, Khlafallah TK, Jeffress JK, Freitag M, Mohn F, Schubeler D, Selker EU (2009) Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. Genome Res 19:427-437

Luan DD, Korman MH, Jakubczak JL, Eickbush TH (1993) Reverse transcription of R2Bm RNA is primed by a nick at the chromosomal target site: a mechanism for non-LTR retrotransposition. Cell 72:595-605

Maheshwari R, Bharadwaj G, Bhat MK (2000) Thermophilic fungi: their physiology and enzymes. Microbiol Mol Biol Rev 64:461-488

McClintock B (1948) Mutable loci in maize. Carnegie Institution of Washington Year Book 47: 155-169

Muszewska A, Hoffman-Sommer M, Grynberg M (2011) LTR retrotransposons in fungi. PLoS One 6:e29425

Nene V, Wortman JR, Lawson D, Haas B, Kodira C, Tu ZJ, Loftus B, Xi Z, Megy K, Grabherr M, Ren Q, Zdobnov EM, Lobo NF, Campbell KS, Brown SE, Bonaldo MF, Zhu J, Sinkins SP, Hogenkamp DG, Amedeo P, Arensburger P, Atkinson PW, Bidwell S, Biedler J, Birney E, Bruggner RV, Costas J, Coy MR, Crabtree J, Crawford M, Debruyn B, Decaprio D, Eiglmeier K, Eisenstadt E, El-Dorry H, Gelbart WM, Gomes SL, Hammond M, Hannick LI, Hogan JR, Holmes MH, Jaffe D, Johnston JS, Kennedy RC, Koo H, Kravitz S, Kriventseva EV, Kulp D, Labutti K, Lee E, Li S, Lovin DD, Mao C, Mauceli E, Menck CF, Miller JR, Montgomery P, Mori A, Nascimento AL, Naveira HF, Nusbaum C, O'Leary S, Orvis J, Pertea M, Quesneville H, Reidenbach KR, Rogers YH, Roth CW, Schneider JR, Schatz M, Shumway M, Stanke M, Stinson EO, Tubio JM, Vanzee JP, Verjovski-Almeida S, Werner D, White O, Wyder S, Zeng Q, Zhao Q, Zhao Y, Hill CA, Raikhel AS, Soares MB, Knudson DL, Lee NH, Galagan J, Salzberg SL, Paulsen IT, Dimopoulos G, Collins FH, Birren B, Fraser-Liggett CM, Severson DW (2007) Genome sequence of Aedes aegypti, a major arbovirus vector. Science 316:1718-1723

Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC, Osiewacz HD, Poggeler S, Read ND, Seiler S, Smith KM, Zickler D, Kuck U, Freitag M (2010) De novo assembly of a 40 Mb eukaryotic genome from short sequence reads: Sordaria macrospora, a model organism for fungal morphogenesis. PLoS Genet 6:e1000891

Nielsen ML, Hermansen TD, Aleksenko A (2001) A family of DNA repeats in *Aspergillus nidulans* has assimilated degenerated retrotransposons. Mol Genet Genomics 265:883-887

Polavarapu N, Marino-Ramirez L, Landsman D, McDonald JF, Jordan IK (2008) Evolutionary rates and patterns for human transcription factor binding sites derived from repetitive DNA. BMC Genomics 9:226

Kunze R (1996) The maize transposable element Activator (Ac). Curr Top Microbiol Immunol 204: 161-194

Wessler SR (2006) Eukaryotic Transposable Elements: Teaching Old Genomes New Tricks.

Oxford University Press

Roper M, Ellison C, Taylor JW, Glass NL (2011) Nuclear and genome dynamics in multinucleate ascomycete fungi. Curr Biol 21:R786-793

Rosa AL, Folco HD, Mautino MR (2004) In vivo levels of S-adenosylmethionine modulate C:G to T:A mutations associated with repeat-induced point mutation in *Neurospora crassa*. Mutat Res 548:85-95

Rossignol JL, Faugeron G (1994) Gene inactivation triggered by recognition between DNA repeats. Experientia 50:307-317

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nat Genet 20:43-45

Selker EU (1990) Premeiotic instability of repeated sequences in *Neurospora crassa*. Annu Rev Genet 24:579-613

Selker EU, Cambareri EB, Jensen BC, Haack KR (1987) Rearrangement of duplicated DNA in specialized cells of *Neurospora*. Cell 51:741-752

Selker EU, Garrett PW (1988) DNA sequence duplications trigger gene inactivation in *Neurospora crassa*. Proc Natl Acad Sci U S A 85:6870-6874

Shen L, Wu LC, Sanlioglu S, Chen R, Mendoza AR, Dangel AW, Carroll MC, Zipf WB, Yu CY (1994) Structure and genetics of the partially duplicated gene RP located immediately upstream of the complement C4A and the C4B genes in the HLA class III region. Molecular cloning, exon-intron structure, composite retroposon, and breakpoint of gene duplication. J Biol Chem 269:8466-8476

Shen MR, Batzer MA, Deininger PL (1991) Evolution of the master *Alu* gene(s). J Mol Evol 33:311-320

Singer MJ, Marcotte BA, Selker EU (1995) DNA methylation associated with repeat-induced point mutation in *Neurospora crassa*. Mol Cell Biol 15:5586-5597

van de Lagemaat LN, Landry JR, Mager DL, Medstrand P (2003) Transposable elements in mammals promote regulatory variation and diversification of genes with specialized functions. Trends Genet 19:530-536

Wang H, Xing J, Grover D, Hedges DJ, Han K, Walker JA, Batzer MA (2005) SVA elements: a hominid-specific retroposon family. J Mol Biol 354:994-1007

Wang X, Hsueh YP, Li W, Floyd A, Skalsky R, Heitman J (2010) Sex-induced silencing defends the genome of *Cryptococcus neoformans* via RNAi. Genes Dev 24:2566-2582

Wessler SR (2006) Transposable elements and the evolution of eukaryotic genomes. Proc Natl Acad Sci U S A 103:17600-17601

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982

Wright SI, Agrawal N, Bureau TE (2003) Effects of recombination rate and gene density on transposable element distributions in *Arabidopsis thaliana*. Genome Res 13:1897-1903

Yuan YW, Wessler SR (2011) The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. Proc Natl Acad Sci U S A 108:7884-7889