Abstract

DIENER, STEPHEN ERICSON. Insight into Filamentous Fungal Secretion and Evolution through Genomic Analysis (Under the direction of Ralph Dean)

Filamentous fungi are of broad economic importance due to their roles in industry, medicine and agriculture. There are several filamentous fungi, such as *Trichoderma reesei,* which have been harnessed as protein factories due to their immense secretion capacity. Unfortunately, their full potential cannot be exploited due to an incomplete understanding of the pathways and genes involved in the fungal secretion system. Through the development of bioinformatic tools and the use of genomic technologies including expressed sequence tags and bacterial artificial chromosomes, the genome of *T. reesei* has been partially characterized and a number of genes involved in the secretion system have been identified.

Pathogenic fungi, such as *Magnaporthe grisea,* are of great economic importance due to their devastating effect on agriculture. *M. grisea* is responsible for the loss of incredible amounts of rice crop yearly and has recently had its genome completely sequenced and annotated. The genome sequence has revealed the set of transposable elements in *M. grisea* which have then been analyzed using gene genealogies and the coalescent. The genealogies and coalescent have revealed that all elements analyzed showed a rapid expansion at some point in the past. This can be explained as a genomic event leading to the acceptance of transposable element activity most likely caused by the loss of genomic defense mechanisms. As a pathogen, the ability to evolve quickly in the face of plant defense mechanisms is essential. Transposable element activity can provide means for rapid evolution and this acceptance may represent a shift of these elements from genomic parasitism to mutualism.

INSIGHT INTO FILAMENTOUS FUNGAL
SECRETION AND EVOLUTION THROUGH
GENOMIC ANALYSIS

by

Stephen Ericson Diener

A dissertation submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of

Doctor of Philosophy
FUNCTIONAL GENOMICS

Raleigh
2005

APPROVED BY:

Chair of Advisory Commitee

Dedication

To my parents, grandfather and wife.

To my parents and grandfather for making education an early priority.

And to my wife for her endless support.

Biography

I was born the first of three children to a West Point graduate and an army nurse. After hopping around the states for a few years my parents decided to leave the army and settle in North Carolina. Elementary, middle and high school went by quickly and I was off to the University of North Carolina at Chapel Hill where my plan was to study biochemistry. Four years and three major changes later, I ended up back where I started graduating with my B.S. in Biochemistry. While at UNC I met my future wife and we were married a few years later in 2003. It took less than a year after graduation to realize that I did not want to be a lab tech for the rest of my life. My interest in genomic analysis and the use of computers for biological research was initiated by an undergraduate research project involving on one of the first completed genomes and ultimately led me to join the Genomics Sciences program at North Carolina State University.

Table of Contents

List of Tables

List of Figures

Chapter 1

Review

Introduction

The advances of the genomic age have provided biological scientists with a wide variety of powerful techniques and technologies. The introduction of high-throughput, low-cost sequencing has had the most wide ranging effect on science. This technology has quickly broadened the scope of research from the characterization of a single gene's action to the identification and profiling of complete pathways and genomes. Expressed sequence tag (EST) sequencing, whole genome shotgun sequencing and transcriptional profiling through micro-array analysis all have shown great promise in furthering our understanding of biological processes and evolution. These genomic technologies, spurred on by the human genome initiative, are of great utility to a number of scientific communities including those who study filamentous fungi. The high gene density and relatively small size of filamentous fungal genomes makes them incredibly amenable to genomic analysis.

The Fungal Kingdom is composed of a wide variety of species, including pathogens and saprophytes, which are economically important in both agriculture and industry. Certain saprophytic fungi such as *Trichoderma reesei* and *Aspergillus nidulans* are known to be extremely efficient protein secretors and this characteristic has been exploited for the production of both homologous and heterologous proteins. Unfortunately, the yield of heterologous protein secretion in filamentous fungi is significantly less than that of homologous proteins. In order to overcome this limitation, it is necessary to understand the basic biology behind fungal secretion. The first section of the following review outlines what is known about the filamentous secretion pathway as well as what steps have been taken to overcome its inherent limitations in *T. reesei*.

Among fungal species, many exist which are the causal agents of crop diseases which lead to the loss of agricultural products. An example of which is *Magnaporthe grisea*, the rice blast fungus, which leads to an estimated yearly loss in rice production sufficient to feed sixty million people. The recent sequencing of this fungi's genome has provided the community with an enormous resource in the fight to control the devastating disease caused by this fungus and understand the evolutionary processes which lead to and maintain pathogenicity. Found within the genome is a wide variety of transposable elements whose effect on the evolution and pathogenicity of *M. grisea* is noted within the literature. Through the use of population genetics techniques, the history, migration, and evolutionary effects of transposable elements can be better understood. Within the second section of the following review, I will briefly cover what is know about fungal transposable elements with an emphasis on the *M. grisea* transposons and their known effects on genomic structure and regulation.

Fungal Secretion

In recent years filamentous fungi have become increasingly important economically; expanding from their centuries old uses in the manufacture of cheese and fermented products for food and beverages into a multi-billion dollar industry producing enzymes, amino acids, polymers and other compounds for use in numerous industries. Filamentous fungi are of high utility to industrial concerns because they provide both a factory for the production of these compounds as well as the raw genetic material for a number of industrially important enzymes. Fungal-derived enzymes are used in such diverse industries as pulp and paper, textiles, detergent manufacture, and alternative fuel [1]. *Trichoderma reesei* is a fungus of particular interest due to its large complement of biomass degrading genes and its ability to secrete them in large quantities. The exploitation of these two characteristic has made *T. reesei* an important industrial fungus.

The explosion of filamentous fungi for industrial applications was ignited by their ability to produce large quantities of specific proteins. Commercial fermentation can result in yields ranging up to 40g/l of the targeted protein [2]. The initial exploitation of these fungi involved identification of strains naturally overproducing an enzyme of interest and improvement of yield through mutagenesis techniques [3]. In cases where a strain overexpressing the target of interest was not available, it became necessary to heterologously express these proteins in other fungi through the use of various gene transfer techniques. The primary issue limiting the expansion of protein production in filamentous fungi is the significant decrease in yield of a heterologous protein when compared to the yield of a homologous protein produced by the same fungus. While it is not uncommon to obtain a yield in the tens of grams per liter for an expressed homologous protein, heterologous yield is usually limited to the single digits or less in grams per liter [4]. In order to overcome this deficiency in heterologous expression, it has become clear that a more complete understanding of the filamentous fungi's secretion machinery is required. The past several decades have provided us with great insights into this field but the knowledge required to completely leverage these fungi's expression potential is still lacking. In order to identify all areas of the pathway which can be altered to increase protein production, thorough understanding of the system is needed. Recent exploration of *T. reesei's* genome has provided a wide variety of insights into the functioning of the filamentous fungus's secretion system as well as the identification of novel biomass degrading enzyme which can be exploited. This recently acquired knowledge combined with the upcoming release of *T. reesei's* annotated genome promise to further our understanding of this fungus's biology and allow us to maximize its potential. The following section provides a brief overview of the filamentous secretion system, which will be looked at in more depth in subsequent sections. The subsequent sections will deal with recent knowledge of the various steps in the secretion process, from induction through glycoslation, with a focus on the state of *T. reesei* as an industrial fungus.

The Secretion System

The majority of our knowledge of filamentous fungi's secretion system comes from the application of models developed in yeast and other higher eukaryotes. The consensus is that the secretion pathway in filamentous fungi is quite similar to that of yeast with few differences. The primary differences exist in scale of protein yield, ultrastructural characteristics of certain organelles and secretion targeting due to the polarized hyphal growth of filamentous fungi. While the possibility of hyphal secretion in filamentous fungi was debated in the past, recent studies tracking protein movement and secretion with GFP fusion proteins has shown that secretion occurs at the hyphal tips [5, 6].

Most secreted proteins follow a distinct path as they are shuttled out of the cell. Passing through the endoplasmic reticulum (ER) where folding and post-translational modifications occur, such as disulfide bond creation, phosphorylation, and glycosylation. The proteins are then loaded into transport vesicles and targeted to the Golgi compartment where additional glycosylation and other post-translational modifications take place. After being packaged yet again the proteins are targeted to their final destination, whether that is the extracellular matrix to be secreted or a cellular compartment to become part of the cellular machinery. For the purposes of this review we will discuss what is known about the filamentous secretion pathway from regulation to the targeting of a nascent protein through the ER and Golgi with a focus on application of this knowledge to overcome production bottlenecks in *T. reesei*.

Regulation of Secretion

Understanding the regulation of the secretion pathway is of great interest to those interested in optimizing the fungal system for secretion of specific proteins. In order to adapt to their changing surroundings fungi have evolved highly inducible promoters to regulate the types of enzymes secreted. Of the proteins secreted by *T. reesei* into the extracellular matrix under cellulase inducing conditions, the CBHI, CBHII, EGI and EGII make up fifty percent [7]. The ability of this fungus to redirect a large percentage of its secretion apparatus to the creation of four enzymes shows its potential as an industrial fungus. The regulation of cellulases in *T. reesei* is a well explored topic due to its potential for expansion of the fungal secretion system.

Plant biomass is primarily composed of cellulose, a polymer of BETA-1-4 linked glucose molecules organized into fibrillar structures. A secondary component of biomass is hemicellulose which is a polysaccharide composed of two or more types of monosaccharides such as D-glucose, D-mannose and D-xylose [8]. In order to liberate the large quantity of energy stored in these polysaccharides, *T. reesei* secretes a variety of cellulases and hemi-cellulases. Its cellulase activity is produced by a number of endoglucanases, responsible for the digestion of branched polysaccharides into linear polysaccharides, and exoglucanases, responsible for degradation of linear polysaccharides into disaccharides. The endoglucanases (EGI/Cel7b, EGII/Cel5a, EGIII/Cel12a, EGIV/Cel61a, EGV/Cel45a) work in turn with the exoglucanases (CBHI/Cel7a, CBHII/Cel6a) to digest cellulose into cellobiose which is thought to be further degraded into glucose by BETA-glucosidases (bgl1/cel3a, bgl/cel1a) [9-16]. Recent work sequencing EST libraries from this fungus has identified a number of additional enzymes involved in this process including three putative endoglucanases (cel74a, cel61b and cel5b) and five putative exoglucanases (cel3b, cel3c, cel1b, cel3d, and cel3e) [17]. This fungus also secretes hemicellulose-degrading enzymes responsible for the digestion of the xylan (Xyn1, Xyn2, Xyn3 and Xyn4) and mannan (Man1) side chains as well as side chains containing acetyl (Axe1), methyglucuronic acid (Glr1), and arabinose (Abf1) moieties. BETA-xylosidase (Bxl1) and ALPHA- galactosidases (Agl1, Agl2 and Agl3) are responsible for digestion of oligosaccharides derived from hemicellulose [18-27]. Foreman et al. also identified a second acetyl xylan esterase (Axe2), a second arabinofuranosidase (Abf2) as well as two other enzymes containing carbohydrate binding domains [17]. A bacterial artificial chromosome (BAC) library was constructed from the genome of *T. reesei* which was then end-sequenced and assembled into a physical map using FPC. A number of sequences annotated with biomass degrading functions were identified in the assembly of BAC ends which were not present in comprehensive EST datasets targeted at identifying these enzymes. These consisted of sequences annotated as BETA-mannosidase, BETA-galactosidase and xylanase regulator 1. Identification of these genes which were lacking in targeted comprehensive EST datasets shows the utility of random survey sequencing. In the same study, hybridization analysis of 34 probes consisting of *T. reesei* sequence annotated as having biomass degrading function to the BACs revealed that some of these genes are clustered within the genome [28]. This clustering of genes of related function could be a remnant of the genes' evolution or provide for co-regulation.

A number of inducers have been identified which lead to the secretion of cellulase enzymes. Oligosaccharides, such as cellobiose, sophorose (BETA-1-2 linked glucose disaccharide), and gentiobiose (BETA 1-6 linked glucose disaccharide) all are shown to induce expression of cellulase genes [12, 29-31]. Of these, sophorose, not the intermediate product of cellulose degradation, cellobiose, is the most potent. Sophorose is not a component of cellulose and, in fact, is a relatively rare disaccharide in nature. Conversion of cellobiose to sophorose is attributed to the transglucosylation activity of BETA-glucosidase. An inhibitor of BETA-glucosidase was shown to inhibit cellulase induction by oligosaccharides but did not affecting induction by sophorose [32-34]. This showed the importance of Bgl in cellulase induction. However, when the BglI gene of *T. reesei* was disrupted there was still beta-glucosidase activity suggesting that the Bgl1 gene is not solely responsible for induction of cellulase expression. A membrane bound constitutive beta-glucosidase initially modifies the inducer which then leads to expression of Bgl1 and rapid induction of the cellulases [35]. Constitutive expression of cellulases allows the fungus to recognize cellulose which is degraded into cellobiose and transglycosylated by membrane bound BETA-glucanase into sophorose which leads to the induction of cellulase expression. Sophorose is assumed to be transported into the fungus where it initiates a signaling pathway leading to the expression of the cellulases.

Analysis of the expression of a number of these genes has elucidated transcription factors and promoter elements involved in the regulation of this pathway. It is clear from the identification of Cre1 binding sites in the promoter of these genes, that the catabolyte repression plays a role in regulating the expression of cellulases based on identification of a source of glucose or other preferred carbon source [36, 37]. Another factor important in the expression of cellulases is the CAE (cbh2 activating element) which engages the cbh2 promoter, shifting its nucleosome positioning which exposes the TATA box allowing for transcription [38]. The transcription factors AceI and AceII are involved in the global regulation of cellulase genes. AceI has been shown to be a repressor of cellulase and xylanase genes in *T. reesei* while AceII led to lower levels of induction of cbhI, cbhII, eglI and eglII on cellulose but no change on sophorose [39, 40].

Analyses of the expression patterns of cellulases and hemi-cellulases under cellulase-inducing conditions have shown coordinate expression of subsets of these genes [41]. Foreman et al. also showed coordinate expression in a microarray analysis of known cellulases, hemi-cellulases and the novel biomass degrading genes identified in the study. While the majority of BETA-glucosidases followed expression patterns consistent with cellulase expression, three of the putative BETA-glucosidases (cel1b, cel3d, and cel3e) were more highly induced by glucose than lactose which may indicate that these particular enzymes are less involved in cellulose degradation than other genes of this class. The other newly identified genes were coordinately expressed with other members of the class as annotated. Interestingly, the genes identified with cellulose binding domain but no other functional annotation, cip1 and cip2, both had very similar expression patterns to cbh1 which supports that these genes have potential roles in biomass degradation. Also noted was that very few of the genes annotated as chaperones and foldases were up-regulated by the induction of cellulase secretion by sophorose or lactose. This contrasts a number of studies which show an increase in the expression of these genes due to overexpression of foreign proteins [42, 43]. However, the discrepancy may be expected since over expressing a modified protein is more likely to induce the unfolded protein response, reviewed in a later section, leading to an increase in those correlated enzymes while naturally inducing a native protein is less likely to induce such a response. Understanding how fungi initiate and regulate protein translation is essential to maximizing the utility of this fungus. The ability to control the expression of secreted proteins through the regulation of growth conditions is used throughout industrial applications [44]. The successfully transcribed target protein's mRNA is subsequently captured by a ribosome where the nascent protein begins it journey through the secretion pathway.

Targeting and Vesicular Trafficking

Protein translocation and targeting throughout the secretion pathway is compulsory to the creation of properly folded and modified proteins as well as ensuring that the protein arrives at the appropriate final destination. Initial targeting of the nascent protein to the ER in *S. cerevisiae* has been described as occurring via two distinct pathways. In the first, the signal recognition particle (SRP)-dependent pathway, targeting to and traversal of the ER membrane occurs co-translationally. In this pathway the signal peptide of the nascent protein is identified by the SRP while the protein is still being translated and the complete ribosomal complex is translocated to the ER membrane where the nascent protein is released into the ER lumen. In the SRP-independent pathway, the targeting to the ER occurs post-translationally. A signal peptide on the nascent protein is recognized by a chaperone (Hsp70) which carries the protein to the ER membrane and facilitates its entry into the ER. The hydrophobicity of the signal peptide determines which of these pathways a certain protein takes into the ER. Proteins which contain less hydrophobic signal sequences are preferentially routed to the ER using the SRP-dependent pathway while those proteins with a more hydrophobic signal peptide can be targeted to the ER using either pathway. Both of these pathways use the Sec61p trans-membrane protein channel as their proteins entrance to the ER. The SRP-dependent pathway requires a membrane bound SRP-recognition protein to bind to the SRP and bring the ER-targeted protein to the Sec61p channel. In the SRP-independent pathway, the chaperone brings the ER-targeted protein to the Sec62p-Sec72p-Sec73p complex which interacts with the membrane bound Sec63p to bring the nascent protein to the Sec61p channel. After being targeted to the Sec61p channel, the lumen localized chaperone BiP, interacting with Sec63p, assists in the translocation of the nascent protein through the channel [45]. A homolog of the *S. cerevisiae* signal recognition partical, SRP54, has been identified in *Aspergillus niger* [46]. Homologs to the BiP protein of *S. cerevisiae* has been identified in a number of Aspergilli as well as *Neurospora crassa. T*he *A. awamori* homologous gene supported its function by complementing a yeast BiP mutant [47-49]. In a comprehensive set of expressed sequence tags, generated by pooling mRNA from *T. reesei* under a number of secretion inducing growth conditions, homologues to both BiP and Sec61p were identified by Diener et al. [50]. While homologues from filamentous fungi of each protein involved in these pathways have not been isolated, identification of certain key homologs suggests that both ER-targeting pathways exist in filamentous fungi.

After successfully being folded and modified in the ER, properly folded proteins are packaged into vesicles and targeted to the golgi or the golgi-like structure found in filamentous fungi. In *S. cerevisae*, the proteins responsible for the creation and routing of these vesicles between organelles have been elucidated in a number of studies. The majority of these responsible proteins are small GTP-binding members of the Ras superfamily. Within this superfamily are two subfamilies; the ARF/SAR and the SEC4/YPT/RAB groupings. The former is involved in the formation of vesicles from the donor organelle and preparation of the vesicle fusion with the acceptor organelle. The latter group is involved in targeting of the vesicle as well as fusion with the acceptor organelle. A number of these have been identified in yeast and associated with many stages of post-ER protein trafficking [51-53]. Homologs of Sar1, the yeast GTPase involved in ER to Golgi transport, have been identified in both *A. niger* and *T. reesei* and subsequent studies have shown similar function [54]. Punt et al. have identified homologs in *A. niger* to 6 other GTPases in this family involved in all stages of protein trafficking [55] A homologue of YPT7 was recently cloned from *A. nidulans* and shown to be involved in vacuolar biogenesis [56]. Further functional description of genes involved in vesicular trafficking is shown in Table 1. *T. reesei* homologues involved in all stages of vesicle transport were identified through characterization of an EST library as well as the random survey technique of BACend sequencing [28, 50].

ER Protein Maturation and Quality Control

The protein maturation process begins once the nascent protein enters the ER. Within the ER, a number of enzymes, such as chaperones and foldases begin work on the nascent protein. BiP, in addition to assisting in translocation of the protein across the ER membrane, participates in folding of the protein, subunit assembly, and degradation of improperly folded proteins [57]. Protein disulfide isomerase (PDI) is a member of the thioredoxin family which is responsible for catalyzing the reduction, oxidation and isomerization of disulfide bonds [58]. This gene of obvious importance to protein maturation has been cloned in a number of filamentous fungi including *A. niger, A. oryzae,* and *T. reesei* [59-62]. Calnexin is another molecular chaperone responsible, in conjunction with its soluble counterpart, calreticulin, for the proper folding, oligomeric assembly and quality control of glycoproteins in the ER. The calnexin pathway is well defined in the mammalian cell where it is responsible for a folding checkpoint. Calnexin ensures that proper folding has occurred after N-glycans have been added to the protein by a glucosyltransferase and partially degraded by glucosidase I and glucosidase II. It determines whether a glycoprotein should continue on in the pathway or be returned to the glucosyltransferase for an additional attempt at proper folding [63]. The gene for calnexin was first found in *S. cerevisiae* and has been cloned in *A. niger* and identified in the sequence of other filamentous fungi including the EST dataset of *T. reesei* [50, 64].

Peptidyl prolyl isomerase (PPI) is responsible for the catalysis of peptide bonds on the N-terminal side of proline residues and has been shown to increase the rate of protein folding in vitro. These proteins are classified into one of two major families based on their ability to bind to either cyclosporin A or FK506. The families are designated the cyclophilins and FK binding proteins (FKBP) respectively. A number of members from each of these families have been identified in yeast [65]. Filamentous fungi also have been shown to contain genes from both classes. *N. crassa* was shown to have an FKBP and both *A. nidulans* and *A. niger* were shown to have a cyclophilin named cypB [66-68]. *T. reesei* also contains three putative PPI genes (homologues to cypB, cypC and FKBP13) annotated from an EST dataset [50].

The ER is also responsible for quality control; it must ensure that the proteins passing through to the next step are properly folded. There are two primary pathways which ensure that the proteins leaving the ER are properly folded. The first is unfolded protein response pathway (UPR) which is responsible for detection of unfolded proteins and induction of folding enzymes.

The second is the ER-associated degradation (ERAD) which detects improperly folded proteins and targets them for degradation.

The UPR pathway is composed of three primary genes in *S. cerevisiae*; IRE1, RGL1 and HAC1. The IRE1 protein recognizes that there are unfolded or improperly folded proteins accumulating in the cell and, in conjunction with the RLG1 protein, induces the expression of HAC1. HAC1 is a transcription factor which binds to the UPR-expression element in the promoters of genes induced by this pathway, such as foldases and chaperones. PTC2 is a protein phosphatase which negatively regulates the UPR by binding to and dephosphorylating IRE1 [69, 70]. The *T. reesei* homologs of IRE1 and PTC2 were recently identified and shown to complement their respective *S. cerevisiae* mutant [71] and a homologue to HAC1 was identified in a *T. reesei* EST dataset [50]. IRE1 was shown to affect the UPR response when overexpressed as expected leading to increased BiP and PDI levels [71].

The ERAD system identifies misfolded proteins and leads to their degradation in the cytosol. The improperly folded proteins are identified in the ER, most likely by resident chaperones, and translocated to the Sec61p channel where it is returned to the cytosol and targeted to 26S proteasome by ubiquitin-conjugating enzymes [72]. The only evidence for existence of this pathway in filamentous fungi is the identification of prs12 of *T. reesei* which is a homolog to a subunit of the 26S proteasome and other 26S subunit homologs in *N. crassa* and *A. nidulans* identified in genomic sequence [73].

Protein Modifications in the Golgi

While a structure matching the classical morphological description of a Golgi body does not exist in all filamentous fungi, a golgi-like organelle which seems to have the same functionality as a Golgi body is present throughout. The presence of this body in filamentous fungi was shown in ultrastructural studies of *A. niger* and *T. reesei* [74]. Filamentous fungi differ from yeast by possessing the glycosylation machinery necessary to create mammalian-like high-mannose N- or O-linked glycosylated proteins [75, 76]. This characteristic makes the filamentous fungus a candidate for hosting mammalian proteins of biomedical interest. Filamentous fungi apparently lack high complexity glycans which frequently occur in mammalian glycoproteins [77, 78]. There is currently ongoing work addressing these deficiencies through pathway modification which will be addressed later in this review. While the initial addition of carbohydrates to a protein can occur in the ER, the majority of glycosylation occurs within the Golgi apparatus. Few of the enzymes involved in these additions are known in fungi. A complement of glycosyltransferases and glycosidases is likely responsible for the building of these of these glycan structures. Dolichol-dependent protein mannosylation is the most well characterized glycosylation pathway in the filamentous fungi. This pathway is described in *T. reesei* whereby O-linked mannosyl groups are synthesized via a dolicholphosphate dependent pathway. The three major enzymes of this pathway, dolichol kinase, mannosylphosphodolichol synthase and GTP:ALPHA-D-mannose-1-phophate guanyl transferase, work in concert to transfer mannose to a protein via a dolicholphosphate intermediate [78]. The proteins responsible for this pathway as well as number of genes with homology to other fungal oligosaccharyl transferases and glycosyltransferases were identified in a *T. reesei* EST dataset. These putative transferases represent possible targets for mutation or overexpression in the hopes of increasing the range of proteins produced by this fungus [50].

Harnessing and Improving the Secretion Pathway

In order to harness the secretion pathway of filamentous fungi for homologous and heterologous protein production, the gene of interest must first be cloned into the host in a way that ensures protein expression at high levels and shuttling into the secretion pathway. The primary methods used to increase mRNA levels in homologous protein production are to increase the gene copy in the strain and introduce a resident strong promoter, such as that of CBHI in *T. reesei*, to the genes of interest. The signal peptides of known secreted proteins have been fused to the recombinant protein in order to target it into the secretion pathway efficiently [79]. Construction of a protease deficient strain and optimization of the growth medium are also techniques which efficiently increase yield of homologous proteins. Fusion of the target gene to a homologous protein known to be efficiently secreted improves the yield of heterologous protein secretion and is the most successful modification aimed at increasing heterologous protein production to date [80, 81]. While the complete biological basis behind this technique is not understood, a recent study showed that having an intact carrier domain increased production of bacterial xylanase in *T. reesei* from 150-300mg/l to 820mg/l compared to the signal sequence alone or an incomplete domain. This suggests that a heterologous protein chimerized with a protein known to fold properly in the system allows a non-native protein to avoid the quality control mechanisms present in the ER so the protein can continue along the secretion pathway [82].

With the steps necessary to harness this protein secretion pathway for homologous protein production defined, work targeted at improving the filamentous system's deficiencies is of great interest to the growth of this system as a protein factory. The discrepancy in yield between a homologous protein and a heterologous protein produced in the same system is an interesting problem. Attempts to increase the yield of heterologous proteins in the filamentous system target at modification of specific steps in the secretory process in an attempt to identify and remove the secretory bottleneck.

Knowledge of the filamentous fungus's hyphal tip secretory preference spearheaded interest in studies of hyperbranching mutants' secretory potential. While these mutants show physical properties which are advantageous during fermentation, such as low viscosity, a correlation between secretion efficiency and tip density was not found [83]. Initial work postulated that low mRNA transcript levels could be the cause of poor yield and methods of increasing transcript levels were developed. While the transcript levels were adequately enhanced, a correlated increase in secretion of the protein of interest was not seen.

Since the bottleneck to high levels of heterologous protein production does not lay in transcriptional efficiency, the majority of work has centered on ensuring that the proper post-translational modifications of the heterologous protein are able to occur. The use of synthetic leader sequences with additional N-glycosylation sites and BiP binding sites was shown to prolong the retention of the heterologous protein in the ER and allow for correct folding in *S. cerevisiae* [84]. The technique of adding N-glycosylation sites was used in a number of filamentous fungi with success. Further analysis of this technique comparing the addition of glycosylation sites on both the C- and N- terminus showed that N-terminal glycosylation leads to retention in the ER and subsequently increased secretion while C-terminal glycosylation did not [48]. Modification of the levels of molecular chaperones and proteins involved in glycosylation pathways is a route being actively investigated. Overexpression of native enzymes involved in the protein maturation process is a popular tactic used to isolate heterologous protein bottlenecks. The ER associated chaperone BiP was overexpressed in *A. niger* and yeast. In yeast, this resulted in increased secretion of Chymosin and hirudin while not affecting other proteins tested [85, 86]. Overexpression of BiP actually reduced the secretion of glucose oxidase in the *A. niger* [87]. PDI over-expression was tested in *A. niger* resulting in increased secretion of the target protein, bovine prochymosin [59]. Over-expression of IRE1 in *T. reesei* did not increase the efficiency of secretion but did upregulate the UPR response and lead to increased PDI and BiP expression in the fungus [71]. It has recently been shown that constitutive induction of the UPR in *A. niger* improves foreign protein production, specifically of *Trametes versicolor* laccase [88]. Calnexin was over-expressed in a strain of *A. niger* leading to an increase in secretion of the heterologous protein manganese peroxidase [89]. While up-regulation of individual enzymes can lead to an increase in the production of heterologous protein, the effect seems to be dependent on the qualities of the gene being over-expressed.

With the majority of secreted proteins being glycosylated, the possibility that the glycosylation capacity of the host is the limitation in heterologous secretion has been explored. However, recent studies suggest that the major limiting factor in heterologous protein production lays in the secretion pathway [5]. The overexpression of yeast mannosylphosphodolichol synthase, a major component in *T. reesei's* glycosylation pathway, led to increased secretion of CBHI with comparable carbohydrate content and no hyperglycosylation [90]. This line of research suggests that the glycosylation pathway's capacity may limit the overexpression of secreted proteins and upregulation of proteins involved in glycosylation is effective in removing the bottleneck. While identification and overexpression of genes involved in glycosylation is an area of promise for strain improvement, knowledge of genes involved in this pathway limits progress.

The filamentous secretion system is a natural choice to expand into the rapidly growing field of therapeutic glycoproteins for human use. As they are natural secretors of numerous glycoproteins with GRAS (generally regarded as safe) status for use in the food industry, the development of this new field seems a natural step. While a number of fungal glycans resemble the mammalian high-mannose glycans, typical mammalian-like complex glycans have not been found [77]. However, in a study of an IgG antibody produced in *A. niger*, pharmacokinetic behavior and activity were found to be similar to the mammalian wild type regardless of incomplete glycosylation which calls into question the necessity of producing an exact mammalian glycoprotein in every case [91]. While this is promising, there are worries that improper glycosylation can lead to an immunogenic response with long term exposure.

In efforts to overcome the lack of mammalian glycans, several groups have worked to modify the glycosylation pathway both in vitro and in vivo. The identification that secreted fungal proteins contained glycans which resemble mammalian glycan intermediates led to the post secretion modification of those glycans to closer resemble mammalian proteins [92]. While this post-secretion processing produced glycoproteins similar to the mammalian original, it required multiple reactions and the final product was produced at a low yield.

A more promising route to fungal secretion of mammalian glycans seems to be one of mutation of the fungus to remove native activity leading to incorrect glycosylation and transformation to include mammalian glycosyltransferase activity. Acetylglucosaminyltransferase I (GlcNAc-TI), which is absent from fungal genomes and necessary for the conversion of oligomannose to the complex N-glycans in higher eukaryotes, from rabbit was transformed into *A. nidulans* and active enzyme was recovered [93]. Human GlcNAc-TI was transformed into *T. reesei* where NMR analysis showed that the expected glycan (GlcNAc-Man(5)-GlcNAc(2)) was formed on recovered CBHI [94]. While production of human glycoproteins in yeast has been actively pursued [95-97], little work on the humanization of filamentous glycoslyation pathways has been published recently.

The secretion system of filamentous fungi is an extremely powerful tool which is posed to become useful to even more industries as its limitations in producing mammalian proteins are removed. Work conducted of the last several years have added an incredible amount to what we understand about the fungal secretion pathway. Recent genomic analyses have identified members of the *T. reesei's* secretion pathway and provided new genes with potential roles involved in secretion and glycosylation as targets for future studies. With the upcoming release of the annotated *T. reesei* genome, even more genes involved in these industrially important pathways will become available. There is great potential to increase heterologous yield and improve this fungus's ability to create mammalian-like glycans. The use of modern genomic techniques to accelerate strain improvement will speed the realization of this potential.

Transposable Elements of *M. grisea*

Mobile pieces of genetic material, known as transposable elements, inhabit the genomes of both eukaryote and prokaryotes. These elements were first predicted by Barbara McClintock in the late 1940s and have been known to exist in plants, bacteria, animals and yeast since the early 1980's [98]. The past two decades have seen an incredible increase in the number of transposable elements known in filamentous fungi [99]. As fungi have increased in economic importance in both industry and agriculture, research has increased into their biology and has lead to the discovery of many elements. The advent of economical genomic sequencing has simplified the identification of transposable elements and allowed for inferences on the evolutionary interplay between these elements and the genome.

Fungal transposable elements are divided into two major groups based on their mode of locomotion and further subdivided within those two groups based on their structure and the presence of certain key domains. Class I TEs transpose using a "copy and paste" method which requires reverse transcriptase to act upon an RNA intermediate to initiate transposition. Class II TEs contain terminal inverted repeats and transpose using a "cut and paste" method where the DNA of the transposon is directly copied to a new location. Class I transposons are composed of three types; LTR transposons which contain long terminal repeats and encode reverse transcriptase, LINE-like retrotransposons lack LTRs but do contain RT and have polyA tails, and SINE-like elements which do not contain RT but appear to be transcribed by RNA polymerase III [100, 101].

A number of transposable elements are known in *Magnaporthe grisea*, the filamentous fungus which causes rice blast disease. Classical genetic techniques have identified over ten transposable elements present in the *Magnaporthe* genus. MAGGY, the most well known *M. grisea* transposable element, is a Gypsy-like element which was first identified by Farman et al and found in high copy number in all rice pathogenic isolates while it was absent or in low copy number in isolates collected from other grasses. The low copy number in non-rice infecting strains is thought to be a result of later horizontal transfer of the element [102]. Several of *M. grisea's* transposable elements have been thoroughly characterized as to their relative activity and host range specificity. Based on the distribution of transposable elements in a number of collected isolates of *Magnaporthe,* the LTR transposons MAGGY and Grasshopper were of limited distribution and determined to be more recently acquired elements while Pot2, Pot3, MG-Sine and MGR583 were widely distributed and thought to be more anciently introduced to the genus [103]. The relative age of introduction of Pot2 and MAGGY was supported by the analysis of chromosome seven sequence which found multiple insertions of MAGGY into Pot2 but no insertions of Pot2 into MAGGY. This suggests that Pot2 predates the introduction of MAGGY or that MAGGY may have an insertion site preference for Pot2 [104].

The recent sequencing and annotation of its 40 Mb genome has provided an accurate inventory of *M. grisea's* transposable elements. Inspection of the genome has shown that a significant percentage, 9.7, is composed of repetitive elements. Representatives from the major classes of transposons were identified within the repetitive elements. In addition to providing a count of known transposable elements within the genome, a number of new transposons were identified in the genome sequence. Pot 4, Pyret-2, Pyret-3, RETRO6, RETRO7, SR2 and SR3 are all novel transposable elements identified from the annotated genome. The inverted repeat DNA transposons consisting of Pot2/3/4 and Occan make up approximately 2% of the genomic sequence. The class of LTR retrotransposons consisting of MAGGY, MGLR-3, Pyret1/2/3 and RETRO5/6/7 is responsible for 3.8% of the genomic DNA. Roughly 1.6% of the genomic sequence is made up by elements in the non-LTR transposon class (LINEs and SINEs), represented by MGL, Mg-SineA and REPBUF [105]

The role that these transposable elements play in evolution of *M. grisea's* genome has not been thoroughly explored. The evolutionary implications of high transposable element activity in a genome are great. Mutational effects of transposable elements are not limited to solely gene or promoter disruptions; accumulation of these elements can lead to drastic genomic changes such as chromosome length polymorphisms. There is evidence for TE's providing homologous regions for intra-genomic recombination of large sections of chromosomes [99]. Within *M. grisea,* clusters of transposable elements have been linked to areas of high recombination and are also implicated in gene duplication [Thon, personal communication]. A number of *M. grisea's* host specificity genes are positioned in areas of high transposon density which increases the likelihood that these genes will be lost or modified in a transposon mediated recombination event [106, 107].

In *M. grisea*, a transposable element, Pot3, has been reported to modify an avirulence gene's transcriptional profile through insertion into the promoter region. The change in expression induced virulence on a previously resistant strain of rice [108]. The fact that rice pathogenic strains primarily propagate asexually, bypassing meiotic quality control mechanisms, increases the chances that such radical mutations become fixed in the population as long as they do not negatively affect vegetative fitness. The bypass of meiosis allows evolution to adapt the fungus to changing environments or host responses much more quickly than if it was bound by the limitations of the quality control mechanisms triggered during sexual reproduction.

The effect that transposable elements have on the genome they occupy has historically been assumed to be detrimental. However, recently it has been suggested that the symbiotic relationship between the genome and transposable elements may be closer to mutualism than parasitism [109]. Evolution is driven by mutation, and transposable elements quickly and efficiently provide that to the genome. In ecologies where the genomic host is under constant selective pressure, the mutational potential that active transposable elements provide can be very advantageous. The evolutionary history of a host-parasite interaction, such as between *M. grisea* and its host, rice, contains numerous instances where the host evolved some type of resistance and the pathogen subsequently overcame that resistance. The selective pressures in this situation are enormous; essentially, both host and pathogen must evolve or become extinct. In this type of highly stressful situation, any trait which increases the ability of a species to rapidly evolve would be tolerated or even selected for.

It can be argued that these repetitive elements play a large role in the continued pathogenicity and rapid adaptability of this fungus. An interesting question is whether these elements are, in part, responsible for the genesis of *M. grisea's* pathogenicity. In order to understand of the effects of transposable elements on the genome of *M. grisea*, it is necessary to decipher the history of these elements within the genome. With the recent sequencing of the genome *M. grisea's* comprehensive set of transposable elements is now available for genealogical analysis. By deciphering the history of these elements, we can begin to understand how the interplay between transposon and genome affects the evolution of this fungus.

Conclusion

Through the use of genomic techniques and technologies this body of work endeavors to expand our knowledge of filamentous fungal genetics. In order to facilitate sequence annotation, we develop a program which streamlines BLAST analysis and result storage and use that program to probe the gene complement of *Trichoderma reesei*. Through creation and functional analysis of expressed sequence tags (EST) and bacterial artificial chromosome (BAC) end sequences we seek an understanding of its secretion system. We also explore the phylogenetic and coalescent relationship of a number of transposable elements in the *Magnaporthe grisea* genome in an effort to understand their effect on the evolution of this fungus.

Table 1.1 – Genes involved in vesicular trafficking

S. cerevisiae Homologue	A. niger homologue	Transport Step	T. reesei Sequence	
Sar1	SarA	ER to Golgi	EstContig175	*
RAB2	SrgD	ER to Golgi	n/p	
Ypt1	SrgB	ER to Golgi	EstContig3974	*
Ypt31	SrgE	Intra Golgi	EstContig280	*
Ypt32		Intra Golgi	EstContig700	*
Ypt6	SrgC	Golgi to vacuole	EstContig3442	*
Ypt7	SrgF	Late endosome to vacuole	n/p	
Ypt 52	n/a	Plasma membrane to endosome/ early to late endosome	EstContig1672	*
Sec4	SrgA	Golgi to plasma membrane	n/p	
Gdi1	n/a	Sec4 GDP dissociation inhibitor	EstContig1224	*
Vps1	VpsA	Vacuolar Biogenesis	EstContig768/3709	*
Sec7p	n/a	Coat protein assembly	BACContig1199	+
Sec8p	n/a	Plasma membrane exocytosis	Trib007E08	+
Sec31p	n/a	COPII coat from ER	BACContig222	+

n/a – no homologues identified
n/p – not present in EST/BAC dataset
* Found in EST dataset; [50]
+ Found in BACends; [28]

References

[1] Stroh, W.H. (1998) Industrial enzymes market - Growth experienced from new products and movement into Global Market. Genetic Engineering News 18, 11

[2] Durand, H., Clanet, M. and Tiraby, G. (1988) Genetic-Improvement of Trichoderma-Reesei for Large-Scale Cellulase Production. Enzyme and Microbial Technology 10, 341-346.

[3] Finkelstein, D., Rambosek, J., Crawford, M., Soliday, C., McAda, P. and Leach, J. (1989) In: Genetics and Molecular Biology of Industrial Organisms Eds.), pp. 295-300. American Society for Microbiology, Washington DC.

[4] Verdoes, J.C., Punt, P.J. and Vandenhondel, C. (1995) Molecular-Genetic Strain Improvement for the Overproduction of Fungal Proteins by Filamentous Fungi. Applied Microbiology and Biotechnology 43, 195-205.

[5] Gordon, C.L., Khalaj, V., Ram, A.F.J., Archer, D.B., Brookman, J.L., Trinci, A.P.J., Jeenes, D.J., Doonan, J.H., Wells, B., Punt, P.J., van den Hondel, C. and Robson, G.D. (2000) Glucoamylase :: green fluorescent protein fusions to monitor protein secretion in Aspergillus niger. Microbiology-Sgm 146, 415-426.

[6] Khalaj, V., Bookman, J.L. and Robson, G.D. (2001) A study of the protein secretory pathway of Aspergillus niger using a glucoamylase-GFP fusion protein. Fungal Genetics and Biology 32, 55-65.

[7] Durand, H., Baron, M., Calmels, T. and Tiraby, G. (1998) FEMS Symp 43, 135-152.

[8] Beily, P. and Tenkanen, M. (1998) Enzymology of hemicellulose degradation. In: Trichoderma and Gliocladium; Enzymes Biological Control and Commercial Applications Ed.^Eds.), Vol. 2, p.^pp. 25-47. Taylor and Francis, London.
[9] Takashima, S., Nakamura, A., Hidaka, M., Masaki, H. and Uozumi, T. (1999) Molecular cloning and expression of the novel fungal beta-glucosidase genes from Humicola grisea and Trichoderma reesei. Journal of Biochemistry 125, 728-736.
[10] Barnett, C.C., Berka, R.M. and Fowler, T. (1991) Cloning and Amplification of the Gene Encoding an Extracellular Beta-Glucosidase from Trichoderma-Reesei - Evidence for Improved Rates of Saccharification of Cellulosic Substrates. Bio-Technology 9, 562-567.
[11] Saloheimo, A., Henrissat, B., Hoffren, A.M., Teleman, O. and Penttila, M. (1994) A Novel, Small Endoglucanase Gene, Egl5, from Trichoderma-Reesei Isolated by Expression in Yeast. Molecular Microbiology 13, 219-228.
[12] Saloheimo, M., NakariSetala, T., Tenkanen, M. and Penttila, M. (1997) cDNA cloning of a Trichoderma reesei cellulase and demonstration off endoglucanase activity expression in yeast. European Journal of Biochemistry 249, 584-591.
[13] Teeri, T.T., Lehtovaara, P., Kauppinen, S., Salovuori, I. and Knowles, J. (1987) Homologous Domains in Trichoderma-Reesei Cellulolytic Enzymes - Gene Sequence and Expression of Cellobiohydrolase-Ii. Gene 51, 43-52.
[14] Saloheimo, M., Lehtovaara, P., Penttila, M., Teeri, T.T., Stahlberg, J., Johansson, G., Pettersson, G., Claeyssens, M., Tomme, P. and Knowles, J.K.C.

(1988) Egiii, a New Endoglucanase from Trichoderma-Reesei - the Characterization of Both Gene and Enzyme. Gene 63, 11-21.

[15] Okada, H., Tada, K., Sekiya, T., Yokoyama, K., Takahashi, A., Tohda, H., Kumagai, H. and Morikawa, Y. (1998) Molecular characterization and heterologous expression of the gene encoding a low-molecular-mass endoglucanase from Trichoderma reesei QM9414. Applied and Environmental Microbiology 64, 555-563.

[16] Ward, M., Wu, S., Dauberman, J., Weiss, G., Larenas, E., Bower, B., Rey, M.W., Clarkson, K. and Bott, R. (1993) In: Biochemistry and Genetics of Cellulose Degradation (Aubert, J.P., Beguin, P. and Millet, J., Ed.^Eds.), Vol. 1, p.^pp. 53-70. Academic Press, New York.

[17] Foreman, P.K., Brown, D., Dankmeyer, L., Dean, R., Diener, S., Dunn-Coleman, N.S., Goedegebuur, F., Houfek, T.D., England, G.J., Kelley, A.S., Meerman, H.J., Mitchell, T., Mitchinson, C., Olivares, H.A., Teunissen, P.J.M., Yao, J. and Ward, M. (2003) Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus Trichoderma reesei. Journal of Biological Chemistry 278, 31988-31997.

[18] MargollesClark, E., Tenkanen, M., Soderlund, H. and Penttila, M. (1996) Acetyl xylan esterase from Trichoderma reesei contains an active-site serine residue and a cellulose-binding domain. European Journal of Biochemistry 237, 553-560.

[19] MargollesClark, E., Saloheimo, M., Siikaaho, M. and Penttila, M. (1996) The alpha-glucuronidase-encoding gene of Trichoderma reesei. Gene 172, 171-172.

[20] MargollesClark, E., Tenkanen, M., Luonteri, E. and Penttila, M. (1996) Three alpha-galactosidase genes of Trichoderma reesei cloned by expression in yeast. European Journal of Biochemistry 240, 104-111.

[21] MargollesClark, E., Tenkanen, M., NakariSetala, T. and Penttila, M. (1996) Cloning of genes encoding alpha-L-arabinofuranosidase and beta-xylosidase from Trichoderma reesei by expression in Saccharomyces cerevisiae. Applied and Environmental Microbiology 62, 3840-3846.

[22] Mach, R.L., Strauss, J., Zeilinger, S., Schindler, M. and Kubicek, C.P. (1996) Carbon catabolite repression of xylanase I (xyn1) gene expression in Trichoderma reesei. Molecular Microbiology 21, 1273-1281.

[23] Zeilinger, S., Mach, R.L., Schindler, M., Herzog, P. and Kubicek, C.P. (1996) Different inducibility of expression of the two xylanase genes xyn1 and xyn2 in Trichoderma reesei. Journal of Biological Chemistry 271, 25624-25629.

[24] Xu, J., Takakuwa, N., Nogawa, M., Okada, H. and Morikawa, Y. (1998) A third xylanase from Trichoderma reesei PC-3-7. Applied Microbiology and Biotechnology 49, 718-724.

[25] Den Haan, R. and Van Zyl, W.H. (2001) Differential expression of the Trichoderma reesei beta-xylanase II (xyn2) gene in the xylose-fermenting yeast Pichia stipitis. Applied Microbiology and Biotechnology 57, 521-527.

[26] Rose, S.H. and van Zyl, W.H. (2002) Constitutive expression of the Trichoderma reesei beta-1,4-xylanase gene (xyn2) and the beta-1,4-endoglucanase gene (egl) in Aspergillus niger in molasses and defined glucose media. Applied Microbiology and Biotechnology 58, 461-468.

[27] Torronen, A., Mach, R.L., Messner, R., Gonzalez, R., Kalkkinen, N., Harkki, A. and Kubicek, C.P. (1992) The 2 Major Xylanases from Trichoderma-Reesei - Characterization of Both Enzymes and Genes. Bio-Technology 10, 1461-1465.

[28] Diener, S.E., Chellappan, M.K., Mitchell, T.K., Dunn-Coleman, N., Ward, M. and Dean, R.A. (2004) Insight into Trichoderma reesei's genome content, organization and evolution revealed through BAC library characterization. Fungal Genetics and Biology 41, 1077-1087.

[29] Mandels, M., Reese, E.T. and Parrish, F.W. (1962) Sophorose as an Inducer of Cellulase in Trichoderma Viride. Journal of Bacteriology 83, 400-&.

[30] Nogawa, M., Goto, M., Okada, H. and Morikawa, Y. (2001) L-Sorbose induces cellulase gene transcription in the cellulolytic fungus Trichoderma reesei. Current Genetics 38, 329-334.

[31] Ilmen, M., Saloheimo, A., Onnela, M.L. and Penttila, M.E. (1997) Regulation of cellulase gene expression in the filamentous fungus Trichoderma reesei. Applied and Environmental Microbiology 63, 1298-1306.

[32] Kubicek, C.P. (1987) Involvement of a Conidial Endoglucanase and a Plasma-Membrane-Bound Beta-Glucosidase in the Induction of Endoglucanase Synthesis by Cellulose in Trichoderma-Reesei. Journal of General Microbiology 133, 1481-1487.

[33] Vaheri, M.P., Vaheri, M.E.O. and Kauppinen, V.S. (1979) Formation and Release of Cellulolytic Enzymes During Growth of Trichoderma-Reesei on Cellobiose and Glycerol. European Journal of Applied Microbiology and Biotechnology 8, 73-80.

[34] Fowler, T. (1993) Deletion of the Trichoderma-Reesei Beta-Glucosidase Gene, Bgl1. Acs Symposium Series 533, 56-65.
[35] Mach, R.L., Seiboth, B., Myasnikov, A., Gonzalez, R., Strauss, J., Harkki, A.M. and Kubicek, C.P. (1995) The Bgl1 Gene of Trichoderma-Reesei Qm-9414 Encodes an Extracellular, Cellulose-Inducible Beta-Glucosidase Involved in Cellulase Induction by Sophorose. Molecular Microbiology 16, 687-697.
[36] Suto, M. and Tomita, F. (2001) Induction and catabolite repression mechanisms of cellulase in fungi. Journal of Bioscience and Bioengineering 92, 305-311.
[37] Mach, R.L. and Zeilinger, S. (2003) Regulation of gene expression in industrial fungi: Trichoderma. Applied Microbiology and Biotechnology 60, 515-522.
[38] Zeilinger, S., Schmoll, M., Pail, M., Mach, R.L. and Kubicek, C.P. (2003) Nucleosome transactions on the Hypocrea jecorina (Trichoderma reesei) cellulase promoter cbh2 associated with cellulase induction. Molecular Genetics and Genomics 270, 46-55.
[39] Aro, N., Ilmen, M., Saloheimo, A. and Penttila, M. (2003) ACEI of Trichoderma reesei is a repressor of cellulase and xylanase expression. Applied and Environmental Microbiology 69, 56-65.
[40] Aro, N., Saloheimo, A., Ilmen, M. and Penttila, M. (2001) ACEII, a novel transcriptional activator involved in regulation of cellulase and xylanase genes of Trichoderma reesei. Journal of Biological Chemistry 276, 24309-24314.

[41] MargollesClark, E., Ilmen, M. and Penttila, M. (1997) Expression patterns of ten hemicellulase genes of the filamentous fungus Trichoderma reesei on various carbon sources. Journal of Biotechnology 57, 167-179.
[42] Conesa, A., Punt, P.J., van Luijk, N. and van den Hondel, C. (2001) The secretion pathway in filamentous fungi: A biotechnological view. Fungal Genetics and Biology 33, 155-171.
[43] Collen, A., Saloheimo, M., Bailey, M., Penttila, M. and Pakula, T.M. (2005) Protein production and induction of the unfolded protein response in Trichoderma reesei strain rut-c30 and its transformant expressing endoglucanase I with a hydrophobic tag. Biotechnology and Bioengineering 89, 335-344.
[44] Bailey, M.J. and Tahtiharju, J. (2003) Efficient cellulase production by Trichoderma reesei in continuous cultivation on lactose medium with a computer-controlled feeding strategy. Applied Microbiology and Biotechnology 62, 156-162.
[45] Sakaguchi, M. (1997) Eukaryotic protein secretion. Current Opinion in Biotechnology 8, 595-601.
[46] Thompson, S.A., Golightly, E.J. and Yaver, D.S. (1995) Nucleotide sequence of the Aspergillus niger srpA gene. Gene 167, 337-338.
[47] Hijarrubia, M.J., Casqueiro, J., Gutierrez, S., Fernandez, F.J. and Martin, J.F. (1997) Characterization of the bip gene of Aspergillus awamori encoding a protein with an HDEL retention signal homologous to the mammalian BiP involved in polypeptide secretion. Current Genetics 32, 139-146.
[48] Sagt, C.M.J., Kleizen, B., Verwaal, R., de Jong, M.D.M., Muller, W.H., Smits, A., Visser, C., Boonstra, J., Verkleij, A.J. and Verrips, C.T. (2000) Introduction of

an N-glycosylation site increases secretion of heterologous proteins in yeasts. Applied and Environmental Microbiology 66, 4940-+.

[49] Techel, D., Hafker, T., Muschner, S., Reimann, M., Li, Y.Z., Monnerjahn, C. and Rensing, L. (1998) Molecular analysis of a glucose-regulated gene (grp78) of Neurospora crassa. Biochimica Et Biophysica Acta-Gene Structure and Expression 1397, 21-26.

[50] Diener, S.E., Dunn-Coleman, N., Foreman, P., Houfek, T.D., Teunissen, P.J.M., van Solingen, P., Dankmeyer, L., Mitchell, T.K., Ward, M. and Dean, R.A. (2004) Characterization of the protein processing and secretion pathways in a comprehensive set of expressed sequence tags from Trichoderma reesei. Fems Microbiology Letters 230, 275-282.

[51] Gotte, M. and Lazar, T. (1999) The ins and outs of yeast vacuole trafficking. Protoplasma 209, 9-18.

[52] Lazar, T., Gotte, M. and Gallwitz, D. (1997) Vesicular transport: how many Ypt/Rab-GTPases make a eukaryotic cell? Trends in Biochemical Sciences 22, 468-472.

[53] Novick, P. and Brennwald, P. (1993) Friends and Family - the Role of the Rab Gtpases in Vesicular Traffic. Cell 75, 597-601.

[54] Veldhuisen, G., Saloheimo, M., Fiers, M.A., Punt, P.J., Contreras, R., Penttila, M. and vandenHondel, C. (1997) Isolation and analysis of functional homologues of the secretion-related SAR1 gene of Saccharomyces cerevisiae from Aspergillus niger and Trichoderma reesei. Molecular & General Genetics 256, 446-455.

[55] Punt, P.J., Seiboth, B., Weenink, X.O., van Zeijl, C., Lenders, M., Konetschny, C., Ram, A.F.J., Montijn, R., Kubicek, C.P. and van den Hondel, C.A.M.J.J. (2001) Identification and characterization of a family of secretion-related small GTPase-encoding genes from the filamentous fungus Aspergillus niger: a putative SEC4 homologue is not essential for growth. Molecular Microbiology 41, 513-525.

[56] Ohsumi, K., Arioka, M., Nakajima, H. and Kitamoto, K. (2002) Cloning and characterization of a gene (avaA) from Aspergillus nidulans encoding a small GTPase involved in vacuolar biogenesis. Gene 291, 77-84.

[57] Pedrazzini, E. and Vitale, A. (1996) The binding protein (BiP) and the synthesis of secretory proteins. Plant Physiology and Biochemistry 34, 207-216.

[58] Noiva, R. (1999) Protein disulfide isomerase: The multifunctional redox chaperone of the endoplasmic reticulum. Seminars in Cell & Developmental Biology 10, 481-493.

[59] Wang, H.M. and Ward, M. (2000) Molecular characterization of a PDI-related gene prpA in Aspergillus niger var. awamori. Current Genetics 37, 57-64.

[60] Saloheimo, M., Lund, M. and Penttila, M.E. (1999) The protein disulphide isomerase gene of the fungus Trichoderma reesei is induced by endoplasmic reticulum stress and regulated by the carbon source. Molecular and General Genetics 262, 35-45.

[61] Jeenes, D.J., Pfaller, R. and Archer, D.B. (1997) Isolation and characterisation of a novel stress-inducible PDI-family gene from Aspergillus niger. Gene 193, 151-156.

[62] Ngiam, C., Jeenes, D.J. and Archer, D.B. (1997) Isolation and characterisation of a gene encoding protein disulphide isomerase, pdiA, from Aspergillus niger. Current Genetics 31, 133-138.

[63] Helenius, A., Trombetta, E.S., Hebert, D.N. and Simons, J.F. (1997) Calnexin, calreticulin and the folding of glycoproteins. Trends in Cell Biology 7, 193-200.

[64] Wang, H., Entwistle, J., Morlon, E., Archer, D.B., Peberdy, J.F., Ward, M. and Jeenes, D.J. (2003) Isolation and characterisation of a calnexin homologue, clxA, from Aspergillus niger. Molecular Genetics and Genomics 268, 684-691.

[65] Gothel, S.F. and Marahiel, M.A. (1999) Peptidyl-prolyl cis-trans isomerases, a superfamily of ubiquitous folding catalysts. Cellular and Molecular Life Sciences 55, 423-436.

[66] Joseph, J.D., Heitman, J. and Means, A.R. (1999) Molecular cloning and characterization of Aspergillus nidulans cyclophilin B. Fungal Genetics and Biology 27, 55-66.

[67] Derkx, P.M.F. and Madrid, S.M. (2001) The foldase CYPB is a component of the secretory pathway of Aspergillus niger and contains the endoplasmic reticulum retention signal HEEL. Molecular Genetics and Genomics 266, 537-545.

[68] Solscheid, B. and Tropschug, M. (2000) A novel type of FKBP in the secretory pathway of Neurospora crassa. Febs Letters 480, 118-122.

[69] Chapman, R., Sidrauski, C. and Walter, P. (1998) Intracellular signaling from the endoplasmic reticulum to the nucleus. Annual Review of Cell and Developmental Biology 14, 459-485.
[70] Welihinda, A.A., Tirasophon, W. and Kaufman, R.J. (1999) The cellular response to protein misfolding in the endoplasmic reticulum. Gene Expression 7, 293-300.
[71] Valkonen, M., Penttila, M. and Saloheimo, M. (2004) The ire1 and ptc2 genes involved in the unfolded protein response pathway in the filamentous fungus Trichoderma reesei. Molecular Genetics and Genomics 272, 443-451.
[72] Hampton, R.Y. (2002) ER-associated degradation in protein quality control and cellular regulation. Current Opinion in Cell Biology 14, 476-482.
[73] Goller, S.P., Gorfer, M. and Kubicek, C.P. (1998) Trichoderma reesei prs12 encodes a stress- and unfolded-protein response inducible regulatory subunit of the fungal 26S proteasome. Current Genetics 33, 284-290.
[74] Hemming, F.W. (1995) Expression and Secretion of Glycoproteins by Hyphal Fungi. Biochemical Society Transactions 23, 180-185.
[75] Goto, M., Ekino, K. and Furukawa, K. (1997) Expression and functional analysis of a hyperglycosylated glucoamylase in a parental host, Aspergillus awamori var. kawachi. Applied and Environmental Microbiology 63, 2940-2943.
[76] Devchand, M. and Gwynne, D.I. (1991) Expression of Heterologous Proteins in Aspergillus. Journal of Biotechnology 17, 3-9.

[77] Maras, M., van Die, I., Contreras, R. and van den Hondel, C. (1999) Filamentous fungi as production organisms for glycoproteins of bio-medical interest. Glycoconjugate Journal 16, 99-107.

[78] Palamarczyk, G., Maras, M., Contreras, R. and Kruszewska, J. (1998) Protein secretion and glycosylation in *Trichoderma*. In: Trichoderma and Gliocladium (Kubicek, C.P. and Harman, G.E., Ed.^Eds.), Vol. 1, p.^pp. 121-138. Taylor & Francis, Bristol, TN.

[79] Penttila, M. (1998) Heterologous protein production in *Trichoderma*. In: Trichoderma and Gliocladium (Harman, G.E. and Kubicek, C.P., Ed.^Eds.), Vol. 2, p.^pp. 365-379. Taylor and Francis, London.

[80] Gouka, R.J., Punt, P.J. and vandenHondel, C.A.M.J.J. (1997) Efficient production of secreted proteins by Aspergillus: Progress, limitations and prospects. Applied Microbiology and Biotechnology 47, 1-11.

[81] Gouka, R.J., Punt, P.J. and VandenHondel, C.A.M.J.J. (1997) Glucoamylase gene fusions alleviate limitations for protein production in Aspergillus awamori at the transcriptional and (post)translational levels. Applied and Environmental Microbiology 63, 488-497.

[82] Paloheimo, M., Mantyla, A., Kallio, J. and Suominen, P. (2003) High-yield production of a bacterial xylanase in the filamentous fungus Trichoderma reesei requires a carrier polypeptide with an intact domain structure. Applied and Environmental Microbiology 69, 7073-7082.

[83] Bocking, S.P., Wiebe, M.G., Robson, G.D., Hansen, K., Christiansen, L.H. and Trinci, A.P.J. (1999) Effect of branch frequency in Aspergillus oryzae on

protein secretion and culture viscosity. Biotechnology and Bioengineering 65, 638-648.

[84] Kjeldsen, T., Pettersson, A.F., Hach, M., Diers, I., Havelund, S., Hansen, P.H. and Andersen, A.S. (1997) Synthetic leaders with potential BiP binding mediate high-yield secretion of correctly folded insulin precursors from Saccharomyces cerevisiae. Protein Expression and Purification 9, 331-336.

[85] Harmsen, M.M., Bruyne, M.I., Raue, H.A. and Maat, J. (1996) Overexpression of binding protein and disruption of the PMR1 gene synergistically stimulate secretion of bovine prochymosin but not plant Thaumatin in yeast. Applied Microbiology and Biotechnology 46, 365-370.

[86] Kim, M.D., Han, K.C., Kang, H.A., Rhee, S.K. and Seo, J.H. (2003) Coexpression of BiP increased antithrombotic hirudin production in recombinant Sacchdromyces cerevisiae. Journal of Biotechnology 101, 81-87.

[87] van der Heide, M., Hollenberg, C.P., van der Klei, I.J. and Veenhuis, M. (2002) Overproduction of BiP negatively affects the secretion of Aspergillus niger glucose oxidase by the yeast Hansenula polymorpha. Applied Microbiology and Biotechnology 58, 487-494.

[88] Valkonen, M., Ward, M., Wang, H.M., Penttila, M. and Saloheimo, M. (2003) Improvement of foreign-protein production in Aspergillus niger var. awamori by constitutive induction of the unfolded-protein response. Applied and Environmental Microbiology 69, 6979-6986.

[89] Conesa, A., Jeenes, D., Archer, D.B., van den Hondel, C. and Punt, P.J. (2002) Calnexin overexpression increases manganese peroxidase production in Aspergillus niger. Applied and Environmental Microbiology 68, 846-851.

[90] Kruszewska, J.S., Butterweck, A.H., Kurzatkowski, W., Migdalski, A., Kubicek, C.P. and Palamarczyk, G. (1999) Overexpression of the Saccharomyces cerevisiae mannosylphosphodolichol synthase-encoding gene in Trichoderma reesei results in an increased level of protein secretion and abnormal cell ultrastructure. Applied and Environmental Microbiology 65, 2382-2387.

[91] Ward, M., Lin, C., Victoria, D.C., Fox, B.P., Fox, J.A., Wong, D.L., Meerman, H.J., Pucci, J.P., Fong, R.B., Heng, M.H., Tsurushita, N., Gieswein, C., Park, M.H. and Wang, H.M. (2004) Characterization of humanized antibodies secreted by Aspergillus niger. Applied and Environmental Microbiology 70, 2567-2576.

[92] Maras, M., Saelens, X., Laroy, W., Piens, K., Claeyssens, M., Fiers, W. and Contreras, R. (1997) In vitro conversion of the carbohydrate moiety of fungal glycoproteins to mammalian-type oligosaccharides - Evidence for N-acetylglucosaminyltransferase-I-accepting glycans from Trichoderma reesei. European Journal of Biochemistry 249, 701-707.

[93] Kalsner, I., Hintz, W., Reid, L.S. and Schachter, H. (1995) Insertion into Aspergillus-Nidulans of Functional Udp-Glcnac - Alpha-3-D-Mannoside Beta-1,2-N-Acetylglucosaminyl-Transferase-I, the Enzyme Catalyzing the First Committed Step from Oligomannose to Hybrid and Complex N-Glycans. Glycoconjugate Journal 12, 360-370.

[94] Maras, M., De Bruyn, A., Vervecken, W., Uusitalo, J., Penttila, M., Busson, R., Herdewijn, P. and Contreras, R. (1999) In vivo synthesis of complex N-glycans by expression of human N-acetylglucosaminyltransferase I in the filamentous fungus Trichoderma reesei. Febs Letters 452, 365-370.
[95] Vervecken, W., Kaigorodov, V., Callewaert, N., Geysens, S., De Vusser, K. and Contreras, R. (2004) In vivo synthesis of mammalian-like, hybrid-type N-glycans in Pichia pastoris. Applied and Environmental Microbiology 70, 2639-2646.
[96] Wildt, S. and Gerngross, T.U. (2005) The humanization of N-glycosylation pathways in yeast. Nature Reviews Microbiology 3, 119-128.
[97] Hamilton, S.R., Bobrowicz, P., Bobrowicz, B., Davidson, R.C., Li, H.J., Mitchell, T., Nett, J.H., Rausch, S., Stadheim, T.A., Wischnewski, H., Wildt, S. and Gerngross, T.U. (2003) Production of complex human glycoproteins in yeast. Science 301, 1244-1246.
[98] Mcclintock, B. (1950) The Origin and Behavior of Mutable Loci in Maize. Proceedings of the National Academy of Sciences of the United States of America 36, 344-355.
[99] Daboussi, M.J. (1997) Fungal transposable elements and genome evolution. Genetica 100, 253-260.
[100] Finnegan, D.J. (1989) Eukaryotic Transposable Elements and Genome Evolution. Trends in Genetics 5, 103-107.
[101] Kempken, F. and Kuck, U. (1998) Transposons in filamentous fungi - facts and perspectives. Bioessays 20, 652-659.

[102] Farman, M.L., Tosa, Y., Nitta, N. and Leong, S.A. (1996) MAGGY, a retrotransposon in the genome of the rice blast fungus Magnaporthe grisea. Molecular & General Genetics 251, 665-674.

[103] Eto, Y., Ikeda, K., Chuma, I., Kataoka, T., Kuroda, S., Kikuchi, N., Don, L.D., Kusaba, M., Nakayashiki, H., Tosa, Y. and Mayama, S. (2001) Comparative analyses of the distribution of various transposable elements in Pyricularia and their activity during and after the sexual cycle. Molecular and General Genetics 264, 565-577.

[104] Thon, M.R., Martin, S.L., Goff, S., Wing, R.A. and Dean, R.A. (2004) BAC end sequences and a physical map reveal transposable element content and clustering patterns in the genome of Magnaporthe grisea. Fungal Genetics and Biology 41, 657-666.

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

[106] Farman, M.L., Eto, Y., Nakao, T., Tosa, Y., Nakayashiki, H., Mayama, S. and Leong, S.A. (2002) Analysis of the structure of the AVR1-CO39 avirulence locus in virulent rice-infecting isolates of Magnaporthe grisea. Molecular Plant-Microbe Interactions 15, 6-16.

[107] Kang, S.C., Sweigard, J.A. and Valent, B. (1995) The PWL host specificity gene family in the blast fungus Magnaporthe grisea. Molecular Plant-Microbe Interactions 8, 939-948.

[108] Kang, S., Lebrun, M.H., Farrall, L. and Valent, B. (2001) Gain of virulence caused by insertion of a Pot3 transposon in a Magnaporthe grisea avirulence gene. Molecular Plant-Microbe Interactions 14, 671-674.

[109] Brookfield, J.F.Y. (2005) The ecology of the genome - Mobile DNA elements and their hosts. Nature Reviews Genetics 6, 128-136.

Chapter 2

Alkahest NuclearBLAST :

a user-friendly BLAST management and analysis system

Stephen E Diener[1*], Thomas D Houfek[2*], Sam E Kalat[2], D E Windham[1], Mark Burke[2], Charles Opperman[2], Ralph A Dean[1§]

[1] Fungal Genomics Laboratory, Center for Integrated Fungal Research, North Carolina State University, Raleigh, NC 27695, USA

[2] Plant Nematode Genetics Group, North Carolina State University, Raleigh, NC 27695, USA

* These authors contributed equally to this work
[§] Corresponding author

Email Addresses
SED: sediener@unity.ncsu.edu
TDH: tdhoufek@unity.ncsu.edu
SEK: samkalat@sneakrets.net
DEW: dewindha@unity.ncsu.edu
MB: mark_burke@ncsu.edu
CO: warthog@unity.ncsu.edu
RAD: ralph_dean@ncsu.edu

Abstract

Sequencing of EST and BAC end datasets is no longer limited to large research groups. Drops in per-base pricing have made high throughput sequencing accessible to individual investigators. However, there are few options available which provide a free and user-friendly solution to the BLAST result storage and data mining needs of biologists. Here we describe NuclearBLAST, a batch BLAST analysis, storage and management system designed for the biologist. It is a wrapper for NCBI BLAST which provides a user-friendly web interface which includes a request wizard and the ability to view and mine the results. All BLAST results are stored in a MySQL database which allows for more advanced data-mining through supplied command-line utilities or direct database access. NuclearBLAST can be installed on a single machine or clustered amongst a number of machines to improve analysis throughput. NuclearBLAST provides a platform which eases data-mining of multiple BLAST results. With the supplied scripts, the program can export data into a spreadsheet-friendly format, automatically assign Gene Ontology terms to sequences and provide bi-directional best hits between two datasets. Users with SQL experience can use the database to ask even more complex questions and extract any subset of data they require. This tool provides a user-friendly interface for requesting, viewing and mining of BLAST results which makes the management and data-mining of large sets of BLAST analyses tractable to biologists.

Background

Recently the number of research groups generating sequence data such as expressed sequence tags (EST) and bacterial artificial chromosome (BAC) ends has increased dramatically due to dropping costs of obtaining DNA sequence. The size of groups involved in such projects ranges from members of sequencing centers with a large bioinformatic support staff to individual researchers with little data management expertise. As it is now typical for even the smallest projects to generate hundreds to thousands of sequences, tools that streamline annotation, data mining and data management have become increasingly important to biologists without large bioinformatic support staffs.

Annotation of a large set of sequences typically involves attempting to subdivide the sequences into common functional categories. The usual course of action followed is to perform one or more BLAST searches against databases to reveal homologies with sequences previously functionally annotated. Using these homologies, categorical annotations such as Gene Ontology terms can be associated with the novel sequences and used to evaluate the sequence dataset. When dealing with the large numbers of sequence reads generated by these projects, storage and compilation of flatfile BLAST results related to this endeavor becomes rather cumbersome. Furthermore, comparative analysis of gene complement between related organisms can provide a number of insights into many areas of biology. Identification of a subset of genes which exist in a pathogen but not in a related non-pathogenic organism can provide targets for other functional analyses such as gene disruption experiments. While there exist other BLAST clustering packages, such as BeoBLAST or provided by NCBI, they do not provide a free open-source complete result management package with storage of results in a relational database (RDMS) [1,7]. The ability to easily locate, data-mine and perform comparative analysis on BLAST results can be substantially simplified through the use of a RDMS. In order to facilitate sequence analysis and comparative genomics we have created NuclearBLAST.

Implementation

The primary design goals for NuclearBLAST were to provide biologists a centralized system where BLAST results can be easily created and retrieved, a relational database storage system which can be easily mined for comparative analyses, and a program which would take advantage of clustered computing resources to increase the throughput of large BLAST jobs. A secondary objective was to design a solution which was open-source and freely available. This led us to preferentially utilizing a number of open-source software packages in the implementation of NuclearBLAST including BioPerl, Apache, PHP and MySQL as well as using Linux as the base operating system [2-5].

A web browser was chosen as the interface for NuclearBLAST since it provides a widely used, recognizable, globally available, platform independent interface. The Apache web server provides access control and encrypted connections if the user desires to secure their installation. BioPerl provides a parsing module for the returned BLAST results which are then loaded into the MySQL database. All information about requested BLAST searches are loaded into the database and the status of each query in a job is tracked in the database. This client server design of the program allows for the use of multiple computers performing the BLAST analyses in a clustered environment by using a job management software package such as PBS (Portable Batch System) [6]. NuclearBLAST can be clustered with as little as several lab computers used as worker nodes in their down time or as much as a dedicated compute farm. A minimal installation of NuclearBLAST requires a typical workstation machine acting as both the server and the worker.

In order to keep the MySQL database to a manageable size and reduce redundant data in the system we opted to use BLAST's database format as the only store of sequence information in the program. Only the sequence names within each dataset and a minimal amount of metadata are stored in the MySQL database. We also minimized the database size by storing all result statistics in the database but excluding the actual alignments. When requested, the alignments are recreated on the fly by extracting the two sequences from the BLAST databases and blasting them against each other using bl2seq.

Results

NuclearBLAST is a free open source batch BLAST analysis, storage and management system which provides a simple platform for performing BLAST analyses and mining of the results. NuclearBLAST is written in Perl which drives NCBI's *blastall* application to conduct its analyses and store all BLAST results in a MySQL database [7]. Users may import datasets, request analyses, and view results in the PHP web interface through their browser. Command-line programs support more complex or special-purpose queries. NuclearBLAST may be installed on a single machine, or it may be set up to distribute BLAST analyses to a cluster of machines by way of a batch queuing system.

Sequence files are imported in FASTA format. NuclearBLAST uses NCBI's *formatdb* utility to parlay sequences into the format required for *blastall* targets. When importing a sequence data set into NuclearBLAST, a user specifies whether the sequence may be used in subsequent searches as a query, a target, or both. Specifying 'both' allows easy reciprocal BLASTs between smaller datasets. These parameters can be set to allow users access to large datasets as targets but not as queries. This guards against the casual mistake of querying very large datasets, such as the Genbank non-redundant protein database, against much smaller datasets (rather than vice versa). If unnoticed, such a mishap can divert a laboratory's analysis system for extended periods of time.

Requesting batch BLAST jobs through the web interface is facilitated by a job request wizard which guides the user through the process. The user first chooses a query dataset (from datasets designated as allowable queries) (Figure 2.1a). The choice of query (specifically, whether it is a nucleotide or protein sequence) automatically limits the menu of programs of the BLAST “family” (BLASTN, TBLASTX, etc) to those appropriate for the type of query sequence (Figure 2.1b). The user's choice of program dictates the sequence type (nucleotide/protein) appropriate for targets; in the next stage of the wizard, when the user selects a target sequence dataset, only choices of the appropriate type appear on the menu (Figure 2.1c). The job request is completed by specifying an e-value threshold for returned results. (More advanced BLAST parameters can be stipulated by using NuclearBLAST's command-line request utility).

After a request is submitted, the status of the job can be continually monitored on the job's main page (Figure 2.2a). Invocations of *blastall* are scheduled by a simple first-in-first-out scheme. NuclearBLAST parses *blastall's* output files using BioPerl and loads their constituent data into the database [8]. The work flow of a job from request to completion can be seen in Figure 2.3.

Results of a completed batch BLAST job can be browsed as ordered by e-value in a paginated fashion with the top hit available on the main result page (Figure 2.2b). Clicking on the query calls up a page which shows all hits to that sequence found, along with a graphical view of where each hit aligned to the sequence (Figure 2.2c). Each hit further contains a link to a page showing the alignment of the HSPs and other associated statistic (Figure 2.2d). In addition to ordered hierarchical browsing, NuclearBLAST offers a facility for searching for strings within text fields associated with the job (names or descriptions of queries or hits). This can help researchers quickly find certain selected targets amidst a large volume of search results.

In addition to providing an easy web-based method of perusing BLAST results, NuclearBLAST supplies a number of command line scripts allowing expanded access to search results in the database. One such script exports a job's results to a tab delimited file for inclusion in a spreadsheet or publication.

Another determines reciprocal best hits between two datasets that have been reciprocally BLASTed. A third script may be used to transfer Gene Ontology (GO) annotations to query sequences that have matches against a database annotated with these terms. As Compugen has made available GO annotations for much of the Genbank non-redundant protein database at the Gene Ontology website, one can easily annotate a large set of sequences with GO terms using NuclearBLAST [9, 10]. This provides the researcher an easy way to categorize their sequences into functional groupings. Future work on NuclearBLAST will extend the mining capabilities available on the command line as well as through the web interface and also expand integration with clustering software.

Conclusion

NuclearBLAST provides a powerful tool to biologists for data mining and comparative genomic analysis of generated sequence. It has shown its utility in prior studies [11]. This program provides a simple interface for performing large batch BLAST searches, effectively manages a large number of search results, presents those results in an intelligibly browsable format, and provides an extensible platform for more thorough data mining of BLAST results.

Availability and requirements

The program, source and full documentation for installation and use are available at http://www.alkahest.org as well as in the additional file section [see Additional File 1].Installation of the software is walked through in the documentation and requires personnel with Unix/Linux installation experience. The software is licensed under the GNU GPL and requires PHP, Perl, Apache, and MySQL.

Acknowledgements

We thank Michael Thon, Patsy Little, members of the Center for Integrated Fungal Research and the Plant Nematode Genetics Group for testing and feedback. This work is supported in part by an NSF IGERT fellowship to S.E.D. through the Functional Genomics graduate program, NSF Plant Genome program (#0115642), UNC office of the President (UNC RA FY 2002), Genencor International, Phillip Morris, and NCAES.

References

[1] Grant JT, Dunbrack RL, Manion FJ, and Ochs MF: BeoBLAST: distributed BLAST and PSI-BLAST on a Beowulf cluster *Bioinformatics* 2002, 18:765-766.

[2] BioPerl Website [http://www.bioperl.org]

[3] Apache Website [http://www.apache.org]

[4] PHP Website [http://www.php.net]

[5] MySQL Website [http://www.mysql.org]

[6] PBS Website [http://www.openpbs.org]

[7] Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ: Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 1997, 25(17):3389-3402. [http://www.ncbi.nlm.nih.gov/blast]

[8] Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD, Birney E: The Bioperl toolkit: Perl modules for the life sciences. *Genome Research* 2002 Oct, 12(10):1611-8

[9] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT: Gene Ontology: tool for the unification of biology. *Nat Genet* 2000, 25(1): 25-29. [http://www.geneontology.org]

[10] Xie HQ, Wasserman A, Levine Z, Novik A, Grebinskiy V, Shoshan A, Mintz L: Large-scale protein annotation through gene ontology. *Genome Res* 2002, 12(5): 785-794.

[11] Diener SE, Dunn-Coleman N, Foreman P, Houfek TD, Teunissen PJM, van Solingen P, Dankmeyer L, Mitchell TK, Ward M, Dean RA: Characterization of the protein processing and secretion pathways in a comprehensive set of expressed sequence tags from *Trichoderma reesei*. *Fems Microbiol Lett* 2004, 230: 275-282

Figure Legends

Figure 2.1 Screenshots of the NuclearBLAST Job Request Wizard

Panel **a** shows the set of possible query sets in the system. Upon choosing one of these, the appropriate BLAST sub-programs are made available in Panel **b**. After choosing one of the sub-programs, Panel **c** arises which allows you to limit the e-value of stored results and gives choices of BLAST target databases in the system which are acceptable based upon prior decisions.

Figure 2.2 Screenshots of the NuclearBLAST web interface.

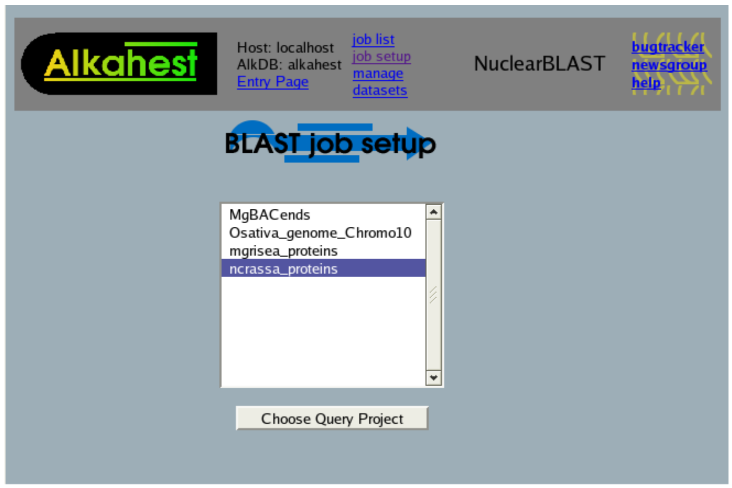
Panel **a** shows the requested jobs and their progress. Clicking on a hyperlinked job number brings up a page, panel **b**, containing the results of multiple query sequences within that requested job. The link on this page showing the number of hits fetches the location and statistics of multiple hits to that query sequence, panel **c**. Examining a hit further provides a view of the HSP locations and statistics for a single hit, panel **d**. Hyperlink of a sequence name retrieves a page with information about that particular sequence.

Figure 2.3 Flow of a NuclearBLAST Job

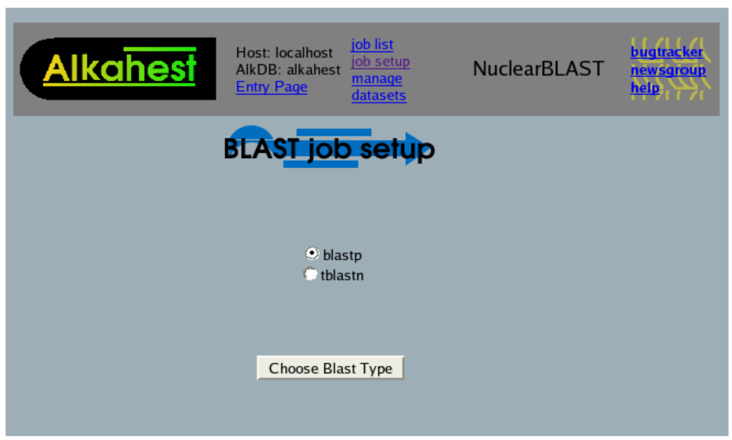
Illustration of the program's work flow when progressing through a requested BLAST analysis

Figure 2.1

a



b



c

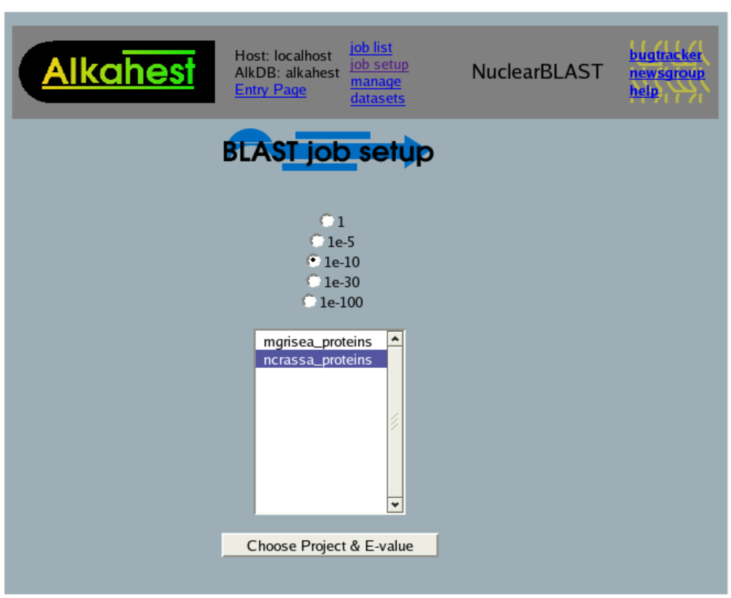
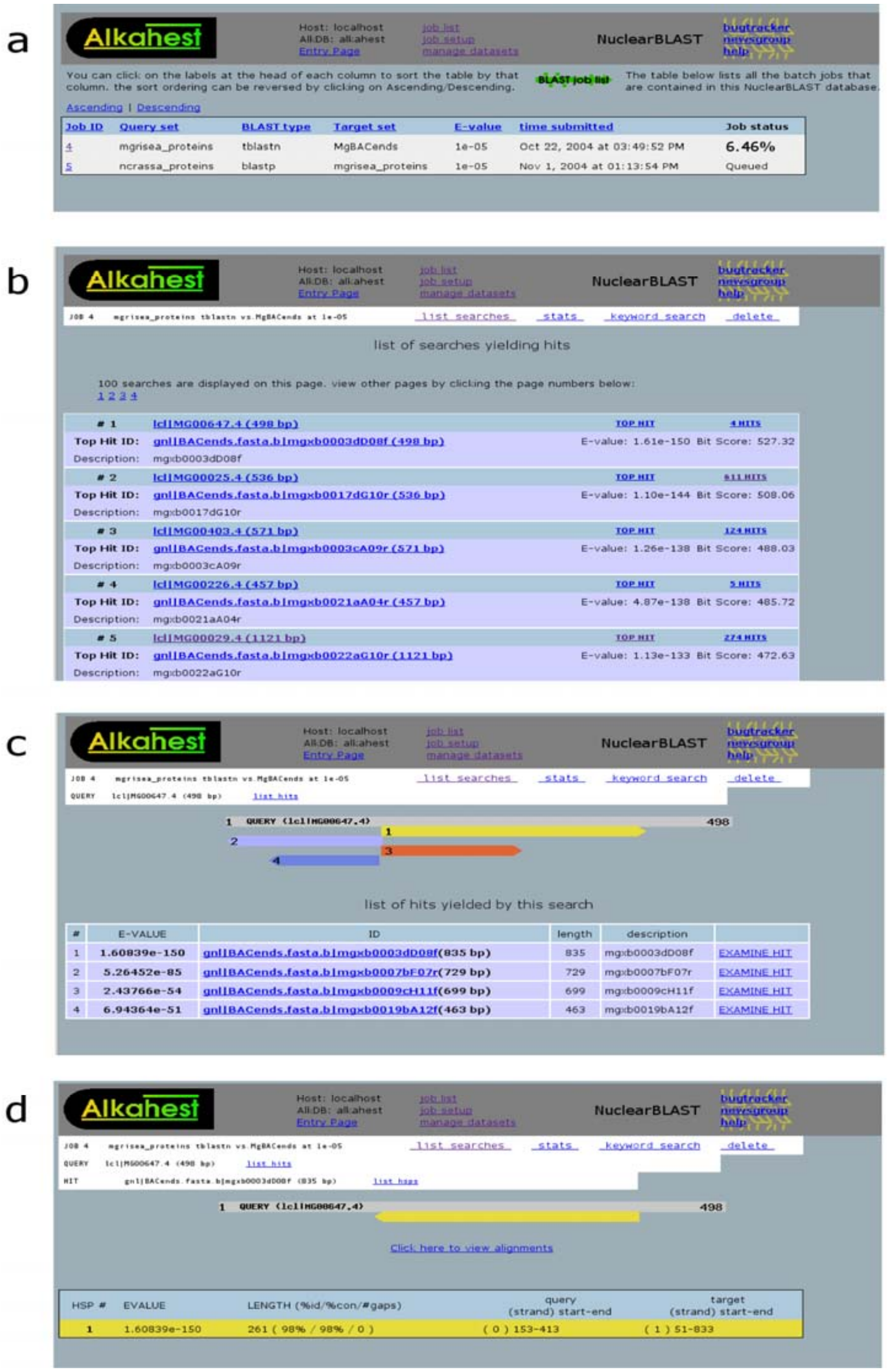
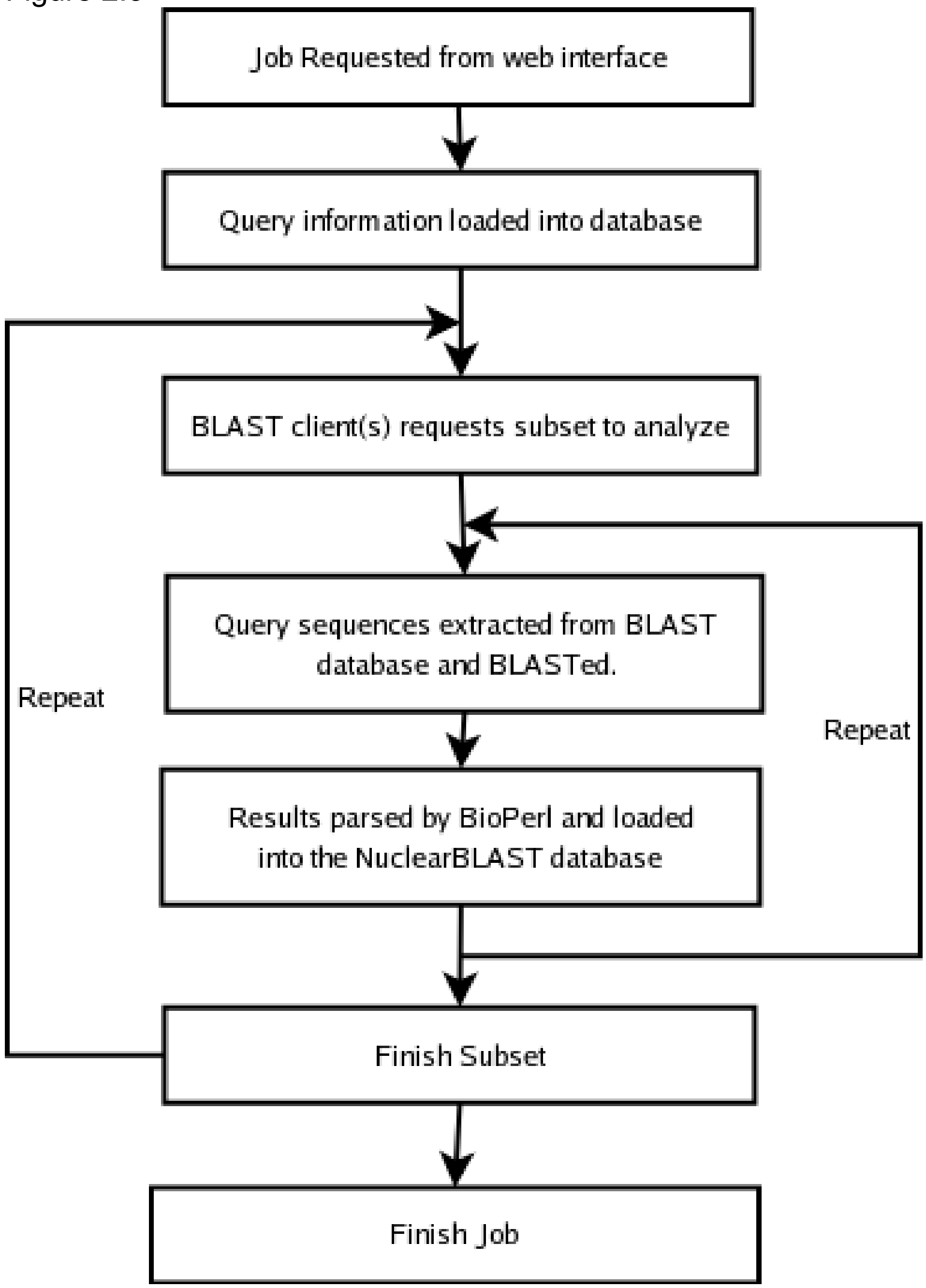


Figure 2.2



a
b
c
d
Alkahest
NuclearBLAST
list of searches yielding hits
list of hits yielded by this search
Click here to view alignments

Figure 2.3



Job Requested from web interface
Query information loaded into database
BLAST client(s) requests subset to analyze
Query sequences extracted from BLAST database and BLASTed.
Results parsed by BioPerl and loaded into the NuclearBLAST database
Finish Subset
Finish Job
Repeat
Repeat

Chapter 3

Characterization of the protein processing and secretion pathways in a comprehensive set of expressed sequence tags from *Trichoderma reesei*

S.E. Diener[1], N. Dunn-Coleman [2], P. Foreman [2], T.D. Houfek [1], P.J.M. Teunissen[3], P. van Solingen [3], L. Dankmeyer [3], T.K. Mitchell[1], M. Ward[2], R.A.Dean[1]

[1] Fungal Genomics Laboratory, North Carolina State University, Raleigh, North Carolina

[2] Genencor International, Inc., Palo Alto, California 94304

[3] Genencor International B.V., 2333 CN Leiden, The Netherlands

Corresponding Author:
Ralph A. Dean
Suite 1200
840 Main Campus Drive,
Raleigh, NC 27606
Phone 919-513-0022,
Fax 919-513-0024
Ralph_Dean@ncsu.edu

Keywords: secretion, protein processing, Trichoderma, EST

Abstract

Trichoderma reesei is a filamentous fungus widely used as an efficient protein producer and known to secrete large quantities of biomass degrading enzymes. Much work has been done aimed at improving the secretion efficiency of this fungus. It is generally accepted that the major bottlenecks in secretion are the protein folding and ornamentation steps in this pathway. In an attempt to identify genes involved in these steps, we have generated and 5' sequenced an EST library of 21,888 clones consisting of a unique set of over 5,000 which were also 3' sequenced. Using annotation tools we associated Gene Ontology terms to 2,732 of the sequences. We were able to identify homologs to the majority of *A. niger's* Srg genes as well as a number of homologs to genes involved in protein folding and ornamentation pathways.

Introduction

Trichoderma reesei is a highly cellulolytic filamentous fungi which secretes a variety of biomass degrading enzymes and is widely used in industrial applications. Strains have been reported to produce up to forty grams of protein per liter in fermentation experiments [1]. *T. reesei* has also attracted interest as a host for heterologous protein expression and secretion. A number of genes encoding enzymes implicated in degradation of plant material have been cloned and characterized from this fungus [2].

Major impediments towards further enhancements in heterologous gene expression are the rates of protein folding and post-translational ornamentation [3]. The secretion system of *T. reesei* is not well understood at the genetic level. Other fungal systems, such as *S. cerevisiae*, have been better characterized and ultrastructural comparisons to *T. reesei* show many similarities [4, 5] . Homologues of several of the gene products identified in *Saccharomyces cerevisiae* as being involved in secretion have been identified in *Aspergillus niger,* but few of these have been positively identified in *T. reesei*. A set of small GTP-binding proteins belonging to the Ras superfamily has been shown in *S. cerevisiae* to be key for intra-organelle transport of nascent proteins. These proteins can be divided into two major groups: the ARF/SAR family, involved in vesicle creation, and the SEC4/YPT/RAB family, involved in targeting and vesicle fusion. Each of these GTP-binding proteins is distributed in a specific cellular location and acts upon vesicle movement in a directional manner [6, 7]. Homologues of these genes have been identified in *A. niger* but, with the exception of SAR1 [8], there has been little work done to identify these in *T. reesei*.

Post-translational modification is an essential step in the processing of a biologically active secreted protein. With an estimated fifty percent of proteins being glyco-proteins, understanding the modification capabilities of *T. reesei* is essential for increasing heterologous protein expression. *T. reesei's* modification machinery produces both N-linked and O-linked oligosaccharides [9]. A study of the biosynthesis of O-linked mannosyl groups has resolved a unique dolicholphosphate (PD) dependent pathway [10]. This pathway is comprised of dolichol kinase, mannosylphosphodolichol synthase (MPD-synthase) and GTP:alpha-D-mannose-1-phosphate guanyl transferase. These enzymes transfer mannose through a dolicholphospate intermediate to add an O-linked mannose [11].

Modification of *T. reesei's* secretion system has brought about strains with increased yield and purity of expressed proteins. Attempts to create more efficient industrial strains have primarily relied upon the methods of random mutagenesis, deletion of highly secreted gene products and up-regulation of genes involved in secretion. Recent interest has been focused on enzymes performing post-translational modifications [12]. Up-regulation of genes involved in the secretion pathway, specifically those involved in post-translational modifications, presents an opportunity for increasing yields beyond what has already been achieved [13]. Through the creation of a comprehensive EST library and subsequent analysis, this study probes for genes involved in the protein processing and secretion pathways of *T. reesei*.

Materials and Methods

Media, strains and culture growth conditions

T. reesei strains used in this study were obtained from the American Type Culture collection. Liquid minimal medium was as described in [14], except that 100 mM PIPPS (Calbiochem) was included to maintain the pH at 5.5. Vogels medium was described by Davis and Serres [15]. YEG medium contains 0.5% yeast extract (Difco), 2% glucose.

cDNA library and sequencing

Two different libraries were produced for this study from *T. reesei* strain QM6a mycelia. For both libraries, mycelia were grown in baffled flasks at 30˚C for 24 h in YEG medium with vigorous aeration and 5 ml of the base culture was added to 50 ml of the secondary medium. LT002 was grown in secondary media with cellulose as the primary carbon source and LT003 was grown in secondary media with a variety of carbon sources including simple sugars as described [2].

Sequencing, bioinformatics and data analysis

cDNA library construction and initial EST sequencing were performed as previously described [2]. Sequences containing >100 bases with Phred quality values $\geq$20, or an average quality $\geq$12 were assembled using the PhredPhrap (version 0.990329) script provided by Consed (version 11.0) [16, 17]. Using custom Perl scripts, representative clones from each of the contigs were chosen and appended to the list of singletons to be re-arrayed and 5' and 3' sequenced. Post sequencing of the rearrayed clones, all sequence was processed and analyzed by Alkahest Interceptor (www.alkahest.org). Sequences containing >100 bases of quality trimmed sequence were extracted, assembled using the CAP3 assembler, and viewed using Consed [18]. Statistics of the assemblies were generated by perl scripts using BioPerl modules [19].

Annotation and Functional Binning

Annotation and functional binning was accomplished using tools provided by the Gene Ontology Consortium [20] in combination with Alkahest NuclearBLAST and InterPro's search tool InterProScan [21]. NuclearBLAST is a batch Blast analysis and storage system developed in our laboratory and available as part of Alkahest (www.alkahest.org) [22] . We annotated the sequences with Gene Ontology (GO) terms using Blast against the GO Consortium's database of manually annotated genes and Genbank's protein database (nr) [20]. Results from these searches with e-values better than 1e-30 were used to annotate the EST sequences with GO terms. Alignments to nr were translated using the annotation file provided by Compugen [23]. InterProScan was used to search the EST sequences for known protein domains and families and translate the annotation into GO terms. The GO term annotations were merged together and loaded into the AmiGO browser and database (www.godatabase.org). Sequences were translated in six reading frames and the first 50 amino acids were searched for existence of a signal peptide using both the HMM and neural network modules of SignalP [24].

Results

Sequencing and Assembly

After initial 5-prime sequencing of the LT002 and LT003 *Trichoderma reesei* EST libraries, separate CAP3 assemblies were produced of each library. Assembly of library LT002 produced 1,345 contigs and 2,597 singlets and assembly of LT003 library resulted in 716 contigs and 1,907 singlets. Results of BlastX against the Genbank non-redundant protein database of the top 20 most redundant sequences from each librariy can be seen in Tables 3.1 and 3.2. Both sets of sequences contained elongation factors, heat shock proteins, and cellobiohydrolases as well as a number of putative proteins from *Neurospora crassa*. Combined assembly of these libraries produced a unique set of 2,088 contigs and 3,341 singlets totaling 5,429 unique 5' sequences.

After re-arraying and 5’ and 3’ sequencing a representative clone from each contig and the singletons the set of trimmed sequences contained 12,783 initial 5-prime sequences, 3,800 re-arrayed 5-prime sequences and 3,647 re-arrayed 3-prime sequences. All sequences were submitted to Genbank and assigned accession numbers CF865336-CF886831. CAP3 assembly resulted in 4,438 contigs and 2,505 singlets totaling 7,943 unique sequences. Within the set of singlets, there were 1,196 3' sequences and 1,309 5' sequences. The 4,438 contigs contained 1,006 instances where the 5’ and 3’ read assembled together. There were 971 distinct contigs which contained at least one set of paired reads. 2,954 and 249 of the 4,438 contigs were composed solely of 5’ or 3’ reads, respectively. The remaining 264 contigs were composed of non-matching 5’ and 3’ reads.

The results of BlastX against the Genbank non-redundant protein database of the 20 most redundant sequences from this assembly can be seen in Table 3.3. The set of Blast hits to these redundant clones, as expected, corresponded well to the separate sets with one exception. Contig3675 aligned to a putative 14-3-3 protein from *T. reesei* which was not seen as a hit in the previous sets.

Functional Binning

Functional binning of this set of ESTs was accomplished by using Blast and Interpro analysis to assign GO terms. Blast searching was performed against the nr protein database and the manually annotated Gene Ontology gene (GO_gene) database. The Blast against the nr protein database produced 4,497 alignments at an evalue of 1e-5 (64.8%). Of these 1,304 were annotated with GO terms using the mapping tables limited to an e-value cutoff of 1e-30. The Blast search against the GO_genes database produced 2,932 alignments at an e-value of 1e-5 (42.2 %). At an e-value of 1e-30 GO terms were annotated to 1,383 of the ESTs from this Blast analysis. Interpro analysis returned at least one protein domain to 3,892 of the assembled sequences. Of these, 1,954 were annotated with GO terms. After merging all GO terms from each of the annotations, 2,732 unique sequences had at least one GO term annotated to them.

The three major ontology subsections contained close to an equivalent number of annotated genes within them: biological process with 2,428, molecular function with 2,544, and cellular component with 1,972 sequences. Within the biological process ontology the subsections containing the majority of genes were metabolism with 1,906 and physiological processes with 2,299 annotated. Almost all of the genes annotated to the cellular component ontology were annotated to the cell subsection (1,817) with a minority of 74 annotated to extracellular. The two primary groups annotated within the molecular function ontology were enzyme, with 1,344, and ligand-binding, with 1,309.

Ontology terms of primary interest to this study are hydrolase, catabolism, transport and all associated children of these. There were 551 genes annotated with hydrolase activity, 334 to catabolism and 479 to transport. The hydrolase category was further broken down into the subcategories of ATPase, GTPase, peptidase, and lyase activity with 101, 28, 197, and 70 annotations, respectively. Catabolism contained 229 genes binned to protein catabolism, 56 to carbohydrate catabolism and 19 to lipid catabolism. The transport category contains protein transport with 123, intracellular with 131, secretory pathway with 32 and vesicle mediated transport with 47 sequences binned within. SignalP analysis of the assembled dataset predicted a signal sequence in the first 50 amino acids in 2,644 unique sequences.

Protein Transport Analysis

Using the AmiGO browser, we identified 479 sequences annotated as having a function in 'transport'. A large number of these sequences were annotated to functions not dealing with protein transport, such as carbohydrate or hydrogen transport, and were not analyzed further. The regions of the ontology which were scrutinized more thoroughly were ones dealing specifically with steps in the locomotion of immature protein through its processing. Intracellular protein transport was a large category consisting of 118. Within that category were 44 sequences annotated as protein targeting. Homologues to a number of sequences involved in vesicle formation, trafficking and docking between organelles were found (Table 3.4). We noted a number of coatomer and clathrin subunits, both involved in vesicle budding from the donor membrane and transport through the protein processing pathway.

Another set of genes of particular interest to this study were those associated with the GO term GTPase. Any gene associated with the GTPase term was a candidate for being a homologue of the SEC4/YPT/RAB or the ARF/SAR family. From the 28 sequences in this category we were able to locate a number which were possible homologues to genes in this family. As can be seen from Table 3.5, we have located homologues for the majority of the *A. niger* Srg gene family (corresponding to the SES4/YPT/RAB family of *S. cerevisiae*) as well as other genes involved in regulation of protein trafficking. We identified homologs to three of the six known Srg genes (SrgB, C, and E) all of which are involved in the various transport steps of vesicles as well as a Sar1 homologue previously identified in *T. reesei*. [8]

Post-translational Modification

The GO category of 'transferase activity' was populated with 294 sequences. Within this the largest GO category was that of 'transferring phosphorous-containing groups' with 120 members. The 'transferring glycosly groups' category contained three major subcategories of interest; 'acetylglucosaminlytransferase activity', 'glucosyltransfease activity', and 'mannosyltransferase activity' with 7, 8 and 6 sequences binned respectively. The primary players in the DP-mannosyl transfer pathway were identified including two DP-mannose mannosyltransferases. A number of glycosyl and oligosaccharyl transferases were also present in the dataset along with a possible homolog to Ktr1, an N- and O-linked mannosyl transferase. Proteins involved in folding such as protein disulfide isomerase (PDI), peptidyl prolyl isomerase related protens cypB/C, and several chaperones were also found in our EST libraries. Sequences of particular interest are outlined in Table 3.6.

Discussion

In this study, *T. reesei* EST libraries were sequenced from two different conditions with the goal of identifying novel genes involved in biomass degradation and the pathways of protein secretion. Assembly of both libraries together produced 5,429 unique 5' sequences from over twenty one thousand clones sequenced. The genomes of *N. crassa* and *M. grisea* both near a size of 40 megabases(Mb) and containing 10,082 and 11,108 predicted proteins respectively [25]. Assuming a similar gene density, with an estimated genome size of 33 Mb, the EST sequences identified in this study represent well over fifty percent of *T. reesei's* genomic complement of coding regions. However, an unknown number of these sequences are non-overlapping regions of the same gene and assembly may merge closely related members of the same protein family misrepresenting the actual number of sequences identified here.

In assembling the combined datasets with 3' sequence, there resulted a discrepancy of 264 contigs containing non-paired 5' and 3' reads. One expects that the 3' read of a clone to assemble in one of three ways; overlapping its 5' read, as an unassembled singlet, or as a contig with other 3' reads. In this case, the 3' reads assembling with non-matching 5' reads can be explained by the presence of a set of 5' reads sequenced from different parts of the same gene, failure of the 3' read of one of these and/or truncation of an EST.

If any expression differences existed between the cellulose-only, LT002, and the simple-sugar doped, LT003, libraries, they could not be seen by comparing the annotation of the most redundant clones. The twenty most sequenced clones from each library were similar in annotation and the number of sequences lacking useful annotation. Certain genes, such as EF1-alpha and CBH, were expected to be seen in this highly redundant set due to their obvious roles in secretion and the targeting of the library. What was not expected was the large percentage (35%) of the most redundant genes which either had no hit to Genbank or were only annotated as *N. crassa* hypothetical proteins. As noted earlier, when the two libraries were combined into a single dataset, a gene not previously seen in the most redundant lists surfaced. This 14-3-3-like protein, ftt2, was previously identified through hybridization to ftt1, a candidate secretion pathway member. However, further studies suggest that ftt2 was not involved in secretion. The fact that this gene was listed as one of the most redundantly sequenced clones in a library highly expressing secretion related proteins supports that this protein may be involved with the secretion pathway [26].

While over 2,700 unique sequences were annotated to Gene Ontology terms, close to two thirds of the sequences were not annotated or were annotated to genes of unknown function, thus they could not be used in binning to GO terms. A large percentage (70%) of the binned sequences were annotated under the major category of 'metabolism' with a subset of 334 involved in 'catabolism' with only 7 involved in 'cellulose catabolism'. A very small number of sequences, 3, were annotated to be involved in 'protein secretion'. The purpose of these libraries was to identify genes involved in cellulose degradation and secretion. While it may seem, looking at these numbers, that the libraries were not targeted to our needs, the Blast results of the most redundant genes show that the library was highly enriched for genes involved in biomass degradation and secretion. The discrepancy between genes annotated to Gene Ontology terms and raw Blast analysis is an artifact of the annotation process due to the lack of widespread GO term assignment to sequences. This illustrates the need for work to create quality annotations on a reference set of fungal genes to enhance future annotation of novel genes as well as further biochemical analysis of these hypothetical genes.

In this study, homologues to a number of required GTPase encoding genes involved in protein trafficking were identified. A number of coatamer subunit genes were also located. The fact that these genes exist in *T. reesei* provides support for the results of previous ultra-structural comparisons of the secretion pathway in filamentous fungi. Not only does it appear, structurally, that the protein's secretory path is similar to yeast and higher eukaryotes but the genes involved in those pathways also seem to be homologous. A number of genes involved in the secretion and post-translational modification machinery were identified in this study. Other genes which are not well characterized but hint at being important in these pathways were also identified. These genes are all targets for expression level manipulation in the hopes of enhancing the secretion capacity. Identification of novel genes involved in the bottleneck pathways of folding and protein modification is the first step in the creation of higher producing strains of *T. reesei*.

Acknowledgements

This research was performed and supported by a collaborative effort between Genencor International and the Fungal Genomics Laboratory of North Carolina State University.

References

[1] Durand, H., Clanet, M. & Tiraby, G. (1988) Genetic improvement of *Trichoderma reesei* for large scale cellulase production, *Enzyme Microbiological Techniques. 10*, 341-345.

[2] Foreman, P., Brown, D., Dankmeyer, L., Dean, R., Diener, S., Dunn-Coleman, N., Meerman, H., Mitchell, T., Mitchinson, C., Olivares, H., Teunissen, P., Yao, J. & Ward, M. (2003) Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus *Trichoderma reesei*, *Journal of Biological Chemistry. 278*, 31988-31997.

[3] Gouka, R., Punt, P. & van den Hondel, C. (1997) Efficient production of secreted proteins by *Aspergillus*: progress, limitations, and prospects, *Applied Microbiology and Biotechnology. 47*, 1-11.

[4] Vorisek, J. (2000) Functional morphology of the secreted pathway organelles in yeast, *Microscopy Research and Technique. 51*, 530-546.

[5] Hemming, F. W. (1995) Expression and secretion of glycoproteins by hyphal fungi, *Biochem. Soc. Trans. 23*, 180-185.

[6] Novic, P. & Brennwald, P. (1993) Friends and family: The role of Rab GTPases in vesicular traffic, *Cell. 75*, 597-601.

[7] Lazar, T., Gotte, M. & Gallwitz, D. (1997) Vesicular transport: How many Ypt/Rab-GTPases make a eukaryotic cell?, *Trends Biochem. Sci. 22*, 468-472.

[8] Veldhuisen, G., Saloheimo, M., Fiers, M., Punt, P., Contreras, R., Penttila, M. & van den Hondel, C. (1997) Isolation and analysis of functional homologues of

the secretion-related SAR1 gene of *Saccharomyces cerevisiae* from *Aspergillus niger* and *Trichoderma reesei*, *Mol Gen Genet. 256*, 446-455.

[9] Salvouri, I., Markham, M., Rauvala, H., Knowles, J. & Kaariainnen, L. (1987) Low molecular weight high-mannose type glycans in a secreted protein of the filamentous fungus *Trichoderma reesei*, *Bio/Technology. 5*, 152-156.

[10] Kruszewska, J., Messner, R., Kubicek, C. & Palamarczyk, G. (1989) O-glycosylation of proteins by membrane fraction of *Trichoderma reesei* QM 9414, *J. Gen. Microbiology. 135*, 301-307.

[11] Palamarczyk, G., Maras, M., Contreras, R. & Kruszewska, J. (1998) Protein secretion and glycosylation in *Trichoderma* in *Trichoderma and Gliocladium* (C. P. Kubicek, G. E. H., ed pp. 121-138, Taylor & Francis Ltd., Padstow UK.

[12] Mantyla, A., Paloheimo, M. & Suominen, P. (1998) Industrial mutants and recombinant strains of *Tricoderma reesei* in *Trichoderma and Gliocladium* (G. E. Harman, C. P. K., ed pp. 291-304, Taylor & Francis, Bristol, PA.

[13] Kruszewska, J., Butterweck, A., Kurzatkowski, W., Migdalski, A., Kubicek, C. & Palamarczyk, G. (1999) Overexpression of the *Saccharomyces cerevisiae* mannosylphosphodolichol synthase-encoding gene in *Trichoderma reesei* results in an increased level of protein secretion and abnormal cell ultrastructure, *Applied and Environmental Microbiology. 65*, 2382-2387.

[14] Ilmen, P., Saloheimo, A., Onnela, M. & Pnettila, M. (1997) Regulation of cellulase expression in the filamentous fungus *Trichoderma reesei.*, *Appl Environ Microbiol. 63*, 1298-306.

[15] Davis, R. & deSerres, F. (1970) Genetic and microbiological research techniques for *Neurospora crassa.*, *Meth Enzymology. 17*, 79-143.

[16] Gordon, D., Abajian, C. & Green, P. (1998) Consed: a graphical tool for sequence finishing, *Genome Research. 8*, 195-202.

[17] Ewing, B. & Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Research. 8*, 186-194.

[18] Huang, X. & Madan, A. (1999) CAP3: A DNA sequence assembly program, *Genome Research. 9*, 868-877.

[19] Stajich, J., Block, D., Boulez, K., Brenner, S., Chervitz, S., Dagdigian, C., Fuellen, G., Gilbert, J., Korf, I., Lapp, H., Lehvaslaiho, H., Matsalla, C., Mungall, C., Osborne, B., Pocock, M., Schattner, P., Senger, M., Stein, L., Stupka, E., Wilkinson, M., & Birney, E. (2002) The Bioperl toolkit: Perl modules for the life sciences., *Genome Research. 12*, 1611-1318.

[20] Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G. & Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium, *Nature Genetics. 25*, 25-29.

[21] Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S., Pagni, M., Peyruc, D., Ponting, C.,

Selengut, J., Servant, F., Sigrist, C., Vaughan, R., and & Zdobnov, E. (2003) The InterPro Database, 2003 brings increased coverage and new features, *Nucleic Acids Res. 31*, 315-8.

[22] Altschul, S., Gish, W., Miller, W., Myers, E. & Lipman, D. (1990) Basic local alignment search tool., *J Mol Biol. 215*, 403-10.

[23] Xei, H., Wasserman, A., Levine, Z., Novik, A., Grebinskiy, V., Shoshan, A. & Mintz, L. (2002) Large-scale protein annotation through gene ontology, *Genome Research. 12*, 785-794.

[24] Nielson, H., Brunak, S. & von Heijne, G. (1999) Machine learning approaches for the prediction of signal peptides and other protein sorting signals, *Protein Engineering. 12*, 3-9.

[25] Galagan, J. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*, *Nature. 422*, 859-868.

[26] Vasara, T., Keranen, S., Penttila, M. & Saloheimo, M. (2002) Characterization of two 14-3-3 genes from *Trichoderma reesei*: interactions with yeast secretory pathway components, *Biochemica et Biophysica Acta. 1590*, 27-40.

Table 3.1 : Results of BlastX of the 20 most redundant clones from LT002 against nr protein database.

Sequence	# Clones	E-value	Description
Contig613	96	0	ELONGATION FACTOR 1-ALPHA [Trichoderma reesei]
Contig702	69	0	Exoglucanase I precursor (Exocellobiohydrolase I) (CBHI) (1,4-beta-cellobiohydrolase) [Trichoderma reesei]
Contig818	65	3E-69	probable V-ATPase, 20K chain [imported] – [Neurospora crassa]
Contig494	60	5E-08	rRNA intron-encoded homing endonuclease [Oryza sativa]
Contig163	51	6E-14	conserved hypothetical protein [Neurospora crassa]
Contig1010	48	-	no alignment
Contig981	47	-	no alignment
Contig948	40	0	QI74 protein [Trichoderma harzianum]
Contig1023	35	7E-21	Transcript Antisense to Ribosomal RNA; Tar1p [Saccharomyces cerevisiae]
Contig383	35	-	no alignment
Contig1201	34	1E-35	HYDROPHOBIN II PRECURSOR (HFBII) [Hypocrea jecorina]
Contig38	32	1E-64	Chain A, Structure Of L-Aminopeptidase D-Ala-EsteraseAMIDASE [Ochrobactrum Anthropi]
Contig744	32	4E-76	putative protein [Neurospora crassa]
Contig1166	31	5E-49	30 KD HEAT SHOCK PROTEIN [Neurospora crassa]
Contig260	31	4E-175	Glyceraldehyde 3-phosphate dehydrogenase 2 (GAPDH2) [Trichoderma koningii]
Contig572	31	-	no alignment
Contig1150	30	0	Actin, gamma [Neurospora crassa]
Contig386	30	-	No alignment
Contig445	30	-	no alignment
Contig1036	29	0	Exoglucanase II precursor (Exocellobiohydrolase II) (CBHII) (1,4-beta-cellobiohydrolase) [Trichoderma reesei]

Table 3.2 : Results of BlastX of the 20 most redundant clones from LT003 against nr protein database.

Sequence	# Clones	E-value	Description
Contig288	48	9E-132	ELONGATION FACTOR 1-ALPHA (EF-1-ALPHA) [Trichoderma reesei]
Contig352	42	3E-69	probable V-ATPase, 20K chain [imported] – [Neurospora crassa]
Contig329	37	1E-145	putative multiple drug resistance protein [Aspergillus fumigatus]
Contig174	29	-	no alignment
Contig269	21	0	QI74 protein [Trichoderma harzianum]
Contig482	21	6E-21	Transcript Antisense to Ribosomal RNA; Tar1p [Saccharomyces cerevisiae]
Contig460	19	4E-14	conserved hypothetical protein [Neurospora crassa]
Contig412	18	5E-137	Exoglucanase I precursor (Exocellobiohydrolase I) (CBHI) (1,4-beta-cellobiohydrolase) [Trichoderma reesei]
Contig673	18	0	ELONGATION FACTOR 2 (EF-2) [Neurospora crassa]
Contig659	17	1E-114	hypothetical protein [Neurospora crassa]
Contig184	16	-	no alignment
Contig147	15	0	heat shock protein 90 [Candida tropicalis]
Contig178	15	0	subtilisin-like serine protease PR1H [Metarhizium anisopliae var. acridum]
Contig59	15	1E-158	Glyceraldehyde 3-phosphate dehydrogenase 2 (GAPDH2) [Trichoderma koningii]
Contig160	14	5E-47	30 KD HEAT SHOCK PROTEIN [Neurospora crassa]
Contig395	14	3E-29	Required for invasion and pseudohyphae formation in response to nitrogen starvation; Muc1p [Saccharomyces cerevisiae]
Contig276	13	7E-74	putative protein [Neurospora crassa]
Contig492	13	0	AMINO-ACID PERMEASE INDA1 [Trichoderma harzianum]
Contig365	12	7E-75	hypothetical protein [Neurospora crassa]
Contig260	12	-	no alignment

Table 3.3 : Results of BlastX of the 20 most redundant clones from all sequence against nr protein database.

Sequence	# Clones	E-value	Description
Contig3156	153	0	ELONGATION FACTOR 1-ALPHA (EF-1-ALPHA) [Trichoderma reesei]
Contig1620	111	3E-69	probable V-ATPase, 20K chain [imported] [Neurospora crassa]
Contig3362	91	0	Exoglucanase I precursor (Exocellobiohydrolase I) (CBHI) (1,4-beta-cellobiohydrolase) [Trichoderma reesei]
Contig2867	77	1 E-18	similar to malonyl-CoA decarboxylase (EC 4.1.1.9) - goose [Rattus norvegicus]
Contig412	75	6E-14	conserved hypothetical protein [Neurospora crassa]
Contig1541	68	1E-145	putative multiple drug resistance protein [Aspergillus fumigatus]
Contig3747	68	0	QI74 protein [Trichoderma harzianum]
Contig2574	61	-	no hit
Contig3840	59	7E-21	Transcript Antisense to Ribosomal RNA; Tar1p [Saccharomyces cerevisiae]
Contig3772	59	-	no hit
Contig658	55	4E-175	Glyceraldehyde 3-phosphate dehydrogenase 2 (GAPDH2) [Trichoderma koningii]
Contig3775	52	-	no hit
Contig1363	49	4E-76	putative protein [Neurospora crassa]
Contig448	49	-	no hit
Contig1083	48	7E-49	30 KD HEAT SHOCK PROTEIN [Neurospora crassa]
Contig2651	48	-	no hit
Contig3771	47	0	subtilisin-like serine protease PR1H [Metarhizium anisopliae var. acridum]
Contig4042	45	-	no hit
Contig4361	44	1E-35	HYDROPHOBIN II PRECURSOR (HFBII) [Hypocrea jecorina]
Contig3675	43	5E-130	14-3-3-like protein [Hypocrea jecorina]

Table 3.4 : Genes involved in protein targeting and vesicle components in *T. reesei* identified using BlastX to nr protein and BlastN to GO_Genes database.

Vesicle Components

Sequence	Function	ID	Organism	Evalue
Contig1195	vesicle component (Erv25)	gi|11595612|emb|CAC18232.1|	*N. crassa*	3E-67
Contig2545	vesicle component (Erv14p)	gi|6321384|ref|NP_011461.1|	*S. cerevisiae*	3E-46
Contig4264	coatomer alpha subunit	gi|3170523|gb|AAC18088.1|	*A. nidulane*	9E-69
Contig1792	coatomer beta subunit	gi|11120716|ref|NP_068533.1|	*Rattus norvegicus*	4E-61
Contig629	Sec26p: COATOMER BETA SUBUNIT	gi|6320444|ref|NP_010524.1|	*S. cerevisiae*	7E-35
Contig2783	putative coatmer delta subunit	gi|19075836|ref|NP_588336.1|	*S. pombe*	8E-64
Contig1294	coatomer beta' subunit (beta' coat protein); secretory pathway	gi|19113604|ref|NP_596811.1|	*S. pombe*	5E-60

Targeting Components

Sequence	Function	ID	Organism	Evalue
Contig1945	probable membrane protein YBL106c	gi|6319362|ref|NP_009444.1|	*S. cerevisiae*	1E-28
Contig3848	gamma-actin vesicle transport	gi|14194449|sp|Q9UVW9|ACTG_CEPAC	*N. crassa*	0
Contig767	myosin I myoA	gi|1078633|pir||A56511	*A. nidulans*	5E-80
Contig2722	Vacuolar protein sorting-associated protein VPS4 (END13 protein)	gi|6325431|ref|NP_015499.1|	*S. cerevisiae*	5E-32
Contig3533	Vacuolar protein sorting-associated protein VPS4 (END13 protein)	gi|6325431|ref|NP_015499.1|	*S. cerevisiae*	3E-91
Contig35	RER1 protein	gi|6226763|sp|O15258|RER1_HUMAN	*H. sapiens*	4E-54
Contig4105	YIP3-like protein (Ypt interacting)	gi|14326101|gb|AAK60139.1|AF365406_1	*C. albicans*	2E-41
Contig1432	ADP-ribosylation factor (arf6) vesicle budding	gi|114128|sp|P26990|ARF6_CHICK	*Gallus gallus*	8E-78
Contig824	ArfGap GTP-ase activating protein	gi|19115755|ref|NP_594843.1|	*S. pombe*	2E-51
tric015xg04	ADP-ribosylation factor (Gcs1p)	gi|6319975|ref|NP_010055.1|	*S. cerevisiae*	3E-33

Fusion

Sequence	Function	ID	Organism	Evalue
Contig352	TRAPP complex 22Kd subunit	gi|6322921|ref|NP_012994.1|	*S. cerevisiae*	1E-53
Contig2782	Synaptobrevin-like V snare protein	gi|19113353|ref|NP_596561.1|	*S. pombe*	1E-65
Contig686	endoplasmic reticulum insertion protein SEC61	gi|28920222|gb|EAA29599.1|	*N. crassa*	3E-98

Table 3.5 : Genes involved in protein transport in *T. reesei* identified using BlastX to nr protein and BlastN to GO_Genes database.

S. cerevisiae Homologue	A. niger homologue	Transport Step	T. reesei Sequence	E-value
Sar1	SarA	ER to Golgi	Contig175	1E-24
RAB2	SrgD	ER to Golgi	n/p	
Ypt1	SrgB	ER to Golgi	Contig3974	1E-89
Ypt31	SrgE	Intra Golgi	Contig280	8E-65
Ypt32		Intra Golgi	Contig700	5E-72
Ypt6	SrgC	Golgi to vacuole	Contig3442	1E-59
Ypt7	SrgF	Late endosome to vacuole	n/p	
Ypt 52	n/a	Plasma membrane to endosome/ early to late endosome	Contig1672	1E-42
Sec4	SrgA	Golgi to plasma membrane	n/p	
Gdi1		Sec4 GDP dissociation inhibitor	Contig1224	3E-148
Vps1	VpsA	Vacuolar Biogenesis	Contig768/3709	1E-50

n/a – no homologues identified
n/p – not present in EST dataset

Table 3.6 : Genes involved in post-translational modification in *T. reesei* identified using BlastX to nr protein and BlastN to GO_Genes database.

DP-dependent glycosylation pathway

Sequence	**Function**	**ID**	**Organism**	**Evalue**
Contig610	mannose phospho-dolichol synthase	gi|11181548|gb|AAG32629.1|AF102883_1	*T. reesei*	7E-138
Contig570	Transfers mannose residues from dolichyl phosphate-D-mannose to specific serine/threonine residues of proteins in the secretory pathway; acts in complex with Pmt2p; Pmt1p	gi|6320107|ref|NP_010188.1|	*S. cerevisiae*	2E-107
tric035xj22.b1.ab1	probable dolichyl-phosphate-mannose-protein mannosyltransferase	gi|9368945|emb|CAB99175.1|	*N. crassa*	8E-103

Oligosaccharyl transferases

Sequence	**Function**	**ID**	**Organism**	**Evalue**
Contig1150	Oligosaccharyl transferase STT3 subunit	gi|6321416|ref|NP_011493.1|	*S. cerevisiae*	2E-81
Contig1274	Oligosaccharyl transferase	gi|11281521|pir||T43370	*S. pombe*	1E-97

Glycosyl transferase / modifiying

Sequence	**Function**	**ID**	**Organism**	**Evalue**
Contig2777	probable glycosyltransferase	gi|11289143|pir||T50453	*S. pombe*	2E-31
Contig4139	possible glycogen phosphorylase	gi|19577353|emb|CAD28434.1	*A. fumigatus*	4E-131
Contig889	possible glycogen phosphorylase	gi|19577353|emb|CAD28434.1	*A. fumigatus*	3E-40
Contig1774	probable mannosyltransferase	gi|19111915|ref|NP_595123.1|	*S. pombe*	1E-85
Contig2587	putative glycosyl transferase	gi|19075208|ref|NP_587708.1|	*S. pombe*	3E-48
Contig1595	mannosyltransferase : Ktr1p	gi|6324673|ref|NP_014742.1|	*S. cerevisiae*	3E-31

Chaperones and Folding

Sequence	**Function**	**ID**	**Organism**	**Evalue**
Contig2100 Contig2574 Contig1275	Hsp70 family : grp78, KAR2, BipA	gi|121575|sp|P16474|GR78_YEAST	*S. cerevisiae*	0
Contig1129	Protein disulfide isomerase	gi|129732|sp|P17967|PDI_YEAST	*S. cerevisiae*	1e-87
Contig3567	cypB – PPI	gi|118090|sp|P23284|PPIB_HUMAN	*H. sapien*	1e-58
Contig4323	cypC – PPI	gi|587533|emb|CAA86500.1|	*S. cerevisiae*	1e-48
tric033xk08.b1.ab1	FKBP13 – protein folding	gi|416992|sp|P32472|FKB2_YEAST	*S. cerevisiae*	1E-21
Contig4240	clxA – calnexin	gi|543920|sp|P27824|CALX_HUMAN	*H. sapien*	1e-51
Contig4136	HAC1 – UPR	gi|1175939|sp|P43567|YFD0_YEAST	*S. cerevisiae*	2e-50

Chapter 4

Insight into *Trichoderma reesei's* genome content, organization and evolution revealed through BAC library characterization

Stephen E. Diener[1], Malathi K. Chellappan[1], Thomas K. Mitchell[1], Nigel Dunn-Coleman[2], Michael Ward[2], Ralph A. Dean[1*]

1 – Fungal Genomics Laboratory, North Carolina State University, 840 Main Campus Drive, Suite 1200 Raleigh, NC 27606, USA

2 – Genencor International Inc. Palo Alto, Ca 94304, USA

* Corresponding Author

Ralph Dean
840 Main Campus Drive,
Suite 1200
Raleigh, NC 27606

Ph: 919-513-0022
Fax: 919-513-0024
Ralph_Dean@ncsu.edu

Keywords: Trichoderma reesei, bacterial artificial chromosome (BAC), physical map, genome organization, secretion, biomass degradation, repetitive element

Abstract

Trichoderma reesei is an important industrial fungus known for its ability to efficiently secrete large quantities of protein as well as its wide variety of biomass degrading enzymes. Past research on this fungus has primarily focused on extending its protein production capabilities, leaving the structure of its 33Mb genome essentially a mystery. To begin to address these deficiencies and further our knowledge of *T. reesei's* secretion and cellulolytic potential, we have created a genomic framework for this fungus. We constructed a BAC library containing 9,216 clones with an average insert size of 125 kb which provides a coverage of 28 genome equivalents. BAC ends were sequenced and annotated using publicly available software which identified a number of genes not seen in previously sequenced EST datasets. Little evidence was found for repetitive sequence in *T.* reesei with the exception of several copies of an element with similarity to the *Podospora anserina* transposon, PAT. Hybridization of thirty-four genes involved in biomass degradation revealed five groups of co-located genes in the genome. BAC clones were fingerprinted and analyzed using fingerprinted contigs (FPC) software resulting in 334 contigs covering 28 megabases of the genome. The assembly of these FPC contigs was verified by congruence with hybridization results.

Introduction

Numerous industrial processes utilize enzymes and specific metabolites which are produced by filamentous fungi. Among the various filamentous fungi, *Trichoderma reesei* has attracted much interest during the last few years due to the wide array of extracellular enzymes it natively secretes as well as its ability to produce heterologous proteins in high quantities (Palamarczyk et al. 1998). *T. reesei* is used commercially for production of cellulases and other enzymes which degrade complex polysaccharides. These enzymes are frequently used in the food, textile and paper industries for processing of raw materials (Biely et al. 1998). A large number of these enzymes have been previously identified and characterized amongst *T. reesei's* complement of genes (Foreman et al. 2003; Diener et al. 2004). The demand for identification of novel biomass degrading enzymes as well as for heterologous protein production at higher efficiency and reduced costs has catalyzed an interest in elucidating the genomic sequence of this fungus.

Over the past several years, sequencing technology has seen large advances in bioinformatics techniques, read length and parallelism. Despite these advances, finishing of a shotgun sequence assembly is a long and laborious process. Identifying the appropriate placement and orientation of each shotgun contig is simplified through the use of a comprehensive physical and genetic map. Physical map construction requires libraries containing large insert sizes at deep genome coverage. *E. coli* based large DNA cloning systems such as Bacterial Artificial Chromosomes (BACs) are particularly valuable for this purpose. Their stability and the ease with which they can be manipulated have made these cloning systems increasingly popular for creating libraries from bacterial, fungal, plant and animal genomes (Tao et al. 1994; Woo et al. 1994; Zhang et al. 1996; Shan et al. 2004). BAC libraries and physical maps have recently been constructed for a number of important fungal species, including *Magnaportha grisea, Aspergillus fumigatus* and *Phytophthora infestans* (Zhu et al. 1997; Randall et al. 1999; Zhu et al. 1999; Pain et al. 2004). In addition to providing a framework, or physical map, for assembly of a genome from whole genome shotgun sequencing, high quality BAC libraries are of great utility for map-based cloning or more directed BAC-by-BAC sequencing projects. BAC end sequencing, in conjunction with a BAC physical map provides a preview of gene content and genome organization.

In this paper, we describe the creation of a *T. reesei* strain QM6A large-insert BAC library and its subsequent analysis. We have assembled a physical map from fingerprinting which covers an estimated 28Mb of genomic sequence. Through hybridization of probes involved in biomass degradation, we have identified genomic clustering amongst these related genes. BAC end sequencing has revealed novel genes, shown evidence for past transposable element activity in the genome and provides markers for genomic assembly. The combination of these results creates a genomic framework for genome assembly and lasting resources for mapping of novel gene products in this valuable industrial fungus.

Materials and Methods

Fungal isolates and culture conditions

T. reesei strain QM6A (ATCC 13631) was used for constructing the BAC library. Fungal spores were grown on 250 ml PDA broth (Potato Dextrose agar) for 1-2 days at $30^{O}C$ at 200 rpm before preparing protoplasts and isolation of HMW DNA.

BAC Vector Preparation

pBeloBAC11 was used as the cloning vector for BAC library. Vector DNA was isolated using alkaline lysis followed by purification in a CsCl gradient. Digestion with *Hind*III was followed by dephosphorylation with Shrimp Alkaline Phosphatase (SAP). Enzymes were de-activated by heating the sample at 65˚C. Test ligations and transformations were performed to evaluate the quality of linearized, dephosphorylated vector DNA.

Isolation of high-molecular-weight DNA (HMW) from *T. reesei* protoplasts

Protoplasts were created following the protocol previously described with minor differences (Penttila et al. 1987). Mycelia were harvested through sterile mira-cloth and washed once with 20 ml of the re-suspension solution. The culture was re-suspended in Driselase solution (5mg/ml in re-suspension solution) and incubated at 28˚C at 250 rpm for 2-3 h. The suspension was filtered through sterile mira-cloth into a centrifuge tube and one volume of overlay solution was slowly added as a cushion on the top. Following centrifugation in a HS-4 rotor at 3079 g for 15 min at 4˚C, protoplasts were collected from the interface between the two phases as a cloudy layer and added to one volume of washing solution. Following centrifugation in a HS-4 rotor at 3079 g for 5 min at 4˚C, the pellet was washed once again with 4-6ml of washing solution. The final protoplast pellet was resuspended in a minimal volume of washing solution to a concentration of 1×10^8/ml. Protoplasts were embedded in 1% low-melting point agarose (Bio Rad, USA) microplugs and incubated in lysis buffer overnight at 37˚C. The plugs were then washed with 50mM EDTA, pH 8.0 and stored at 4˚C.

Hind III partial digestion and size selection of HMW DNA

Before partial digestion, agarose plugs containing high molecular weight (HMW) DNA were equilibrated in TE buffer for 6 h at room temperature and then at 4˚C overnight. Optimal partial digestion conditions were determined by digesting the chopped plugs with different enzyme concentrations (0.1 to 6.0 Units). Plugs were placed in eppendorf tubes containing 10x restriction buffer and *Hind*III enzyme and incubated on ice for 30 min. Digestion was performed at 37˚C for 30 min and the reaction was stopped by adding 20µl of 0.5 M EDTA, pH 8.0 on ice. Once the conditions were optimized, large-scale digestions were carried out.

Partially digested HMW DNA was separated by pulsed field gel electrophoresis (PFGE) in 1% agarose gel (CHEF-DR II drive module, Bio Rad) under the following conditions: 120˚C angle, 14˚C, 6V/cm, switch time 1-40s, run time 17-20 h, in 0.5 x TBE. After electrophoresis, the gel zone between 100 and 400 kbp was removed using glass cover-slip based on a MidRange II PFG Marker (New England Biolabs, Inc.) and divided into two horizontal sections. The gel slices were then subjected to a second size selection with the following conditions: 14˚C, 6V/cm, switch time 1-5s, run time 14 h, 0.5 x TBE. The gel region containing HMW DNA (100 to 400 Kbp) was excised and stored at 4˚C in TE.

Ligation and Transformation of size selected HMW DNA

HMW insert DNA was electro-eluted from the agarose gel slices as previously described (Strong et al. 1997). After electro elution, the DNA concentration was assayed on an agarose gel. Ligation was performed with 20ng vector and 200 ng insert DNA in a 120µl reaction volume at 16˚C overnight. The ligation mixture was transferred to a Millipore dialysis filter (Millipore #VSWPO1300, pore size 0.025µm), which was placed on 0.5 X TE in a petridish at 4˚C for 2 h to remove the ligation buffer. One microliter of dialyzed mixture was transformed in to *E. coli* DH10B competent cells (Gibco/BRL) using electroporation following standard procedures (Zhu et al. 1997). Transformed cells were immediately resuspended in 1 ml of SOC medium and incubated at 37˚C for 1 h with shaking at 200 rpm. Transformants were selected on LB supplemented with 12.5mg/ml chloramphenicol, X-GAL and IPTG and grown at 37˚C overnight. Recombinant white colonies were picked using a Genetix Q-bot into 384-well microtiter plates (Genetix) containing LB plus freezing media and 12.5mg/ml chloramphenicol. After overnight incubation, plates were stored at –80˚C.Three copies of the library were made and stored at –80˚C.

Isolation and Analysis of individual BAC clones

BAC DNA was isolated from 5 ml overnight cultures in LB with 12.5µg/ml chloramphenicol by alkaline lysis. The DNA pellet was resuspended in 40µl of TE. Ten µl DNA was digested with *Not*I and separated by CHEF gel electrophoresis (Bio Rad, USA). DNA was also digested with *Hind*III and separated on 0.7% agarose gel. The conditions for CHEF gel electrophoresis were: 1% agarose gel, 0.5X TBE, 14˚C, 6V/cm, switch time 5-15s, 120˚ angle, for 4 h.

BAC library Screening

BAC clones were arrayed in double spots using a 4X4 array on high-density nylon filters (22.2 cm^2) placed on LB containing 12.5mg/ml chloramphenicol using a Genetix Q-bot. This pattern allows 18,432 clones to be represented per filter, representing two copies of the twenty-four plate library double-spotted per filter. Colony filters were removed after overnight incubation at 37˚C and fixed using standard methods (Zhang et al. 1996). The filters were either used directly or stored at 4˚C. Filters were pre-hybridized in a hybridization bottle containing 20 ml of hybridization buffer (20X SSC, 0.5 M Na_2HPO_4, 100X Denhardts, 10% SDS) and 1 ml of denatured herring sperm DNA at 65˚C overnight. PCR primers were designed for genes of interest. For genes in which no EST had been identified, primers were designed from Genbank entries and PCR was performed with genomic DNA as the template. For genes which an EST had been identified, PCR was performed with the cDNA clone as template and using vector specific primers. The probes were then labeled with [^{32}P]-dCTP according to the manufacture's recommendation (Prime-a-Gene Labeling System, Promega). These were individually added to the pre-hybridization buffer and hybridization was carried out at 65˚C overnight with gentle shaking. Following hybridization, filters were washed twice with 1XSSC plus 0.1% SDS for 20 min each. After washing, filters were wrapped in saran wrap and exposed to a phosphor-screen for 1h followed by imaging.

Fingerprinting and BAC end sequencing

BAC clones were inoculated into 96-deep well plates containing selective medium. Following overnight incubation at 37˚C, DNA was isolated using the alkaline lysis method. For fingerprinting, one eighth of the template was digested with *Hind*III size separated by high resolution agarose gel electrophoresis, stained with SYBR gold and a digital image was captured using Kodak Digital Science (KDS1D, version 3.0.1). Images were processed and contigs were assembled using Image and fingerprinted contigs (FPC) software (Sanger

Center, UK) as described (Soderlund et al. 1997; Soderlund et al. 2000). The remaining template was divided in two and subjected to cycle sequencing reactions for both forward and reverse primers using Big-Dye terminator chemistry. The reactions were set up in 20ul volume (10ul of DNA, 3ul of Big Dye, 1.25ul of 10x Big Dye buffer, 1ul of primer, 4.75ul of deionized water).

Both ends of the BACs were sequenced primers using T7 short (5′-TAATACGACTCACTATAGGG-3′) as the forward primer and primer A (5′-TATGACCATGATTACGCCAAGC-3′) as the reverse primer. PCR was performed in a Gene Amp PCR system 9700 thermocycler (Applied Biosystems) using the following program (95˚C, 5 min; followed by 95˚C, 15 sec; 51˚C,10 sec; and 60˚C, 4 min for 75 cycles). Unincorporated dye terminators were removed following cleanup of cycle sequencing reactions using DyeEx 96 Kit (QIAGEN). Reaction products were analyzed using an ABI PRISM 3700 sequencer. The resulting chromatograms were run through the Alkahest Interceptor (www.alkahest.org) sequencing pipeline for quality and contaminant trimming and then assembled using CAP3 (Huang et al. 1999).

Test Statistic Description

In order to determine if the number of clones per contig in the assembly of BAC end sequences was distributed as expected for a random sampling of sequences we utilized a chi squared goodness-of-fit test statistic against the Poisson distribution.

The Poisson distribution is defined as follows

$$p(x;\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \quad x = 0,1,2...$$

Let λ be equivalent to the average number of sequences expected per bin. In our case λ is the number of quality sequence reads (r) divided by the number of possible sequence positions in the genome (n). Assuming a genome size of 33Mb and that there is a restriction site every 4096 bases there are approximately 8,056 sites. These sites can be sequenced in either direction which provides 16,113 possible sequence positions. The random variable, x, is equivalent to the number of times a particular sequence is seen. P(x;λ) equals the probability of seeing a particular sequence x times given that the average number of randomly selected sequences per possible sequence position is equal to λ (r/n). In order to test the observed data against the data predicted by the Poisson distribution, we performed a X^2 goodness-of-fit test.

$$\mathrm{X}^2 = \sum_{all} \frac{(observed - \exp ected)^2}{\exp ected}$$

The null hypothesis is that the distribution of clustered sequences follows the Poisson distribution and has no bias with the alternative hypothesis being that there is a bias in the sampling of sequences.

Rejection of null hypothesis occurs when $\mathrm{X}^2 \geq \mathrm{X}^2{}_{\alpha,k-1}$

Functional Annotation

Gene Ontology terms were assigned to the assembled dataset using merged annotations from BlastX to the Genbank nr protein database, BlastX to the Gene Ontology manually annotated protein sequence database and Interpro analysis as previously described in Diener et al. 2004 (Altschul et al. 1990; Mulder et al. 2003), These annotations were loaded into the Amigo browser to aid analysis (Ashburner et al. 2000; Harris et al. 2004).

Results

Construction of a *T. reesei* BAC Library

We have constructed a *Hind* III BAC library for an industrially important fungus *T. reesei* strain QM6A. HMW DNA was isolated from protoplasts. DNA was partially digested with *Hind* III and fragments ranging from 100 - 400 kb were size selected twice using PFGE, electro-eluted, ligated to vector DNA, dialyzed and transformed into *E. coli* competent cells (DH10B). Size selected HMW DNAs could be stored at 4˚C for up to a month. We also found that two size selections of partially digested HMW DNA were important in order to remove smaller DNA fragments. One microliter of ligated mixture produced about 150 recombinant clones. A library of 9,216 clones was generated and stored in twenty four 384-well plates.

Determination of the insert size of the BAC Library

The average insert size of the library was determined by digestion of 120 randomly selected BAC DNAs by *Not*I and fractionated by PFGE (Figure 4.1). The insert size of each clone was grouped into 10 kb intervals and plotted against the frequency of each group of clones represented in the library (Figure 4.2). Based on this analysis, the average insert size was estimated to be 125 kb with a range of 34-300 kb. Less than 4% of the clones did not contain either inserts or vector bands. The coverage of the library was estimated at 28 genome equivalents, assuming a haploid genome size of 33Mb (Carter et al. 1992).

BAC Library Screening

To determine the location and organization of genes of interest within the genome, the *T. reesei* BAC library was screened with 34 different probes derived from EST or Genbank sequences annotated as having a biomass degrading function. Probes were generated by PCR using vector specific primers or gene specific primers as described under Materials and Methods. Table 4.1 shows the Blast annotation of the probes used as well as results of library screening. Four hundred and seventy five positive BAC clones were identified with the range of 1-34 clones for each probe, averaging 14. All probes hybridized to at least one BAC clone in the library. In several instances, probes hybridized to an overlapping set of BAC clones. Probes were grouped into 5 clusters based on hybridization to two or more common BACs (Table 4.2). Eleven genes were included within these clusters ranging from two to four genes per cluster. For example, cluster one contained the cellulose transcription factors Ace1 and Ace2 and cluster four contained two beta-glucosidases, Cel1b and Cel3b (Takashima et al. 1999; Saloheimo et al. 2000; Aro et al. 2001; Foreman et al. 2003).

Fingerprinting and FPC assembly

A total of 6,144 BAC clones in the library were fingerprinted using *Hind*III enzyme. Following empirical evaluation, tolerance and cut-off values of 3 and 10^{-16} were set for contig assembly. Four thousand thirty seven (4,037) fingerprints were assembled into contigs after manual removal of poor quality fingerprints. FPC assembly returned 334 contigs, covering an estimated 28Mb of genome sequence. The largest contig contained 53 clones and had an estimated size of 428kb. Distribution of the number of BACs in contigs can be seen in Figure 4.3.

In order to verify the assembly of BACs by FPC, the BAC clones identified with the various probes were compared to the fingerprint analysis results. Expected results were that the majority of BACs identified by a single probe would be located in the same FPC contig. Thirty out of 34 probes had at least one identified BAC assigned to a contig. The remaining 4 probes did not identify any BACs located in contigs. Nine probes identified >5 BACs, ranging from 6-12, which were placed into FPC contigs. In 8 out of 9 of these cases, the majority of the BACs identified were located in the same FPC contig.

BAC end sequencing

BAC clones from seventeen 384 well plates from the library were subjected to end sequencing in both forward and reverse directions. Alkahest Interceptor exported 6,083 quality reads from the sequenced clones. All sequences reported here have been deposited in Genbank (Accessions CL529768-CL536556). The average trimmed read length of those sequences was 415 bases after removal of vector and low quality bases. The BAC ends were assembled by CAP3 into 1,200 contigs and 3,084 singlets producing a unique set of 4,284 sequences. The majority of contigs (818) contained only 2 clones, 256 contigs contained 3 clones, 73 contained 4 clones, and 52 contained between 5 and 9 clones. Contig consensus sequences and clone to contig mappings are available as supplementary data. Using a Poisson distribution with lambda set as the number of quality BAC end sequences divided by twice the number of possible restriction sites, the expected number of contigs containing five or more sequences is less than one. The chi squared goodness-of-fit value (4,738) was not significant which indicates a skew in the distribution of sequences.

Sequence Annotation

Assembled sequences were aligned against the Genbank nr protein database and the Gene Ontology (GO) Consortium GO_Genes protein database using the BlastX algorithm, limited to hits having better than 1e-5 expect values, resulting in a total of 1,323 (31%) and 655 (15%) hits respectively. The union set of sequences which had Blast hits to either Genbank nr or the GO_Genes database contained thirteen hundred and sixty three sequences. GO terms were assigned to 190 BAC end sequences using the alignments with nr protein and 157 using alignments to the GO_Genes database. InterProScan assigned GO terms to 395 unique proteins. Merging of these annotations produced 532 proteins with at least one GO term annotation. A compilation of major GO categories can be seen in Table 4.3 along with comparison to the GO annotation of an EST dataset using the same protocol (Diener et al. 2004). The major categories of biological process, cellular component, and molecular function contained 462, 314, and 499 sequences annotated to a child GO term within them. Comparison, category by category, of the annotation of an EST dataset to the annotation of BAC end sequences returned a relatively consistent BAC/EST annotation ratio close to 20% (Table 4.3) (Diener et al. 2004). A much lower percentage of BAC end sequences (12%) were annotated than EST sequences (39%) likely due to the less gene-directed nature of BAC construction. Furthermore, BlastN of the 4,284 assembled BAC end sequences against the assembled EST dataset at 1e-5 produced 919 (22%) alignments. From the 3365 BAC end sequences which did not align to the EST dataset, 335 had been annotated with GO terms. A further 532 BAC end sequences which were not found in the EST dataset matched to sequences in the Genbank nr protein database but were not assigned GO terms. Sequences in this subset were annotated with a wide variety of functions but most interesting were those identified as being involved with biomass degradation and secretion. We annotated a number of sequences as being related to these pathways. For example, BAC end sequences were found with homology to beta-mannosidase(Ctg508), beta-galactosidase(Ctg1273), xylanase regulator 1 (Ctg1448), farnesyltransferase(Ctg310), Sec7p(Ctg1199), Sec31(Ctg222) and Sec8(7E08.g2).

Repetetive DNA sequences such as transposable elements are commonly encountered in fungi and influence genome evolution (Daboussi 1997). To investigate the presence of such elements in *T.* reesei, a Blastn of all contigs and singlets against known fungal repetitive elements (Daboussi et al. 2003) returned 7 alignments spanning at least 250 bases and 60% identity. Five contigs and 2 singlets, containing a total of 19 sequences, were identified as having repetitive elements in them. All of these matches were to the *Podospora anserina* Fot1-like DNA transposon, Pat (Hamann et al. 2000). While several of these sequences aligned to overlapping regions of PAT, multiple sequence alignment of these BAC end sequences showed significant differences between them supporting that these elements are distinct entities.

Discussion

Creation of large-insert libraries is an important step in characterization of an organism's genomic content. BAC libraries are widely used as a method for physical analysis of genome structure and provide a tremendous wealth of information which can be exploited for gene discovery. Several fungal BAC libraries have been previously reported and have aided in characterization of their respective genomes (Zhu et al. 1997; Shan et al. 2004). Construction and analysis of this BAC library provides a framework for further genomic analysis and gene discovery in the filamentous fungus *T. reesei*. We present a library with an average insert size of 125 kbp which provides a total of 28 genome equivalents arrayed in 24 microtitre plates. The coverage of this library provides a high likelihood of representing the vast majority of the genome in a manageable size. The library's sequence redundancy and coverage is supported by the identification of at least one BAC clone from 34 independent hybridizations.

Fingerprinting of large insert BAC clones allows for the assembly of these BACs into a genome-wide physical map (Soderlund et al. 2000). These maps are critical for assembly of shotgun genomic sequence, positional cloning and understanding the relative organization of genes and markers. Out of 4,037 fingerprinted BAC clones, 334 BAC contigs were assembled using FPC. While the majority of these contigs contain fewer than 5 BACs, the larger contigs provide a quality framework for the assembly of genomic sequence while the smaller ones will help with ordering of local regions. In addition to fingerprinting the BAC clones, end sequencing was also carried out resulting in 4,284 unique sequences. These unique sequences provide a marker every 7.7Kb on average, calculated using an estimated genome size of 33Mb. The known distance between sequenced BAC ends are useful in the assembly of genomic contigs into supercontigs as shown in the assembly of *Neurospora crassa* (Galagan et al. 2003). The distribution of number of clones per contig was shown to be skewed toward higher numbers of clones per contig. This can be explained by restriction site hot spots over-representing certain regions of the genome in the BAC library or transposable elements causing over-representation of their sequence. The number of repetitive elements identified in the BAC end sequences is quite small when compared with other fungal species (Daboussi et al. 2003). The only evidence for transposable elements in the BAC end sequences were the 17 sequences with homology to the transposable element Pat from *Podospora anserina*. Only two contigs from the most redundant fifty-two contigs were contained in this set indicating that if the skew in clones per contig is caused by repetitive element activity, the elements found in *T. reesei* have not been previously identified. The majority of sequences in the top 52 contigs have no sequence similarity to the Genbank nr database or are only similar to hypothetical proteins from the *N. crassa* genome. Pat was shown to be a degenerate element inactivated by repeat induced point mutation (RIP) activity through base comparison to FOT1 (Hamann et al. 2000). Similar base content and transition-transversion ratio exist in the sequences found in the BAC end sequences which supports RIP-like activity having occurred in this fungus as well. RIP activity has been noted in a number of other filamentous fungi further supporting these results (Nakayashiki et al. 1999; Noubissi et al. 2000; Graia et al. 2001). The low number of transposable elements also supports an epigenetic silencing mechanism. The low number of repetitive sequences is fortunate in that it will reduce the number of ambiguous regions stemming from assembly of shotgun sequence.

These sequences were annotated with GO terms and compared to a set of similarly annotated EST sequences. The BAC end sequences were annotated to the same GO category at roughly 20% the frequency of the EST set. A similar ratio amongst the GO categories suggests that the BAC end sequences represent a widely distributed region of the genomic sequence. Over three hundred BAC end sequences were annotated with GO terms which did not have BLAST similarity to an EST sequence. The actual number of novel genes identified by sequencing BAC ends is likely higher due to the low percentage of the Genbank protein database annotated with GO terms. There are, in fact, 867 BAC end sequences which aligned to a sequence in the Genbank nr protein database but not to one of the EST sequences previously presented (Diener et al. 2004). The libraries used in that EST dataset were targeted for genes involved in biomass degradation and secretion. The EST dataset was highly enriched for the genes of interest yet, in our BAC end sequences, we were able to identify both biomass degrading and secretion related genes that were not previously seen in the EST dataset. In genomes containing high gene density, genome scanning techniques such as BAC end sequencing is an efficient method for gene discovery. It is of particular efficacy when searching for genes expressed in unknown conditions or low quantities. The end sequences we have produced enhance gene discovery and serve as markers for the construction of a high-resolution sequence map. The BAC end sequences, in conjunction with assembled FPC contigs provide the organization of genes across the entire genome as well as a wealth of information for the future assembly of genomic sequence.

Hybridization of 34 different probes derived from genes involved in biomass degradation to the BAC library produced four hundred and seventy-five positive BAC clones. We found that groups of these related genes hybridize to a similar set of BAC clones. It has been shown that genes involved in biomass degradation are co-expressed (Foreman et al. 2003). The co-location of these genes may provide a functional advantage for expression regulation. Duplication events of biomass degrading enzymes have been previously reported in other fungi (Harhangi et al. 2003). They showed evidence, in *Piromyces sp*., for expansion of a gene family through duplication resulting in genomic clusters of related cellulases in a head to tail orientation. The relationship between the genes within several of these clusters can be explained as a clustering of genes known to be co-regulated and/or an ancient gene duplication event. While it is possible that related genes had cross-hybridized, sequence identity between probes was not greater than 45 percent in any group significantly reducing the chances of cross-hybridization. These hybridizations also provide support for the assembly of BACs by FPC. Based on the comparative analysis between FPC contigs with the hybridization data, we found that contigs from FPC were in close agreement with co-hybridized clusters. While agreement of the hybridizations results with the FPC assembly was not universal, the discrepancies can be explained by the incomplete coverage of quality fingerprints.

These data reveal the utility of this BAC library for use in assembly of the *T. reesei* genome and isolation of genomic sequences of industrial interest. This BAC library provides a wealth of information for use in current and future studies of the *T. reesei* genome. This framework is particularly timely due to the current work in progress by the Joint Genome Institute aimed at sequencing the genome of *T. reesei* to a 10X shotgun coverage.

Acknowledgements

Construction and characterization of the *Trichoderma reesei* BAC library was performed at the Fungal Genomics Laboratory of North Carolina State University with financial support from Genencor International Inc. We would like to thank Jeff Thorne for his help with the test statistic.

References

Altschul, S., Gish, W., Miller, W., Myers, E., Lipman, D., 1990. Basic local alignment search tool. J Mol Biol. 215, 403-410.

Aro, N., Saloheimo, A., Ilmen, M., Penttila, M., 2001. ACEII, a novel transcriptional activator involved in regulation of cellulase and xylanase genes of *Trichoderma reesei*. J Biol Chem 276, 24309-24314.

Ashburner, M., Ball, C., Blake, J., Botstein, D., Butler, H., Cherry, J., Davis, A., Dolinski, K., Dwight, S., Eppig, J., Harris, M., Hill, D., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J., Richardson, J., Ringwald, M., Rubin, G., Sherlock, G., 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature Genetics 25, 25-29.

Biely, P., Tenkanen, M., 1998. In: Harman, G.E., Kubicek, C.P. (Ed.)^(Eds.), *Trichoderma and Gliocladium*: Enzymes, Biological Control and Commercial Applications, vol. 2. Taylor and Francis, London, pp. 25-47

Carter, G.L., Allison, D., Rey, M.W., Dunn-Coleman, N.S., 1992. Chromosomal and genetic-analysis of the electrophoretic karyotype of *Trichoderma reesei* - Mapping of the cellulase and xylanase genes. Molecular Microbiology 6, 2167-2174.

Daboussi, M.J., 1997. Fungal transposable elements and genome evolution. Genetica 100, 253-260.

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

Diener, S.E., Dunn-Coleman, N., Foreman, P., Houfek, T.D., Teunissen, P.J., van Solingen, P., Dankmeyer, L., Mitchell, T.K., Ward, M., Dean, R.A., 2004. Characterization of the protein processing and secretion pathways in a comprehensive set of expressed sequence tags from *Trichoderma reesei*. FEMS Microbiol Lett 230, 275-282.

Foreman, P., Brown, D., Dankmeyer, L., Dean, R., Diener, S., Dunn-Coleman, N., Meerman, H., Mitchell, T., Mitchinson, C., Olivares, H., Teunissen, P., Yao, J., Ward, M., 2003. Transcriptional regulation of biomass-degrading enzymes in the filamentous fungus *Trichoderma reesei*. Journal of Biological Chemistry 278, 31988-31997.

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., Jaffe, D., FitzHugh, W., Ma, L.J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C.B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M.A., Werner-Washburne, M., Selitrennikoff, C.P., Kinsey, J.A., Braun, E.L., Zelter, A., Schulte, U., Kothe, G.O., 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, R.L., Perkins, D.D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R.J., Osmani, S.A., DeSouza, C.P., Glass, L., Orbach, M.J., Berglund, J.A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D.O., Alex, L.A., Mannhaupt, G., Ebbole, D.J., Freitag, M., Paulsen, I., Sachs, M.S., Lander, E.S., Nusbaum, C., Birren, B., 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422, 859-868.

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.

Hamann, A., Feller, F., Osiewacz, H.D., 2000. The degenerate DNA transposon Pat and repeat-induced point mutation (RIP) in *Podospora anserina*. Mol Gen Genet 263, 1061-1069.

Harhangi, H.R., Akhmanova, A., Steenbakkers, P.J., Jetten, M.S., van der Drift, C., Op den Camp, H.J., 2003. Genomic DNA analysis of genes encoding (hemi-)cellulolytic enzymes of the anaerobic fungus *Piromyces sp*. E2. Gene 314, 73-80.

Harris, M.A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G.M., Blake, J.A., Bult, C., Dolan, M., Drabkin, H., Eppig, J.T., Hill, D.P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J.M., Christie, K.R., Costanzo, M.C., Dwight, S.S., Engel, S., Fisk, D.G., Hirschman, J.E., Hong, E.L., Nash, R.S., Sethuraman, A., Theesfeld, C.L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S.Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E.M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32 Database issue, D258-261.

Huang, X., Madan, A., 1999. CAP3: A DNA sequence assembly program. Genome Research 9, 868-877.

Mulder, N., Apweiler, R., Attwood, T., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., Bucher, P., Copley, R., Courcelle, E., Das, U., Durbin, R., Falquet, L., Fleischmann, W., Griffiths-Jones, S., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lopez, R., Letunic, I., Lonsdale, D., Silventoinen, V., Orchard, S., Pagni, M., Peyruc, D., Ponting, C., Selengut, J., Servant, F., Sigrist, C., Vaughan, R., and, Zdobnov, E., 2003. The InterPro Database, 2003 brings increased coverage and new features. Nucleic Acids Res 31, 315-318.

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

Noubissi, F.K., McCluskey, K., Kasbekar, D.P., 2000. Repeat-induced point mutation (RIP) in crosses with wild-isolated strains of *Neurospora crassa*: evidence for dominant reduction of RIP. Fungal Genet Biol 31, 91-97.

Pain, A., Woodward, J., Quail, M.A., Anderson, M.J., Clark, R., Collins, M., Fosker, N., Fraser, A., Harris, D., Larke, N., Murphy, L., Humphray, S., O'Neil, S., Pertea, M., Price, C., Rabbinowitsch, E., Rajandream, M.A., Salzberg, S., Saunders, D., Seeger, K., Sharp, S., Warren, T., Denning, D.W., Barrell, B., Hall, N., 2004. Insight into the genome of *Aspergillus fumigatus*: analysis of a 922 kb region encompassing the nitrate assimilation gene cluster. Fungal Genet Biol 41, 443-453.

Palamarczyk, G., Maras, M., Contreras, R., Kruszewska, J., 1998. Protein secretion and glycosylation in *Trichoderma*. In: C. P. Kubicek, G.E.H. (Eds.), Trichoderma and Gliocladium, vol. 1. Taylor & Francis Ltd., Padstow UK, pp. 121-138

Penttila, M., Nevalainen, H., Ratto, M., Salminen, E., Knowles, J., 1987. A Versatile Transformation System for the Cellulolytic Filamentous Fungus *Trichoderma reesei*. Gene 61, 155-164.

Randall, T.A., Judelson, H.S., 1999. Construction of a bacterial artificial chromosome library of *Phytophthora infestans* and transformation of clones into *P. infestans*. Fungal Genetics and Biology 28, 160-170.

Saloheimo, A., Aro, N., Ilmen, M., Penttila, M., 2000. Isolation of the ace1 gene encoding a Cys(2)-His(2) transcription factor involved in regulation of activity of the cellulase promoter cbh1 of *Trichoderma reesei*. J Biol Chem 275, 5817-5825.

Shan, W.X., Hardham, A.R., 2004. Construction of a bacterial artificial chromosome library, determination of genome size, and characterization of an Hsp70 gene family in *Phytophthora nicotianae*. Fungal Genetics and Biology 41, 369-380.

Soderlund, C., Humphray, S., Dunham, A., French, L., 2000. Contigs built with fingerprints, markers, and FPCV4.7. Genome Research 10, 1772-1787.

Soderlund, C., Longden, I., Mott, R., 1997. FPC: a system for building contigs from restriction fingerprinted clones. Computer Applications in the Biosciences 13, 523-535.

Strong, S.J., Ohta, Y., Litman, G.W., Amemiya, C.T., 1997. Marked improvement of PAC and BAC cloning is achieved using electroelution of pulsed-field gel-separated partial digests of genomic DNA. Nucleic Acids Res 25, 3959-3961.

Takashima, S., Nakamura, A., Hidaka, M., Masaki, H., Uozumi, T., 1999. Molecular cloning and expression of the novel fungal beta-glucosidase genes from *Humicola grisea* and *Trichoderma reesei*. J Biochem (Tokyo) 125, 728-736.

Tao, Q., Zhao, H., Qui, L., Hong, G., 1994. Construction of a full Bacterial Artificial Chromosome (BAC) library of *Oryza sativa* genome. Cell Research 4, 127-133.

Woo, S.S., Jiang, J.M., Gill, B.S., Paterson, A.H., Wing, R.A., 1994. Construction and Characterization of a Bacterial Artificial Chromosome Library of *Sorghum bicolor*. Nucleic Acids Research 22, 4922-4931.

Zhang, H.B., Choi, S.D., Woo, S.S., Li, Z.K., Wing, R.A., 1996. Construction and characterization of two rice bacterial artificial chromosome libraries from the parents of a permanent recombinant inbred mapping population. Molecular Breeding 2, 11-24.

Zhu, H., Blackmon, B.P., Sasinowski, M., Dean, R.A., 1999. Physical map and organization of chromosome 7 in the rice blast fungus, *Magnaporthe grisea*. Genome Research 9, 739-750.

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

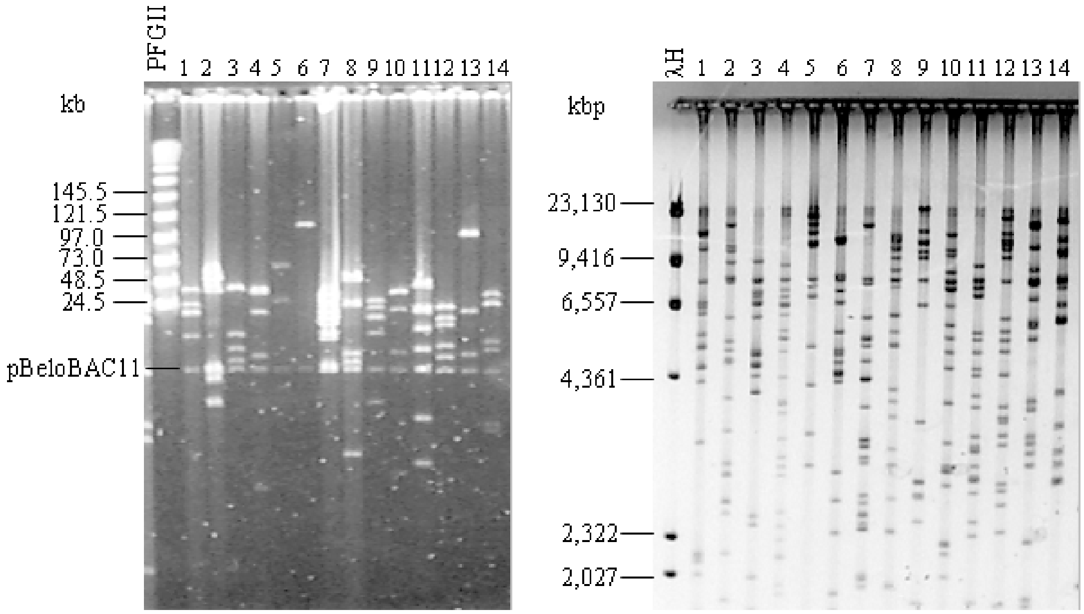
Figure Legends

Figure 4.1.
Restriction analysis of randomly selected clones from the *T. reesei* BAC library. A. *Not*I-digested BAC clones analyzed by pulsed-field gel electrophoresis (PFGE). B. *Hind*III-digested BAC clones separated by standard gel electrophoresis. The 7.4-kb common band is from the vector, pBeloBAC11. PFGII; Midrange II (NEB) molecular-weight marker.

Figure 4.2.
Insert size distribution of 120 randomly selected clones from *T.reesei Hind*III BAC library.

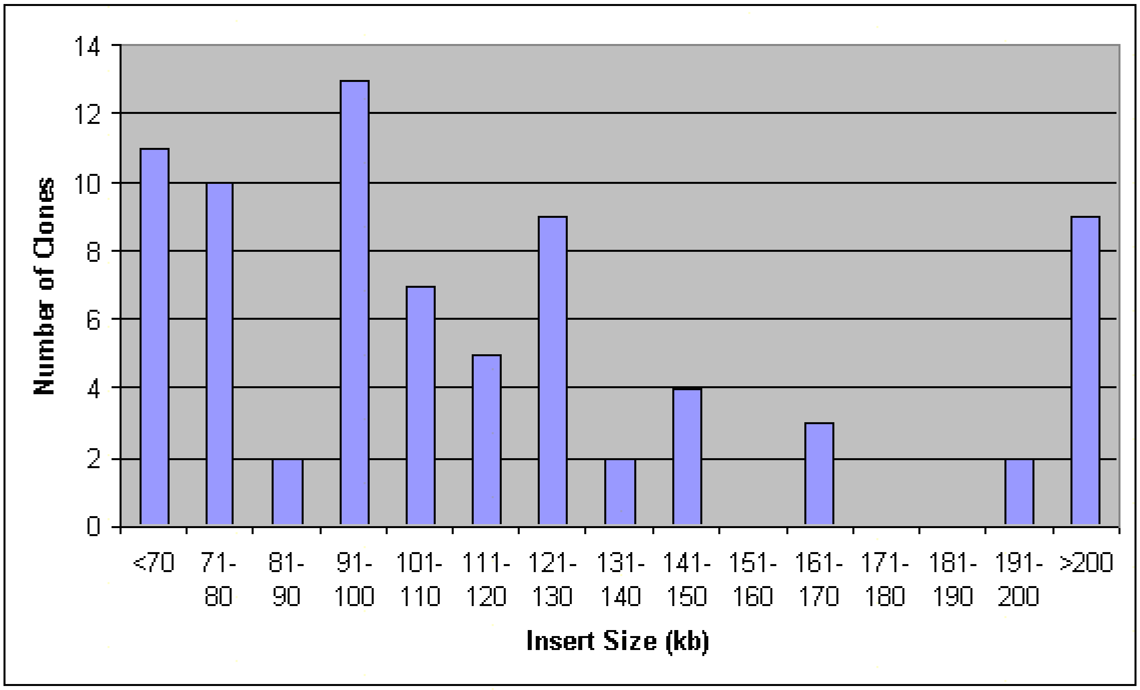
Figure 4.3.
Distribution of the number of BACs contained within assembled FPC contigs

Figure 4.1:



PFGII
1 2 3 4 5 6 7 8 9 10 11 12 13 14
kb
145.5
121.5
97.0
73.0
48.5
24.5
pBeloBAC11
λH
1 2 3 4 5 6 7 8 9 10 11 12 13 14
kbp
23,130
9,416
6,557
4,361
2,322
2,027

Figure 4.2



14
12
10
8
6
4
2
0
Number of Clones
<70
71-80
81-90
91-100
101-110
111-120
121-130
131-140
141-150
151-160
161-170
171-180
181-190
191-200
>200
Insert Size (kb)

Figure 4.3

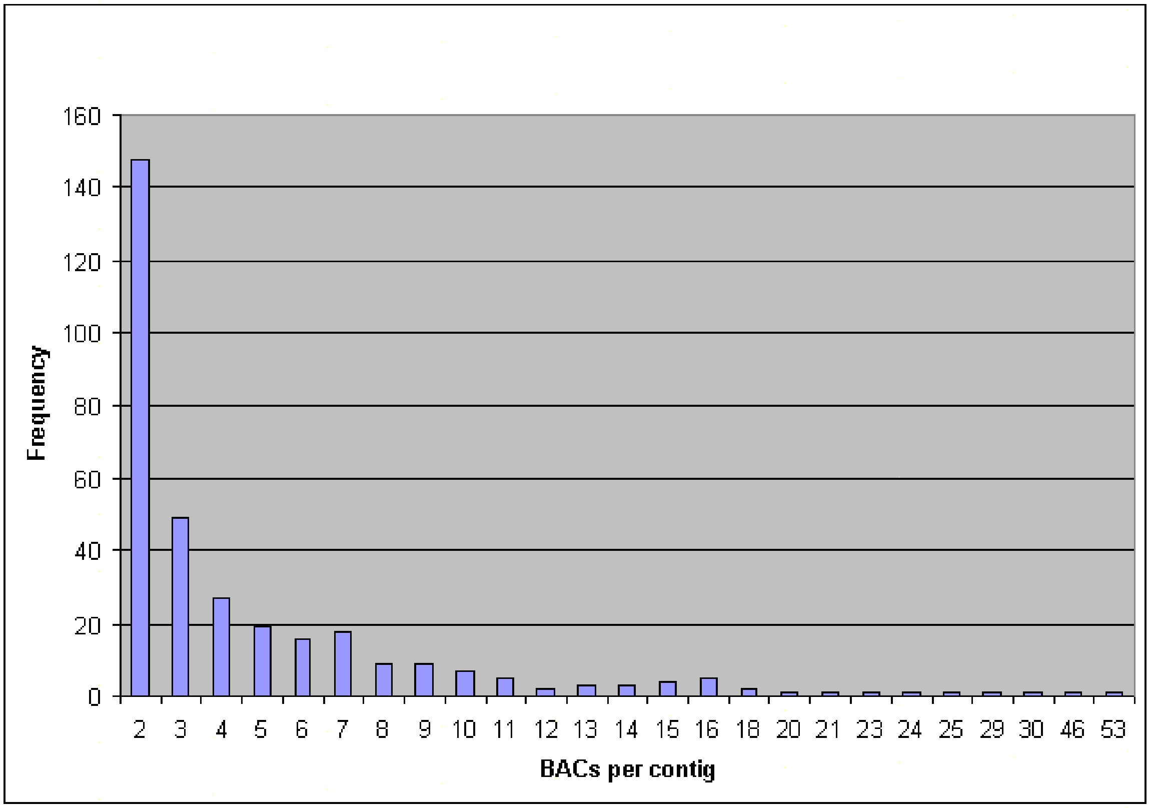

160
140
120
100
80
60
40
20
0
Frequency
2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 18 20 21 23 24 25 29 30 46 53
BACs per contig

Table 4.1. Identification of BACs containing genes involved in biomass degradation.

Probe[1]		Gene Name – Function	Number Hits[2]
Abf1 (CF866662)		Arabinofuranosidase I	8
Ace1 (AF190793)	*	Cellulase transcriptional regulator	20
Ace2 (AF220671)	*	Cellulase transcriptional regulator	19
Agl1 (Z69253)	*	Alpha-galactosidase 1	16
Agl2 (Z69254)	*	Alpha-galactosidase 2	18
Agl3 (CF876376)		Alpha-galactosidase 3	15
Axe1 (CF865918)		Acetyl xylan esterase 1	34
Bgl1 (CF869264)		Beta-glucosidase 1	7
Bgl2 (CF869916)		Beta-glucosidase 2	19
Cel3b (AY281374)		Beta-glucosidase 3	15
Cel3c (AY281375)		Beta-glucosidase 4	26
Cel1b (AY281377)		Beta-glucosidase 5	8
Bxl1 (CF871285))		Beta-xylosidase 1	6
Cbh1/cel7a (CF869152)		Cellobiohydrolase I	5
Cbh2/cel6a (CF867172)		Cellobiohydrolase II	1
CreA (CF875780)		Carbon catabolite repression	9
Der1 (CF865806)		ER degradation	7
Egl1/cel7b (CF874531)		Endoglucanase I	14
Egl2/cel5a (CF868148)		Endoglucanase II	6
Egl4/cel6a (CF873473)		Endoglucanase IV	17
Egl5/cel45a (CF870056)		Endoglucanase V	18
Cel61b (AY281372)		Endoglucanase VII	21
Cel5b (AY281373)		Endoglucanase VIII	1
Hap2 (AF283501)	*	CCAAT-binding factor	34
Hap3 (CF868163)		CCAAT-binding factor	14
Cip1 (AY281370)		Cellulose induced protein	6
Man1 (CF871143)		endo-beta-mannanase	17
Nit3 (CF865713)		Nitrate reductase	2
Nir1 (CF871961)		Nitrate assimilation regulatory protein	11
Snf4 (CF874017)		Glucose derepression	16
Ubc13 (CF874652)		Ubiquitin-conjugating enzyme	16
Xyn1 (CF867983)		Xylanase I	18
Xyn2 (CF865758)		Xylanase II	10
Xyn3 (CF865613)		Xylanase III	21

1 – Identity and Genbank accession of EST probe used to hybridize to BAC library as identified by Blast analysis to nr database
2 – number of positively hybridized BAC clones
* – Probe derived from PCR of *T. reesei* genomic DNA using primers designed from Genbank sequence

Table 4.2. Grouping of BACs identified as containing genes involved in biomass degradation.

Grouping	Probe	BACs shared by all probes	Unique BACs
1	Ace1	01E10, 11O14, 12G22, 12O14	03G13,03P23,04L02,04N05, 04C14,05C19,05A19,05N24, 08O14,09N14,10C21, 10M12,10J14,11G22, 12G22,12O14,14K16, 15G4,16L7,20K4
	Ace2		02L02,02N05,02C14,04M12, 04J14,04C21,06L07,10A19, 10C19,10N24, 12G22,12O14,13K16, 15K4,17G15,19N14, 21G04
2	Xyn3	1N10,3C05,3C17,3F10,4C13,4C14, 4P12,5C10,9E12,10O14,11B24,12I09, 13P18,14D16,18N18,19H24,20A06, 20K02, 21L11	13B12
	Cel5b		10K06
3	Egl4	2G19,3M12,4L15,4L16,5B20,8F17,	02I18,11L22,11P12,11K09, 12N05,13A17,13L17,13O10 18J05,19G09,20G23
	Cip1		
4	Cel3b	1N07,3C18,4I23,4I24,4O03,5A21,5H03,	09E14,11H01,13D16,19K5, 19K18,20I16,21G15, 21M17
	Cel1b		03K19
5	Egl2	12N19, 3I15,4E08,11O4, 19H8	05D02*
	Bgl1		03D24
	Bx11		05D02*

* - shared amongst subset of cluster

Table 4.3. GO annotation of BAC end sequences and EST data

	BAC end	EST[1]	Percent[2]
Total Sequences	4284	6943	
Gene Ontology	532	2732	19.5
Percentage	12.4	39.3	
-- biological process	462	2428	19.0
---- cellular process	171	853	20.0
------ cell growth/maintenance	150	777	19.3
------ cell cycle	28	154	18.2
------ transport	93	479	19.4
---- development	8	79	10.1
---- physiological processes	438	2299	19.1
------ death	0	9	0.0
------ stress response	33	75	44.0
------ metabolism	353	1906	18.5
---- biological process unknown	14	154	9.1
-- cellular component	314	1972	15.9
---- cell	292	1817	16.1
------ external encapsulating structure	0	12	0.0
---- extracellular	7	74	9.5
---- unlocalized	8	55	14.5
---- cellular comp. Unknown	12	128	9.4
-- molecular function	499	2544	19.6
---- binding activity	202	1309	15.4
------ nucleic acid binding	96	559	17.2
---- cell adhesion molecule	0	8	0.0
---- catalytic activity	275	1344	20.5
---- chaperone	7	44	15.9
---- defense activity	1	5	20.0
---- enzyme activity regulator	8	24	33.3
---- molecular function unknown	12	120	10.0
---- motor activity	6	13	46.2
---- signal transducer activity	23	93	24.7
---- structural molecule activity	19	133	14.3
---- toxin activity	1	7	14.3
---- transcription regulator	15	104	14.4
---- transporter activity	112	493	22.7

1 – Data from Diener et al., 2004
2 – Percentage of annotated sequences in a GO category comparing EST annotation to BAC end sequence annotation

Chapter 5

Ancestral history of transposable elements within the genome of *M. grisea* inferred using gene genealogies and the coalescent

Diener S.E., Carbone I., Thon M., Dean R.A.

Abstract

The recent sequencing and annotation of *M. grisea* has uncovered the complete set of transposable elements within the genome. The geneaological and coalescent histories of these transposable elements can provide insight into the interplay between genome and mobile element. We analyzed five different transposable elements which are present in the genome in significant numbers using gene geneaologies and the coalescent. This analysis expanded traditional population genetics techniques into the realm of self-replicating genetic elements within a genome. We identified blocks of recombination within several of the transposable elements and were able to run coalescent analysis on each transposable element. The results of the coalescent analysis suggest that a major genomic event occurred in the past which allowed for the rapid expansion of all transposable elements studied. An event such as the loss of RIP either directly through mutation or via a shift in the primary mode of reproduction to asexual, thus bypassing the pre-meiotic stage necessary for RIP activity, could have lead to such an expansion. The loss of genomic control of these transposable elements could represent an acceptance of the action of transposable elements as beneficial to the fungus as active transposable elements can assist in the rapid evolution required for the fungus to retain its pathogenicity.

Introduction

Magnaporthe grisea, the filamentous fungus that causes rice blast disease, has recently had its 40 Mb genome sequenced and annotated. Inspection of the genome has shown that a sconsiderable percentage, 9.7, is composed of repetitive elements. Three major classes of transposons were identified within the repetitive elements. The inverted repeat DNA transposons consisting of Pot2/3/4 and Occan make up approximately 2% of the genomic sequence. The class of LTR retrotransposons consisting of MAGGY, MGLR-3, Pyret1/2/3 and RETRO5/6/7 is responsible for 3.8% of the genomic DNA. Roughly 1.6% of the genomic sequence is made up by elements in the non-LTR transposon class (LINEs and SINEs), represented by MGL, Mg-SineA and REPBUF [1] The role that these transposable elements play in evolution of *M. grisea's* genome has not been thoroughly explored. There is evidence for TE's providing homologous regions for intra-genomic recombination of large sections of chromosomes[2]. Within *M. grisea,* clusters of transposable elements have been linked to areas of high recombination and are also implicated in gene duplication [Mike's Paper]. A number of *M. grisea's* host specificity genes are positioned in areas of high transposon density which increases the likelihood that these genes will be lost or modified in a transposon mediated recombination event [3, 4].

In *M. grisea*, there is evidence of a transposable element, Pot3, modifying an avirulence gene's transcriptional profile through insertion into the promoter region. The change in expression induced virulence on a previously resistant strain of rice [5]. The evolutionary implications of high transposable element activity in a genome are great. Mutational effects of transposable elements are not limited to gene or promoter disruptions; accumulation of these elements can lead to drastic genomic changes such as chromosome length polymorphisms[2]. The fact that, in nature, rice pathogenic strains typically propagate asexually, bypassing meiotic quality control mechanisms, increases the chances that such radical mutations become fixed in the population as long as they do not negatively affect vegetative fitness. The bypass of meiosis allows evolution to adapt the fungus to changing environments or host responses much more quickly than if it was bound by the limitations of sexual reproduction.

The effect that transposable elements have on the genome they occupy has historically been assumed to be detrimental. However, recently it has been suggested that the symbiotic relationship between the genome and transposable elements may be closer to mutualism than parasitism [6]. Evolution is driven by mutation, and transposable elements provide that to the genome. In ecologies where the genomic host is under constant selective pressure, the mutational potential that active transposable elements provide can be very advantageous. The evolutionary history of a host-parasite interaction, such as between *M. grisea* and its host, rice, contains examples of a host's evolution of resistance and the pathogen's subsequent overcoming of that resistance. The selective pressures in this situation are enormous; essentially both host and pathogen must evolve or become extinct. In this type of highly stressful situation, any trait which increases the ability of a species to rapidly evolve would be tolerated or even selected for. It can be argued that these repetitive elements play a large role in the continued pathogenicity and rapid adaptability of this fungus. An interesting question is whether these elements are, in part, responsible for the genesis of *M. grisea's* pathogenicity. In order to understand the effects of transposable elements on the genome of *M. grisea*, it is necessary to decipher their history within the genome.

Methods using gene genealogies and the coalescent are well described for reconstructing the history and movement of genetic material throughout a population [7]. In this study, we use coalescent methods to analyze populations of transposable element within the genome of *M. grisea*. Understanding the history of transposable elements within a genome can provide insight into how these elements interact with the genome and how they are involved in the generation or maintenance of pathogenicity. Through the use of these population based methods, we explore the history of transposable elements within the genome of *M. grisea*.

Materials and Methods

Sequence acquisition and annotation

An assembly of the genomic shotgun sequence (Ver 2.3) was acquired from the Broad Institute's *M. grisea* database (http://www.broad.mit.edu/annotation/fungi/magnaporthe/). All copies of repetitive elements of interest were identified and extracted from the genome sequence using RepeatMasker (A. Smit, unpublished) and repetitive elements sequence compiled from Genbank. Using files made available by the Broad Institute, the chromosomal location of each extracted repetitive sequence was noted.

Transposable element analysis

Sequence Preparation

Analysis of these elements followed a protocol similar to that described in Carbone et al. [7]. SNAP workbench was used to facilitate the analysis through data management and automation of subsections in the protocol. The following description is an overview of the protocol followed for each transposable element. The specific steps followed for each element's analysis will be completely described the results section. Sequences identified using RepeatMasker were aligned with alike elements using CLUSTALW and the resulting alignment was manually inspected and modified [8]. These alignments were collapsed into unique haplotypes with recoding of indels using SNAP Map. Recoding of indels takes advantage of all available variation to ensure that low frequency variants were not mistakenly excluded in the calculation of test statistics and subsequent coalescent analysis. Base substitutions were categorized based on whether they are phylogenetically informative, the type of substitution whether it was a transition or a transversion using SNAP Map and SITES. A compatibility matrix was generated from the collapsed haplotypes using SNAP Clade and graphically illustrated using SNAP Matrix. Incompatibilities in the data manifested in the haplotypes' compatibility matrices were resolved in a number of ways dependent on what patterns the data presented. In some cases, an obvious block of incompatibility, likely corresponding to a recombination event, presented itself and a sub-region of the data consistent within itself was used for data analysis. In other cases due to a number of factors it was not as easy to identify a compatible region for subsequent analysis. In those cases it was necessary to analyze the sequences clade by clade to identify the largest clade which is most compatible. The unweighted parsimony method in Phylip was used to create the most parsimonious trees [9]. The strict consensus of the returned trees was used as the basis for the clade by clade analysis of incompatible sequences. As a last resort, a site or at most two were removed from the sequences to ensure the compatibility of the dataset. These sites were usually incompatible with only two to three other sites and this incompatibility could be attributed to the effect of repeat induced point mutations (RIP) or a sequencing error.

Tests of neutrality and Subdivision

For each transposable element, we estimated the population mutation rate, Θ, from the number of segregating sites [10], and performed tests of neutrality using Fu's website at (http://hgc.sph.uth.tmc.edu/fu/genealogy/test2/welcome.html). These tests provide population parameter estimates and allow for inference on the population growth patterns through comparison of the results and significance of the various *a priori* estimates [11, 12]. In the event that there is subdivision in the data, these tests of neutrality lose power in identifying the evolutionary processes occurring. Therefore, an alternative test for population subdivision should be used to prevent false positives. Hudson's test for population differentiation is historically used to determine whether geographically distinct populations show genetic subdivision. In this study, transposable elements were tested for geographic subdivision based on their chromosomal location to determine if there is any preference for local over global transposition in this fungus. The collapsed sequences were converted to a distance matrix using Seqtomatrix and tested using Hudson's Permtest to identify any chromosomal subdivision in the data [13]. Hudson's test uses a Monte Carlo simulation of a non-parametric permutation method to estimate test statistics and determine if the dataset shows significant evidence for population subdivision. Identification of subdivision is important for later coalescent work as the proper choice of model, either panmictic or subdivided, in the Genetree software is essential for appropriate analysis which does not violate the model's assumptions.

Coalescent Analysis

Genetree was used for the coalescent analysis and to generate all potential trees for the sequences and up to 10 million runs were used to determine the most likely tree from this set as compared to the unrooted tree [14]. After identification of the most likely tree ten million runs were used to generate ages of the mutations on the chosen tree. The most likely coalescent tree with predicted ages of mutations was then graphically represented by SNAP Workbench.

Results

Sequence acquisition

RepeatMasker identified and extracted all representatives of transposable elements from *M. grisea's* genome using a database of known elements from this species. Table 5.1 shows the count and chromosomal distribution of elements analyzed in this study. The numbers shown in the table are not the element copy number in the genome. Most of the representatives used are partial copies; there are few instances of completely assembled elements within the shotgun genome assembly. Elements used in this study were chosen from the complete set because their high numbers of representatives made it more likely that there would be an aligned region amenable to analysis. Pot2 and Pot 4 were extracted as a single set but were manually separated into distinct entities after obtaining an acceptable multiple sequence alignment of the Pot2/4 dataset proved impossible. Multiple sequence alignment of the combined set as well as phylogenetic analysis clearly distinguished the two elements making the separation trivial.

Tests of neutrality and subdivision

A number of tests of neutrality assuming constant population size, lack of recombination and lack of migration [11, 12, 15] were performed on each transposable element population after the final dataset had been determined [Table 5.2]. The results of these tests can be interpreted to provide *a priori* inferences into the population processes occurring in the datasets. A test returning significant results may suggest any of the following; a departure from neutrality, population subdivision, population growth, or background selection [11, 16, 17]. Through the comparison of Tajima's D, Fu and Li's D* and F*, and Li's Fs, the processes of population growth and background selection can be distinguished. When only Fu and Li's F* and D* are significant, background selection is suggested. When Fu's Fs is significant, population growth or genetic hitchhiking is supported.

As can be seen in Table 5.2, in all of the tested transposable elements, Tajima's D was significant, indicating a deviation from neutrality caused by either sudden population expansion or by selection of a specific allele over others. Since Fu's Fs is also significant, highly significant in all but the P1 region of Maggy, the deviation from neutrality supports one of two processes, either population growth or genetic hitchhiking. Tests for subdivision in all transposable elements were not significant, allowing for the use of the panmictic model for coalescent analysis.

Maggy

The sequences identified by RepeatMasker as being full length copies of the MAGGY element were aligned using CLUSTALW. This alignment was collapsed into unique haplotypes with recoding of indels and categorization of base substitution using SNAP Map and SITES. A compatibility matrix was generated from the collapsed haplotypes using SNAP Clade and graphically illustrated using SNAP Matrix. This matrix represents the support or conflict among the variable sites in the genetic material visually and can be used to easily identify regions of genetic material undergoing reticulate evolution [18].

This matrix was used to identify compatible regions for further analysis using the partial elements available in the genomic sequence.

All elements identified as MAGGY were aligned using CLUSTALW and the resulting alignment was manually edited. The two regions identified for further analysis by the prior full-length copy analysis were extracted from the complete alignment using SNAP Workbench. Protein domains within the MAGGY element were identified and annotated using Iprscan on version 9 of the Interpro database [19, 20].

Of the total 229 sequences identified within the genome as having homology to the MAGGY element, 6 were complete copies of the transposon. A histogram showing the representation of regions of MAGGY is shown in Figure 5.1. The lack of full length representatives of the element requires identification of regions within the transposon to be analyzed fully. The full length sequences were aligned using CLUSTALW and the resulting distribution of haplotypes, base substitution and insertion/deletion events is show in Figure 5.2. Compatibility analysis of this dataset shows a single region of incompatibility from 3416 to 3918 within the MAGGY sequence [Figure 5.3]. According to Interpro analysis, this region encompases the majority of the transferase domain of Maggy's Pol gene, identified from base pairs 3348 to 3733. The compatibility analysis suggests that the region containing the transferase domain has an alternate genetic history from the rest of the element. This region, designated D1, and a 5' region in the Pol gene, from base 1770 to 2145 within the element, designated P1, were chosen for further analysis since they represent both distinct histories identified in the MAGGY element.

The 5' region, P1, was extracted from a multiple sequence alignment of all MAGGY sequence within the genome. The P1 region had 24 representatives spanning an alignment of 384 nucleotides. The distribution of haplotypes, base substitution and recoded insertion/deletion events resulting from collapse of the sequences can be seen in Figure 5.4. The compatibility matrix in Figure 5.5 shows that there is a single site causing incompatibility between the haplotypes. This site, 17, was removed from the alignment and resolved the compatibility issues.

Potential trees were created and 1 million runs with a theta value of 8.03 predicted the most likely tree numbered 10 out of 30 possibilities with a likelihood of 9.1350e-16 which was closest to the unrooted likelihood of 1.5301e-15. The ages of mutations were generated also using 10 million runs and a graphical representative of the coalescent tree can be seen in Figure 5.6. The coalescent tree shows a slow increase in haplotypes followed by a large expansion in the number of haplotypes after the occurrence of mutation 29.

The transferase region, D1, was extracted from the multiple sequence alignment of all MAGGY elements. This region had 30 representatives whose alignment was 623 nucleotides long. The distribution of haplotypes, base substitution and recoded insertion/deletion events resulting from collapse of the sequences can be seen in Figure 5.7. The compatibility matrix in Figure 5.8 shows many conflicting sites which are not easily segregated into a block. A look at the phylip generated parsimony tree [Figure 5.9] shows that haplotype H2 is much more diverged than the remaining haplotypes. Inspection of the base substitutions in Figure 5.7 shows that H2 contains many more substitutions than an average haplotype and that the majority of substitutions are transitions. The action of RIP leads to numerous transition events in a repeated genomic sequence. The pattern of substitution seen in the two haplotypes is consistent with a RIPped sequence and as such we have excluded the sequences from those haplotypes from our analysis. The exclusion of this haplotypes caused the dataset to be completely compatible.

Hudson's test for subdivision in the data showed no evidence for subdivision which allows for the use of the panmictic coalescent model. Potential trees were created and 1 million runs using a theta value of 8.08 predicted the most likely tree numbered 7 out of 32 possibilities with a likelihood of 4.8921e-21 which was closest to the unrooted likelihood of 1.8511e-20. The ages of mutations were generated also using 10 million runs and a graphical representative of the coalescent tree can be seen in Figure 5.10. The coalescent tree of D1 shows a different topology than the P1 region. Instead of starting out with relatively few lineages and having a major expansion of a from a single haplotypes, there is a much larger number of lineages and there are several smaller expansions.

MGSR2

There were 48 instances of MGSR2 identified in the genome of *M. grisea.* These sequences were aligned using clustalw and produced an 1159 base long alignment. Collapsing these sequences into haplotypes recoding indels returned 19 different haplotypes with two haplotypes, 10 and 13, containing the majority of sequences, 15 and 13 respectively [Figure 5.11]. A compatibility matrix was generated and a 170bp incompatible block can be seen in the 3' region [Figure 5.12]. This 3' block of incompatible sites likely represents a recombination event and the larger 5' 989 base pair region was used for further analysis of this element.

Potential trees were created using Genetree and 1 million runs using a Wattersons estimated theta value of 7.21 predicted the most likely tree numbered 12 out of 32 possibilities with a likelihood of 3.0829e-24 which was closest to the unrooted tree likelihood of 9.6052e-24. The ages of mutations were generated using 10 million runs and a graphical representative of the coalescent tree can be seen in Figure 5.13. The coalescent tree shows a pattern similar to P1. Again, the lineage begins with relatively few haplotypes and there is a major expansion in lineages just following a mutation. In this case it was mutation 16.

Pot2

The 62 sequences identified as Pot2 created an alignment of 2070 bases which was collapsed into haplotypes recoding indels using SNAP Map [Figure 5.14]. Compatibility analysis of the sequences in Figure 5.15 showed that the inconsistencies could not easily be rectified and required that the dataset be looked at clade by clade. Phylip was used to create a most parsimonious phylogenetic tree of the data and each clade was analyzed for compatibility within itself. Clade 4, containing 19 sequences composed of haplotypes 10, 11, 14, 15, 17, 20, 3, 31, 32, 33, 34, 5, 7, and 8, was the largest clade which had no incompatible sites within it and was used as the dataset for further analysis.

Twenty eight potential trees were generated by Genetree and one million repetitions were performed to identify the tree with the highest likelihood. Tree 1, with a likelihood of 1.4954e-25 had the highest likelihood of being the true tree when compared to the unrooted tree's likelihood of 6.4934e-24. The ages of mutations were generated using 10 million runs and a graphical representative of the coalescent tree can be seen in Figure 5.16. This transposable element also shows the same pattern of starting with few lineages and rapidly expanding after a mutation, 9.

Pot3

A 1912 base pair alignment was produced from the 35 representatives of Pot3 identified within the genome. This alignment was collapsed recoding indels using SNAP Map into 25 distinct haplotypes [Figure 5.17] and compatibility analysis of the dataset showed that the sequences were completely compatible [Figure 5.18].

Potential trees were created using Genetree and 1 million runs using a Wattersons estimated theta value of 10.93 predicted the most likely tree numbered 42 out of 45 possibilities with a likelihood of 1.2454e-19 which was closest to the unrooted tree likelihood of 2.0718e-19. The ages of mutations were generated using 10 million runs and a graphical representative of the coalescent tree can be seen in Figure 5.19. The coalescent history of Pot3 shows a large number of lineages present from the beginning with little local expansion.

Occan

Thirty three representatives of Occan were identified within the genome and they produced a 2781 base pair alignment. This alignment was collapsed recoding indels using SNAP Map [Figure 5.20] and a compatibility matrix was produced from the resulting 31 haplotypes to determine if the genetic histories of the sequences were in agreement. As can be seen in Figure 5.21, there exists a good deal of conflict in the data which could not easily be resolved with the removal of a single site or recombination block. In order to obtain a compatible set of data, phylip was used to create the most parsimonious phylogenetic tree and each resulting clade was inspected for compatibility within itself. Clade 4, containing 20 of the 33 sequences consisting of haplotypes 1, 10, 11, 13, 14, 15, 16, 17, 2, 20, 21, 22, 24, 28, 29, 30, 31, 4, and 9, was the largest clade with the least amount of inconsistencies, only requiring the removal of only site 5 to be completely compatible [Figure 5.22], and was chosen as the dataset for subsequent coalescent analysis.

Using Genetree, all potential trees for this dataset were generated and 10 million repetitions using a Watterson's estimated theta value of 18.32 predicted the most likely tree numbered 15 out of 65 possibilities with a likelihood of 8.6016e-31 which was closest to the unrooted tree likelihood of 8.6557e-31. The ages of mutations were generated also using 10 million runs and a graphical representative of the coalescent tree was created and can be seen in Figure 5.23. Occan's coalescent tree shows the similar pattern of few lineages present initially followed by a few branches and then a large expansion following mutation number 36.

Discussion

This study sought to explore the genetic history of *M. grisea's* transposable elements in the hopes of better understanding their role in genomic evolution and pathogenicity. The insertion of these elements has been show to cause major genomic events including modification in gene expression, disruption of coding regions and even large polymorphism event. We focused on the most prolific transposable elements in *M. grisea*, MAGGY, Pot2, Pot3, Occan and MGSR2 with expectation that the larger sample sizes would provide more genetic variation and subsequently more insight into the history of transposable elements in this fungus's genome. We would have liked to date the times of introduction of each transposable element to identify if there is a correlation with a transposable element introduction and host shift to rice. The time to most recent common ancestor, which is available through a simple calculation once the coalescent history has been resolved, would provide an estimate of the timing of each elements' introduction to the genome. An important variable in that calculation, the generation time of each element, is unknown and could be impossible to determine so an estimate based on coalescent theory is unavailable.

MAGGY, an LTR transposon, is the most prolific within the genome, with an estimate of over one hundred copies being previously reported and over two hundred partial copies identified within the genome assembly in this study. Compatibility analysis of the full length copies of element identified a region in the Pol gene of roughly four hundred bases which was incompatible with the other 3 kilobases of MAGGY sequence. This region, identified as a transferase domain by Interpro analysis, was extracted from the other partial copies of the MAGGY element identified in the genome and all regions were compatible with each other. The coalescent tree for this portion of the element showed little branching and primarily single mutations leading to the differences in haplotypes. The fact that this region shows incompatibility with the rest of the element but did not show any incompatibility within itself joined with the minimal mutations seen in the region, makes a case for selective pressure on this region possibly as domain integral to the Pol gene's function in transposition. The incompatibility with the rest of the regions likely arose from a recombination event which resulted in the addition of this region to the Pol gene at some point in the history of MAGGY.

The P1 region of MAGGY, Occan, Pot2, and MGSR2 all show a very similar coalescent tree topology in which a relatively few initial lineages exist and there is a massive expansion of a single lineage at some point. An explanation of this occurrence is that the transposable elements existed in the genome for a period of time slowly increasing their number at a rate which avoided detection of genomic repair until some catastrophic genomic event led to the loss of genomic repair mechanisms and allowed the massive expansion of these elements. The most well known filamentous fungal genomic defense mechanism is RIP which identifies repetitive genomic DNA and causes a large number of transitions in both copies of the repeated element. RIP has been shown to only occur at a point just prior to meiosis in the sexual cycle of *Neurospora crassa* [21, 22]. Filamentous fungi have evolved the ability to reproduce both sexually and asexually in order to ensure the survival of their genetic material. However, the majority of pathogenic *M. grisea* isolates tested only show asexual reproduction and, until recently, no natural examples of sexual reproduction had been reported though there was indirect evidence for the existence of the sexual cycle at some point in the fungus's history [23, 24]. At a point in *M. grisea's* history, it seems to have shifted to primarily asexual reproduction. A genomic event which removed the fungus's ability to reproduce sexually would also have the secondary effect of disabling its primary defense against repetitive DNA, RIPing. The loss of a sexual cycle would bypass the pre-meiotic stage required for RIP to identify and disable repetitive elements. Prior to the loss of RIP and the sexual cycle, transposable elements could have remained essentially inactive in the genome only reproducing infrequently, coinciding with asexual reproduction.

Pot3 showed a slightly different topology than the majority of *M. grisea's* transposable elements. There was evidence for a major expansion but the element's history didn't begin with a few lineages as the rest of the transposable elements in this study had. The expansion seems to have occurred as soon as the transposable element was introduced into the genome. An explanation of this different topology could be that the Pot3 element was more recently introduced to the genome, possibly very near or coinciding with the shift to primarily asexual reproduction or that the element was completely inactive if it was resident in the genome for a period of time prior to the asexual shift. Couch et al. support that the Pot3 element has recently expanded in their study examining host shift in *M. grisea.* They present evidence that following the host shift to cutgrass and torpedo grass the copy number of Pot3 in the genome increased.

The *a priori* tests of neutrality suggested that these transposable elements had experienced a sudden population expansion or genetic selection which is supported by the topology of the tree. A single lineage from each TE seems to have been selected and expanded as a result of the change in *M. grisea's* ability to protect its genome. The selection could have been caused by the action of genomic defenses which left only a single active copy in the genome that was then free to propagate after loss of RIP.

The addition of multiple copies of transposable elements to the genome prepared the fungus for future mutation in response to the selective pressures of being a pathogen. The rapid expansion of these elements in the genome provided *M. grisea* with an evolutionary advantage and is likely temporally related to the evolution of pathogenicity or a host shift by the fungus.

Conclusion

Through the novel application of population genetics models and theories to a population of transposable elements within the *M. grisea* genome, we were able to increase our understanding of these element's histories. The analysis identified recombination events in several elements and showed that all elements studied had undergone a major expansion in the past. There is some evidence that the Pot3 expansion followed the host shift to cutgrass and torpedo grass. Further field work will be required to determine if the expansion of other transposable elements in *M. grisea* can be linked to pathogenic events. The coalescent techniques used in this study have provided interesting and novel insights into the history of transposable elements in this genome. The role that these elements play in the genomic evolution of *M. grisea* is obvious but their exact historical role in the evolution of pathogenicity remains to be sufficiently answered.

References

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

[2] Daboussi, M.J. (1997) Fungal transposable elements and genome evolution. Genetica 100, 253-260.

[3] Kang, S.C., Sweigard, J.A. and Valent, B. (1995) The PWL host specificity gene family in the blast fungus Magnaporthe grisea. Molecular Plant-Microbe Interactions 8, 939-948.

[4] Farman, M.L., Eto, Y., Nakao, T., Tosa, Y., Nakayashiki, H., Mayama, S. and Leong, S.A. (2002) Analysis of the structure of the AVR1-CO39 avirulence locus in virulent rice-infecting isolates of Magnaporthe grisea. Molecular Plant-Microbe Interactions 15, 6-16.

[5] Kang, S., Lebrun, M.H., Farrall, L. and Valent, B. (2001) Gain of virulence caused by insertion of a Pot3 transposon in a Magnaporthe grisea avirulence gene. Molecular Plant-Microbe Interactions 14, 671-674.

[6] Brookfield, J.F.Y. (2005) The ecology of the genome - Mobile DNA elements and their hosts. Nature Reviews Genetics 6, 128-136.

[7] Carbone, I., Liu, Y.C., Hillman, B.I. and Milgroom, M.G. (2004) Recombination and migration of Cryphonectria hypovirus 1 as inferred from gene genealogies and the coalescent. Genetics 166, 1611-1629.
[8] Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) Clustal-W - Improving the Sensitivity of Progressive Multiple Sequence Alignment through Sequence Weighting, Position-Specific Gap Penalties and Weight Matrix Choice. Nucleic Acids Research 22, 4673-4680.
[9] Felsenstein, J. (1993) PHYLIP (Phylogeny Inference Package) version 3.5c. In: Ed.^Eds.), Vol. p.^pp. Distributed by the author. Department of Genetics, University of Washington, Seattle.
[10] Watterson, G.A. (1975) Number of Segregating Sites in Genetic Models without Recombination. Theoretical Population Biology 7, 256-276.
[11] Fu, Y.X. (1996) New statistical tests of neutrality for DNA samples from a population. Genetics 143, 557-570.
[12] Tajima, F. (1989) Statistical-Method for Testing the Neutral Mutation Hypothesis by DNA Polymorphism. Genetics 123, 585-595.
[13] Hudson, R.R., Boos, D.D. and Kaplan, N.L. (1992) A Statistical Test for Detecting Geographic Subdivision. Molecular Biology and Evolution 9, 138-151.
[14] Griffiths, R.C. and Tavare, S. (1994) Ancestral Inference in Population-Genetics. Statistical Science 9, 307-319.
[15] Fu, Y.X. and Li, W.H. (1993) Statistical Tests of Neutrality of Mutations. Genetics 133, 693-709.

[16] Simonsen, K.L., Churchill, G.A. and Aquadro, C.F. (1995) Properties of Statistical Tests of Neutrality for DNA Polymorphism Data. Genetics 141, 413-429.

[17] Fu, Y.X. (1997) Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics 147, 915-925.

[18] Jakobsen, I.B., Wilson, S.R. and Easteal, S. (1997) The partition matrix: Exploring variable phylogenetic signals along nucleotide sequence alignments. Molecular Biology and Evolution 14, 474-484.

[19] Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. and Lopez, R. (2005) InterProScan: protein domains identifier. Nucleic Acids Research 33, W116-W120.

[20] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A.N., Orchard, S., Pagni, M., Pointing, C.P., Quevillon, E., Selengut, J., Sigrist, C.J.A., Silventoinen, V., Studholme, D.J., Vaughan, R. and Wu, C.H. (2005) InterPro, progress and status in 2005. Nucleic Acids Research 33, D201-D205.

[21] Foss, E.J., Garrett, P.W., Kinsey, J.A. and Selker, E.U. (1991) Specificity of Repeat-Induced Point Mutation (Rip) in Neurospora - Sensitivity of Non-Neurospora Sequences, a Natural Diverged Tandem Duplication, and Unique DNA Adjacent to a Duplicated Region. Genetics 127, 711-717.

[22] Cambareri, E.B., Jensen, B.C., Schabtach, E. and Selker, E.U. (1989) Repeat-Induced G-C to a-T Mutations in Neurospora. Science 244, 1571-1575.
[23] Ikeda, K., Nakayashiki, H., Kataoka, T., Tamba, H., Hashimoto, Y., Tosa, Y. and Mayama, S. (2002) Repeat-induced point mutation (RIP) in Magnaporthe grisea: implications for its sexual cycle in the natural field context. Molecular Microbiology 45, 1355-1364.
[24] Mekwatanakarn, P., Kositratana, W., Phromraksa, T. and Zeigler, R.S. (1999) Sexually fertile Magnaporthe grisea rice pathogens in Thailand. Plant Disease 83, 939-943.

Table 5.1: Distribution of TEs within the *M. grisea* genome

Chromosome						
	Maggy	Pot2	Pot3	Pot4	Occan	MGSR2
1	23	7	4	9	4	2
2	58	12	5	9	5	8
3	42	13	5	11	7	8
4	14	11	4	15	6	9
5	12	4	3	8	0	1
6	9	2	4	7	4	4
7	26	8	8	16	5	7
Unknown	55	5	2	13	2	9
Total	229	62	35	88	33	48

Table 5.2 – Population statistics and tests of neutrality for selected transposable elements in *M. grisea*

	Population Statistics				Θ	Tests of Neutrality								Hudson's
Element	length	number	Seg sites	k		Tajima's T		Fu and Li's D*		Fu and Li's F*		Fu's Fs		
Maggy I1	623	30	20	2.343	3.80 / 8.08	-1.819	**	-2.809	*	-2.354	*	-14.577	***	ns
Maggy P1	384	24	13	3.13	4.20 / 8.57	-1.819	**	-3.069	***	-2.893	***	-4.667	**	ns
MGSR2	989	48	43	2.076	6.22 / 18.48	-2.258	***	-4.44	***	-4.078	***	-21.379	***	ns
Pot2	2066	19	18	3.918	7.09 / 15.45	-2.071	***	-3.138	***	-3.006	***	-14.335	***	ns
Pot3	1912	35	32	2.034	8.96 / 24.28	-2.245	***	-4.219	***	-3.906	***	-19.651	***	ns
Occan	2779	33	19	5.537	13.34 / 30.16	-2.158	***	-3.238	***	-3.114	***	-10.076	***	ns

Theta estimate Tajima / Watterson

ns	not significant
*	<0.05
**	<0.02
***	<0.005

Figure 5.1: Histogram of Maggy regions represented in the genome

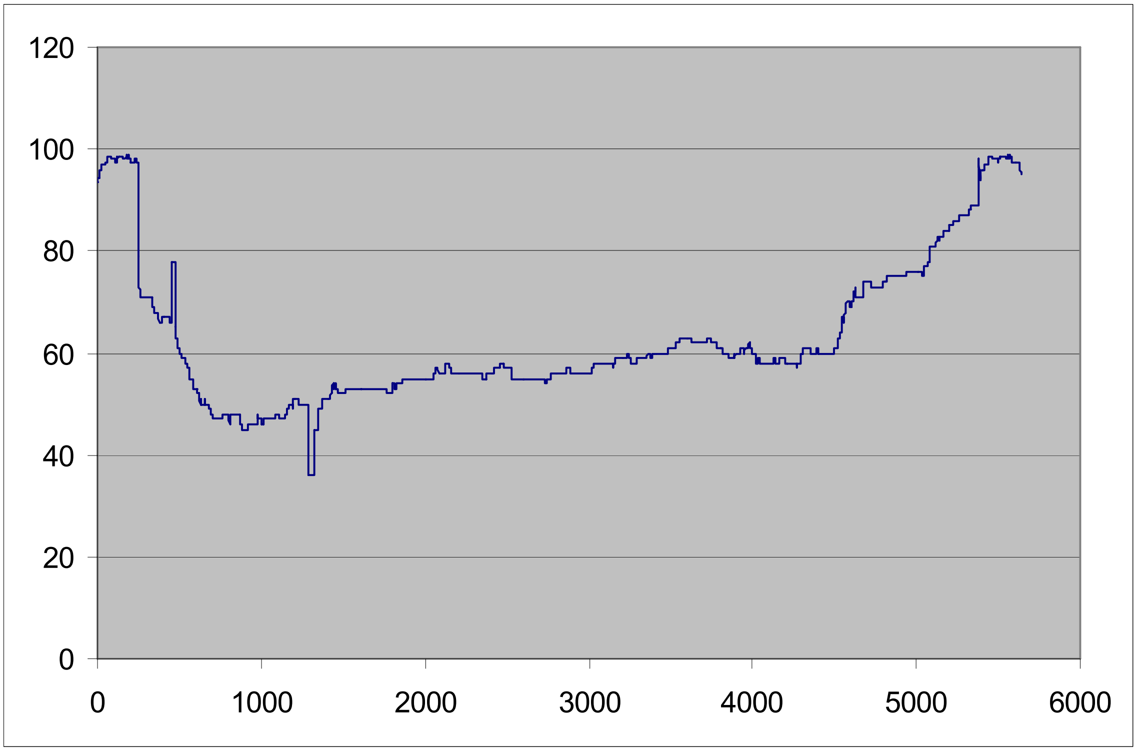


Figure 5.2: Distribution of haplotypes and mutations in the full length copies of Maggy

```
Position                                  1111111111111111111111111111111111111111111112222222222222222333333333333333444444444555555555
                             12345667888889990000111112222222333333344444444444455555556667000001136677790122444456778923344599011112366
                            33464188462456670132569224622456924556990012556788904568890001445670566777682121349421731129032076245881001
                            13394012706206894295504231328081806250891565287346498667913694591698394434534529694568325856849784339515570

Site Number                                                                                                                     11111111
                                     111111111122222222223333333333444444444455555555556666666666777777777788888888889999999999900000000
                            12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012345678901234567

Consensus                   T1CTTC111CG1GG11GTG1CT11T1TTACT1TGTA1CGCCCTGGCTTAGCGAGCA1AC1AGCGCCCTCC111A111112GAG1TGCC1GC11G12GGGC1G11ATA
Site Type                   t-ttvt---tt-tt--tvt-tt--t-ttttt-tttt-vttttttttttvvtttttt-tt-tttvtttttt---v------ttt-tttt-tt--t--tttt-t--ttt
Character Type              - -i--   -- i-  i-- ii  i iiiii iiii iiiiiiiiiii--ii---- -- --i-i-ii--   i      i-- -i-- i-  -  ---- -  -i-
Code Key                    *_****___**_**__***_**__*_*****_****_*******************_**_**********___*______***_****_**__*__****_*__***
H1          (  1) ................A.................................................T.........1.......T.A...A.1..........
H2          (  1) C2.C........A......2TC.2C2CCGTC.CACG.AATTTCAATCC..TA..........T.TTTC....2T2..........A..2.....2..........C.
H3          (  1) ..T..T2....2.A...G....2............................................T22.........G.......AT........T..22...
H4          (  1) ...CA.........22..A...........C.....2...................2.T2.A..........2...222....2.A......2..............
H5          (  1) ...C...22...A.......TC..C.CCGTC2CACG.AATTTCAATCC..TAGATG.G..G.T.T.TC....2T2.....A.....T..A.2....AAA..A..GC.
H6          (  1) .........TA.....A..............................TT............T...........2....A.A.C...............2.....G
```

Figure 5.3: Compatibility matrix of full length copies of Maggy

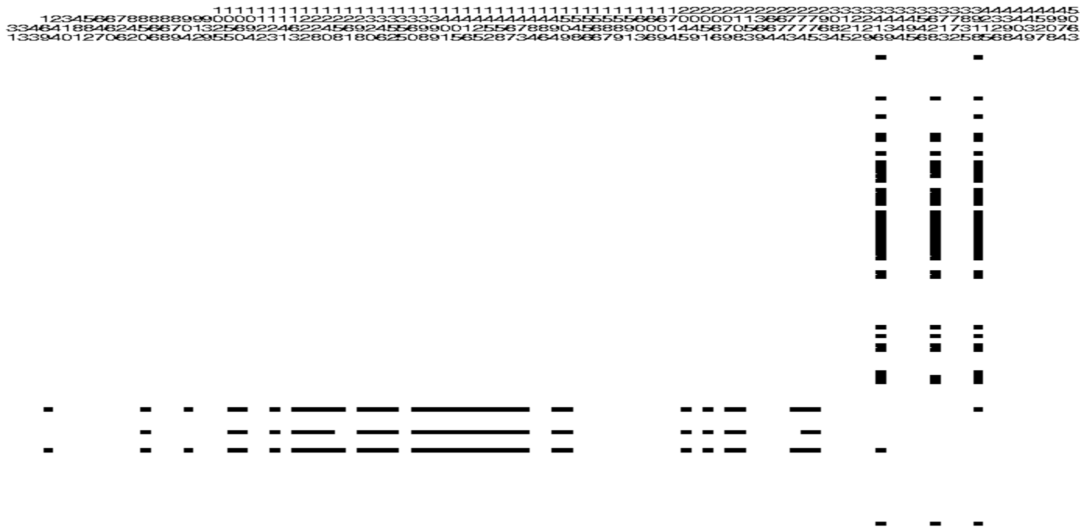


Figure 5.4: Distribution of haplotypes and mutations in the Maggy P1 region

```
Position                          1111222223333
                 133333348888899990126555781145
                 523568903456802350338268710186

Site Number               111111111122222222223
                 123456789012345678901234567890

Consensus        GC1A1CT1CGACAGA1T1121CGC2CT1GC
Site Type        vv-v-tt-vvtvvtt-t----tvt-tt-tt
Character Type   -------------------------ii-i-
H1          (  1) ...................1....1.....
H2          (  1) T.............................
H3          (  1) .G.T..C2......................
H4          (  1) ..2.2T...........2...........
H5          (  1) ........GCGACAG2C...........A.
H6          (  1) .................2............
H7          (  5) ..............................
H8          (  1) ....................2.........
H9          (  6) .....................T.T.TC...
H10         (  1) ......................T.......
H11         (  1) .........................TC...
H12         (  1) ...........................2..
H13         (  1) ............................A.
H14         (  2) .............................T
```

Figure 5.5: Compatibility matrix of the Maggy P1 region

```
                    111111111122
 sites    123456789012345678901

      1   .....................
      2   .....................
      3   .....................
      4   .....................
      5   .....................
      6   .....................
      7   .....................
      8   .....................
      9   .....................
     10   .....................
     11   .....................
     12   .....................
     13   .....................
     14   ................x....
     15   .....................
     16   ................x....
     17   .............x.x.xxx.
     18   ................x....
     19   ................x....
     20   ................x....
     21   .....................
```

Figure 5.6: Coalescent Tree of the Maggy P1 region

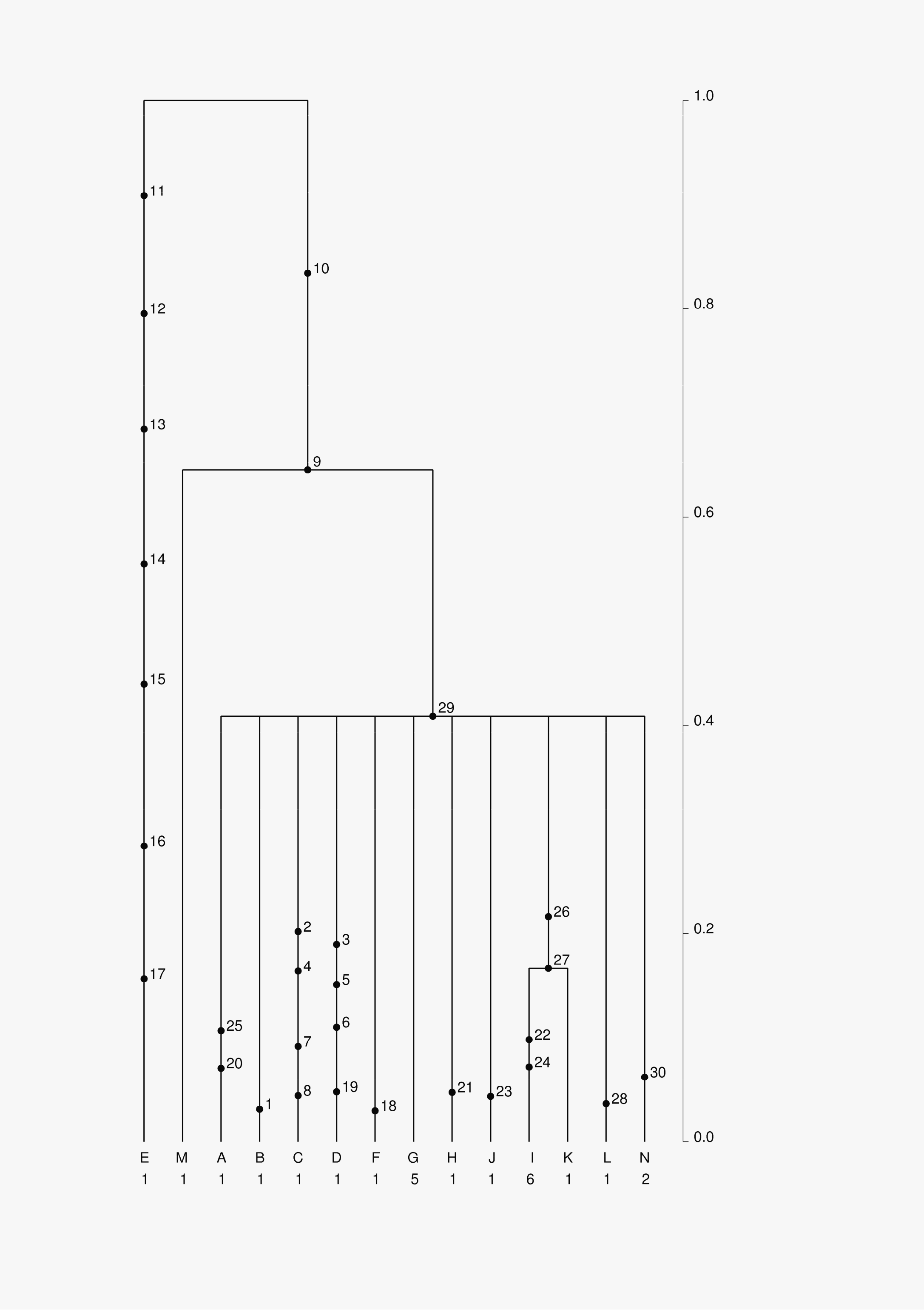


Figure 5.7: Distribution of haplotypes and mutations in the Maggy D1 region

```
Position             11111111222222222233333333444444555556
                  48914556667022345668922334789357888012231
                 980351031299306169038638160234436135731287

Site Number               111111111122222222223333333333444
                 123456789012345678901234567890123456789012

Consensus        GACGCCCGGGGCGGCCCACATCGGCGGGCGTGCGCGCGCCGG
Site Type        tvttttvtvttttttttttvttttttvttttttttttttttt
Character Type   -i--------i------i----ii-i----ii-i--i-----
Code Key         ******************************************
H1         (  1) ..........A........................A......
H2         (  1) A.TA.T...AA.AATTT.T...AA.AA.TA..TAT.TATTAA
H3         (  1) .C............................CA..........
H4         (  1) .C....GA......................CA..........
H5         (  1) ....T..................A..................
H6         (  1) ........T.A......G.GA...............T.....
H7         (  2) ......................A.T.................
H8         (  1) ...........T...........A..................
H9         (  1) .....................T.A..................
H10        (  2) .......................A.A.......A........
H11        ( 16) ..........................................
H12        (  1) ...............................A..........
H13        (  1) .......................A..................
H14        (  1) ..........A......G.........T........T.....
H15        (  1) ..........A...............................
```

Figure 5.8: Compatibility matrix of the Maggy D1 region

```
                   111111111122222222223333333333444
 sites    123456789012345678901234567890123456789012

     1    ..........................................
     2    ..........................................
     3    ..........................................
     4    ..........................................
     5    ..........................................
     6    ..........................................
     7    ..........................................
     8    ..........................................
     9    ..........................................
    10    ..........................................
    11    ......................xx.x.......x........
    12    ..........................................
    13    ..........................................
    14    ..........................................
    15    ..........................................
    16    ..........................................
    17    ..........................................
    18    ..........................................
    19    ..........................................
    20    ..........................................
    21    ..........................................
    22    ..........................................
    23    ..........x............x.x.......x..x.....
    24    ..........x...........x.............x.....
    25    ..........................................
    26    ..........x...........x.............x.....
    27    ..........................................
    28    ..........................................
    29    ..........................................
    30    ..........................................
    31    ..........................................
    32    ..........................................
    33    ..........................................
    34    ..........x...........x.............x.....
    35    ..........................................
    36    ..........................................
    37    ......................xx.x.......x........
    38    ..........................................
    39    ..........................................
    40    ..........................................
    41    ..........................................
    42    ..........................................
```

Figure 5.9: Maximum parsimony phylogenetic tree of the Maggy D1 region

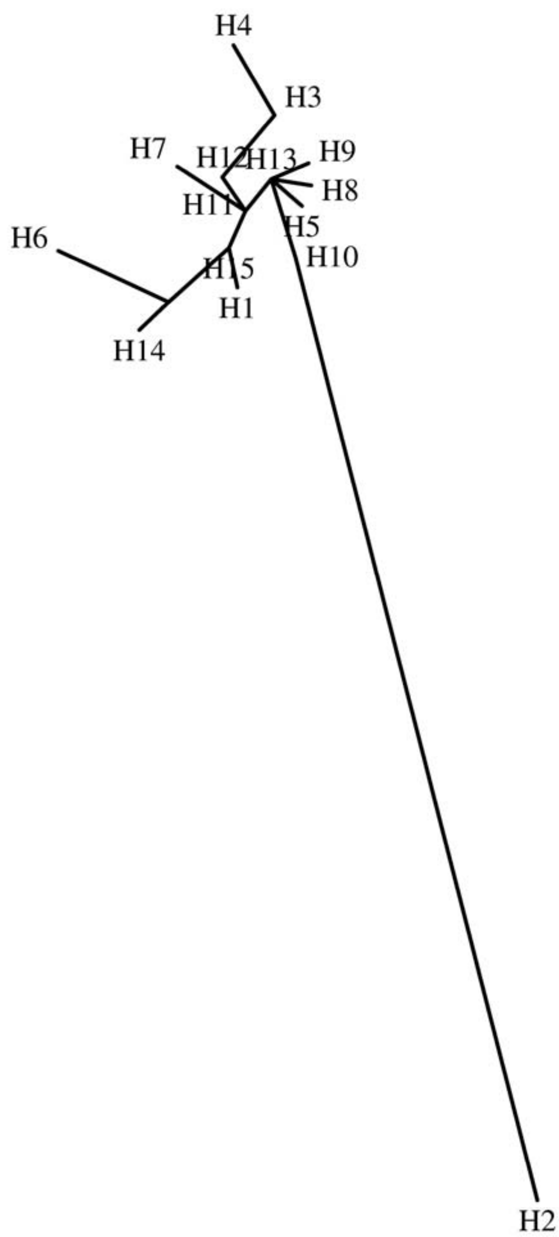


Figure 5.10: Coalescent tree of the Maggy D1 region

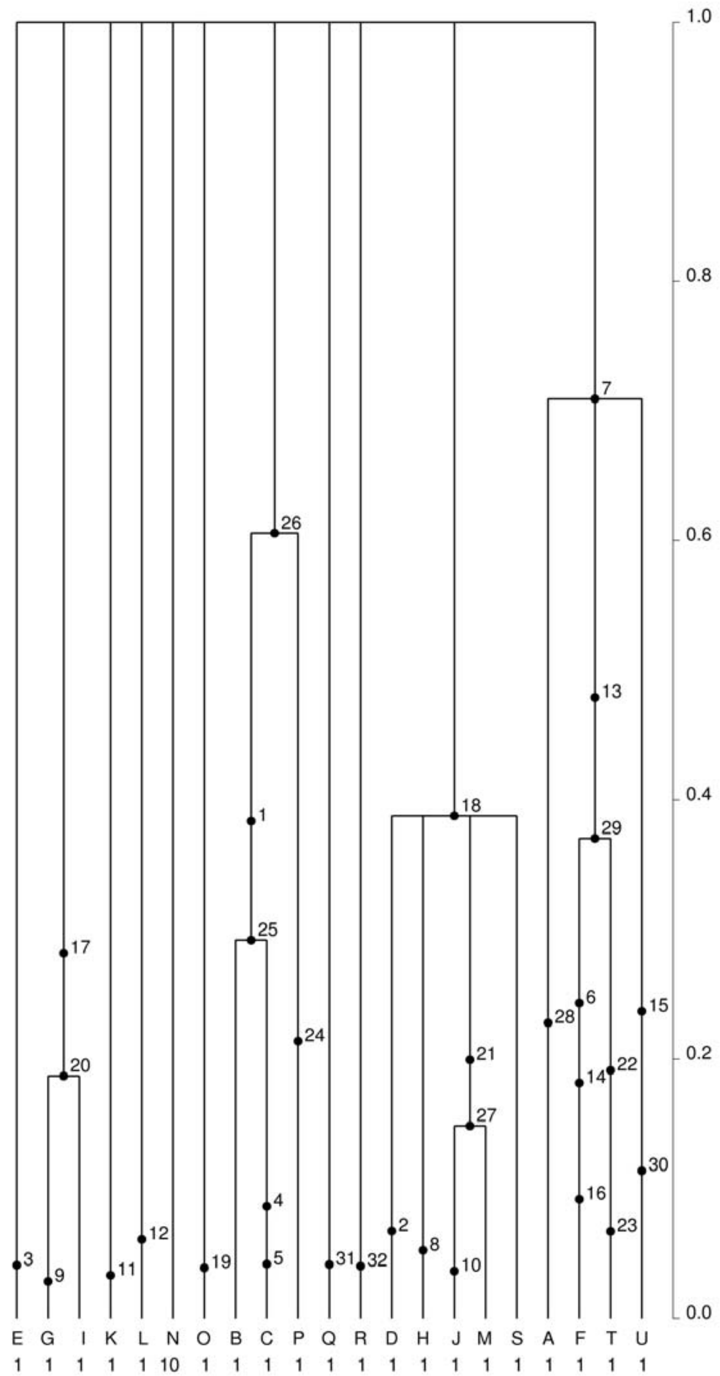


Figure 5.11: Distribution of haplotypes and mutation in MGSR2

```
Position                111111111223333334457789999
                    1349011377888141555586842750234
                   71036724802678034046716916583492

Site Number                 11111111112222222222333
                   12345678901234567890123456789012

Consensus          TCGGGACTCAGTCTCCCACGGCCGGGGCCGCG
Site Type          vvtvvvtvtvttttttvvvvvttttttttttt
Character Type     ------i----i---i-----i------i---
H1         (  2)   ...........C...T................
H2         (  1)   AG.............T................
H3         (  1)   ..A.......A..............A......
H4         (  1)   ...T...G.......T..........A.TA..
H5         (  1)   ....T........C.................A
H6         (  1)   .....C..T.......................
H7         (  2)   ......T.........................
H8         (  1)   ......T...........ATT...........
H9         (  1)   .........C............T.........
H10        ( 15)   ...............T................
H11        (  1)   ............T..T................
H12        (  1)   ..............T.................
H13        ( 13)   ................................
H14        (  1)   ................A....T.A........
H15        (  1)   .................T..............
H16        (  1)   .....................T..........
H17        (  1)   ...............T...........T..T.
H18        (  2)   ...............T............T...
H19        (  1)   ...........C...T........A.......
```

Figure 5.12: Compatibility matrix of MGSR2

```
                  1111111111222222222233333333334
 sites   1234567890123456789012345678901234567890

     1   ........................................
     2   ........................................
     3   ........................................
     4   ........................................
     5   ........................................
     6   ........................................
     7   ........................................
     8   ........................................
     9   ........................................
    10   ........................................
    11   ........................................
    12   ..................................x.x.x.
    13   ........................................
    14   ........................................
    15   ........................................
    16   ................................x.x.x.x.
    17   ........................................
    18   ........................................
    19   ........................................
    20   ........................................
    21   ........................................
    22   ........................................
    23   ........................................
    24   ........................................
    25   ........................................
    26   ........................................
    27   ........................................
    28   ........................................
    29   ..................................x...x.
    30   ........................................
    31   ........................................
    32   ........................................
    33   ...............x......................x.
    34   ........................................
    35   ...........x...x............x.......x.x.
    36   ........................................
    37   ...........x...x..................x...x.
    38   ........................................
    39   ...........x...x............x...x.x.x...
    40   ........................................
```

Figure 5.13: Coalescent tree of MGSR2

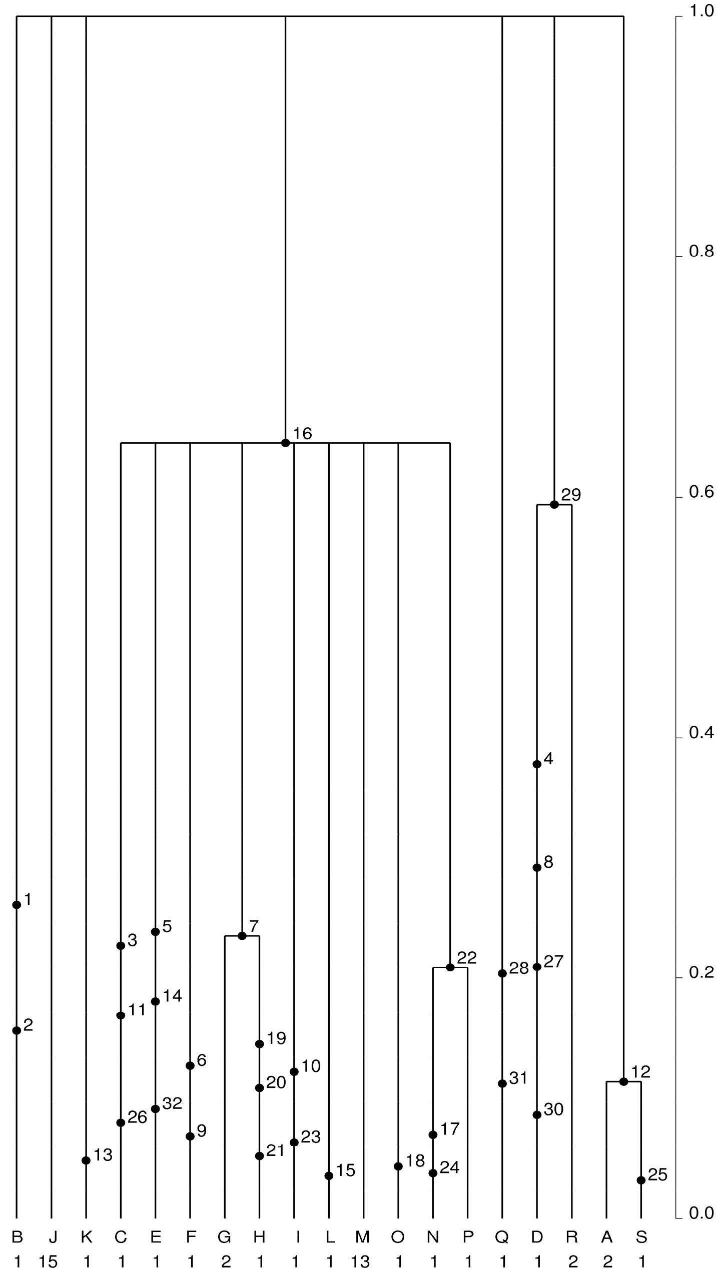


Figure 5.14: Distribution of halotypes and mutations in Pot2

```
Position                                                11111111111111111111111111111111111111111111111111111111122222
                      122222333455555567777999990001111122222444444555555555556677777777778888888888999900000
                 14678513466367733344601269246677790022435888113356000011245774823334568991113478899044644556
                 81903516212661712401633032286734504529795478346783134534763249804583139562389325901149529392

Site Number               11111111112222222222333333333344444444445555555555666666666677777777778888888888999
                 12345678901234567890123456789012345678901234567890123456789012345678901234567890123456789012

Consensus        ACCTGCTCCATTAAGCATAAATAGCGCAACAGAGCTATAGGACGACATCGACACATATTAACAGTTTATTGCGTTTGGGGGAGGGGGAGTCA
Site Type        tvvttvtttvvttttvvttttvtvtvtvtvvvttvttvvvttvvtvttvvvvvvttvvtvvvtvvvtvtvttvvvvtttttttttvvttvvv
Character Type   i---i------iii----i-i---ii--iii---------------i-i-----ii--i--i-----i--i--------i-----ii-----
H1         (  8) ............................................................................................
H2         (  1) G..................................................ACA......................................
H3         (  1) GG..A.........................................G.......GC...........C..A..............T......
H4         (  1) G...A...........................................A..................C..A.....................
H5         (  1) G...A......CG.................................G.......GC...........C..A..............T......
H6         (  1) G...A......CG.....G........TGAC.............................................................
H7         (  1) G...A............C............................G.......GC...........C..A..............T......
H8         (  1) G...A...................T.....................G.......GC...........C..A..............T......
H9         (  1) G...A............................A..G.C...............GC...........C..A.....AAAA...A.T......
H10        (  5) G...A.........................................G.......GC...........C..A..............T......
H11        (  1) G...A.........................................G.......GC...........CC.A..............T......
H12        (  1) G...A.......................................................................................
H13        (  1) G...A.................................................GC...........C..A..............T......
H14        (  1) G..........CG.....G...........................G.......GC...........C........................
H15        (  1) G.............A..........T....................G.......GC...........C..A..............T......
H16        (  1) G..................G..........................G................TGGC..A..GGG.................
H17        (  2) G........................T....................G.......GC...........C..A..............T......
H18        (  1) G...............................................A..................C........................
H19        (  1) G...........................................................................................
H20        (  1) ..GC.....................T............................GC...........C..A..............T......
H21        (  1) ....A......CG.....G....T....GAC.............................................................
H22        (  1) .....A......................................................................................
H23        (  1) ......C...........................GC....A.....G.......GC...............TC...................
H24        (  1) .......T.....................................................G..............................
H25        (  1) ........T...................................................................................
H26        (  1) .........TG.............T...................GA..............................................
H27        (  2) .............G..............................................................................
H28        (  1) .............G......G.........................G.......GC....................................
H29        (  1) .............G......G.......................................................................
H30        (  1) ...............AC....................................................................T......
H31        (  1) .....................A...T....................G.......GC...........C..A..............T......
H32        (  1) ......................G..T....C...............G.......GC...........C..A..............T......
H33        (  1) .........................T....................G.......GC...........C..A.............AT......
H34        (  1) .........................T....................G.......GC...........C..A..............T......
```

Figure 5.14 (cont.)

```
H35        (  1) ..........................T.....................................................AGA....GAGAC
H36        (  1) ...............................TG...........................................................
H37        (  1) .....................................A.......................G..............................
H38        (  1) .......................................T....................................................
H39        (  1) .........................................G.C..GC.TT.....CA.TT......C........................
H40        (  1) ..........................................G.................................................
H41        (  2) ..............................................G.......GC....................................
H42        (  1) ......................................................GC....................................
H43        (  1) ..........................................................C...........................C.....
H44        (  1) ..........................................................C....................A......C.....
H45        (  1) ..............................................................G.............................
H46        (  1) ......................................................................................C.....
```

Figure 5.15: Compatibility matrix of Pot2

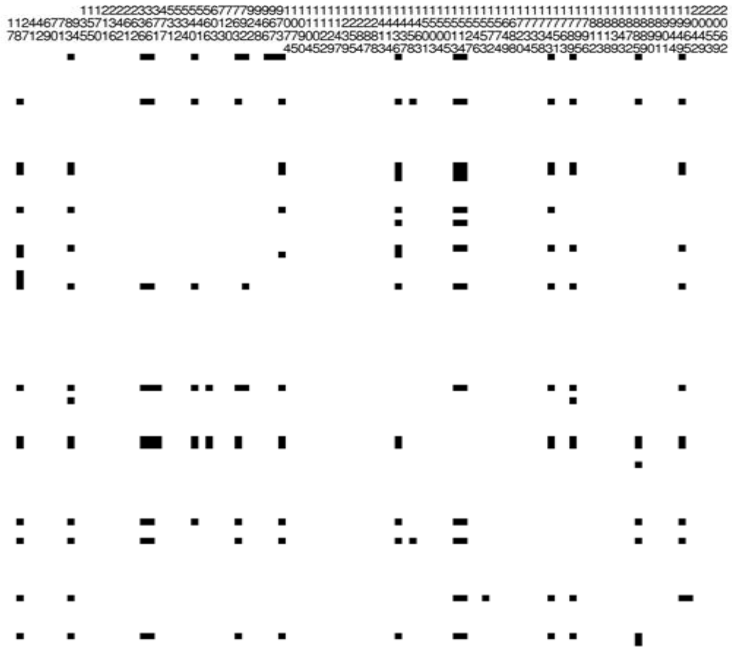


Figure 5.16: Coalescent tree of Pot2

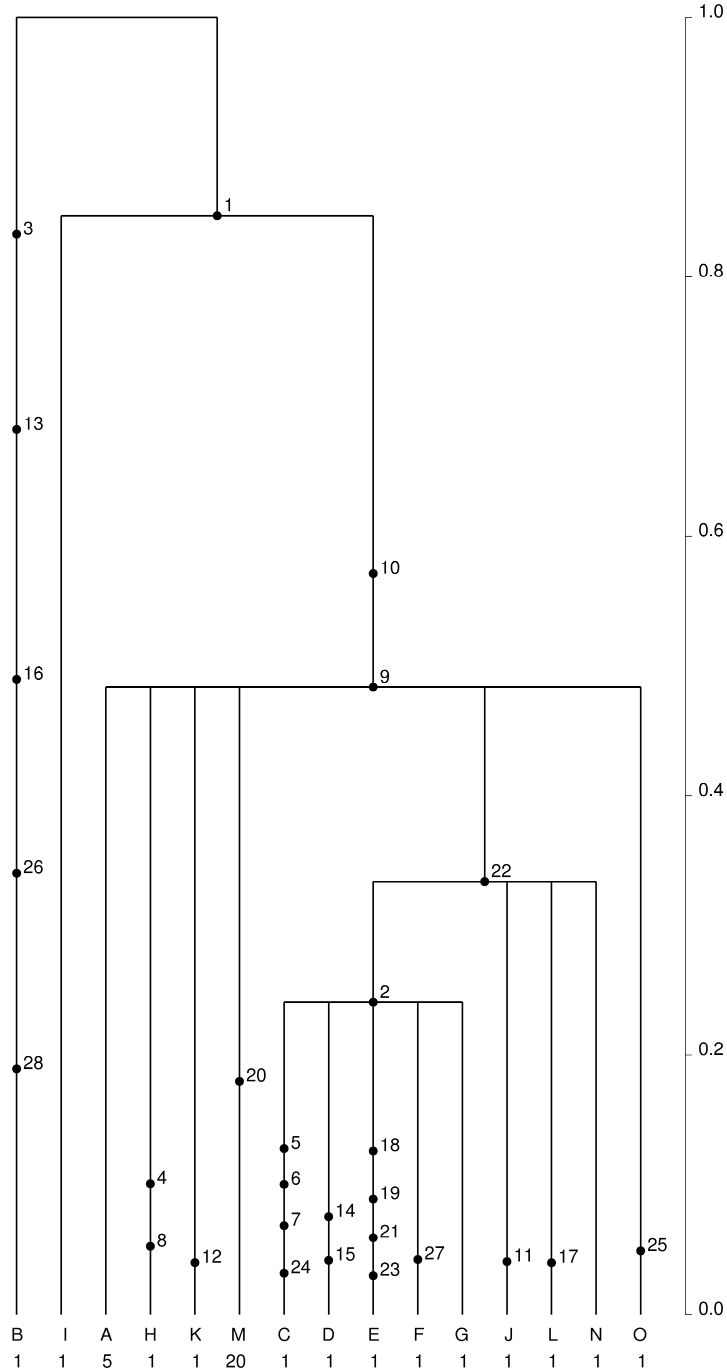


Figure 5.17: Distribution of haplotypes and mutation in Pot3

```
Position                                  11111111111111111111
                       122223555555555588880112233445666777889
                  37899711678122466799355726869343710014559990
                 708138837516007726416478266823571486749024565

Site Number               111111111122222222223333333333444444
                 123456789012345678901234567890123456789012345

Consensus        1C111AACA1A1111111111AGACCCATTCACCCAGTACCCCCA
Site Type        -t---vvvt-t----------vttttvvvtttvvttttttvvtvv
Character Type   i--------------i-------------i------------iii
H1          (  1) ........................T................G...
H2          (  2) 2............................................
H3          (  1) .T..............................GG...........
H4          (  1) ..222.....................................T.C
H5          (  1) .....C.......................................
H6          (  1) ......CGG.G...........................GTA....
H7          (  1) .........2.2.................C...............
H8          (  1) ............2......................GAC.....A.
H9          (  1) .............2...............................
H10         (  1) ..............2..............................
H11         (  1) ...............2.............................
H12         (  1) ...............2............A................
H13         (  1) ................2............................
H14         (  1) .................2..2.............T..........
H15         (  1) ..................2............G.............
H16         (  1) ...................2.......C.................
H17         (  1) .....................T.......................
H18         (  1) ......................AG.....C...............
H19         (  9) .............................................
H20         (  1) .........................T...................
H21         (  1) ..........................A..................
H22         (  1) .............................C...............
H23         (  2) ..............................T..............
H24         (  1) ..........................................T.C
H25         (  1) ...........................................A.
```

Figure 5.18: Compatibility matrix of Pot3

```
The sequences show no incompatibility
                   111111111122222222223
 sites    123456789012345678901234567890

      1   ..............................
      2   ..............................
      3   ..............................
      4   ..............................
      5   ..............................
      6   ..............................
      7   ..............................
      8   ..............................
      9   ..............................
     10   ..............................
     11   ..............................
     12   ..............................
     13   ..............................
     14   ..............................
     15   ..............................
     16   ..............................
     17   ..............................
     18   ..............................
     19   ..............................
     20   ..............................
     21   ..............................
     22   ..............................
     23   ..............................
     24   ..............................
     25   ..............................
     26   ..............................
     27   ..............................
     28   ..............................
     29   ..............................
     30   ..............................
```

Figure 5.19: Coalescent tree of Pot3

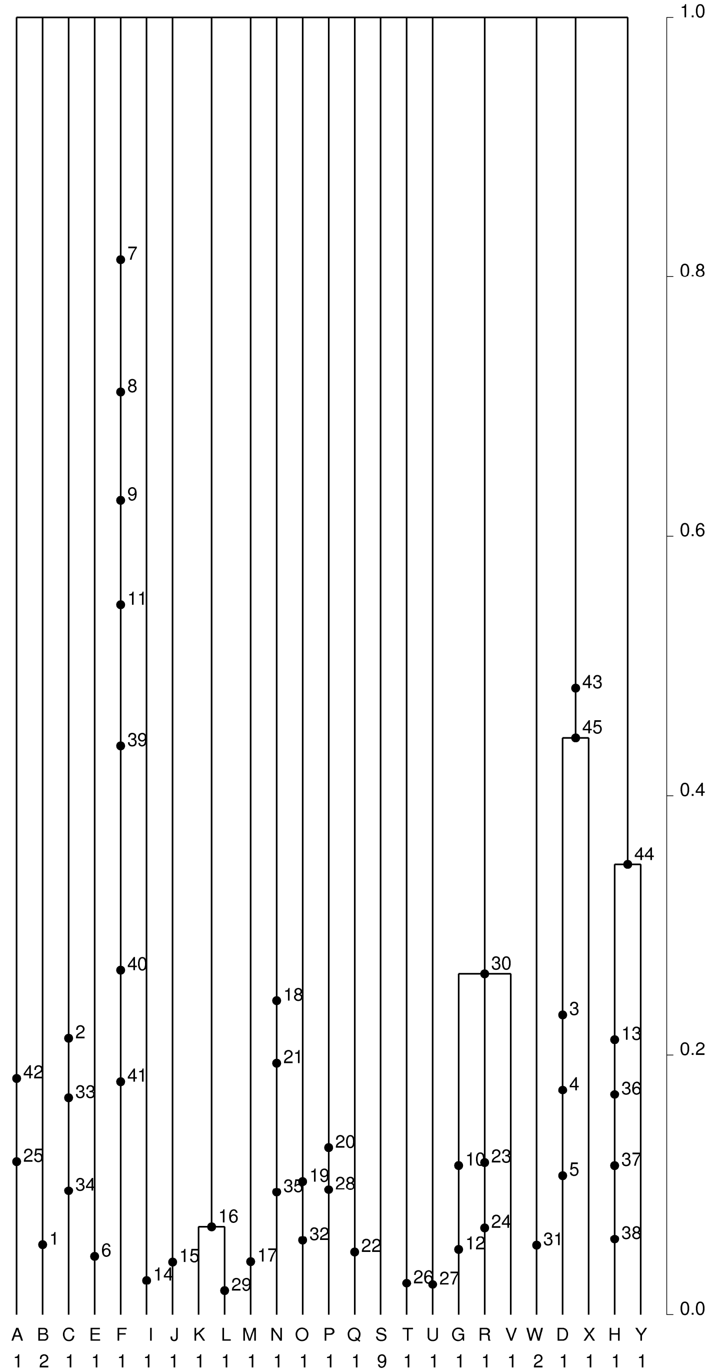


Figure 5.20: Distribution of haplotypes and mutation in Occan

```
Position                        1111111111111111111111111111112222222222222222222222
                  1112335557778990000111112233333344555567788990000222234455555566777
                  99991684682283884677045682903445724134600857396688336900905566916245
                  35793972034550386101249378353232433121370437483756059958432835669630

Site Number                1111111111222222222233333333334444444444555555555566666666
                  12345678901234567890123456789012345678901234567890123456789012345678

Consensus         CAGCAATCGGCTTCGCCTGGAGAGCAGGCCAAGAGTTCCCCTGCGGGCCGCTCCCGGGAATCCCACCG
Site Type         tvvtvtvttvttvvtttvvvttvtvvtttttvtttvtttttttvttttvvtvttttttttttvtvttt
Character Type    ----i-i--------------i------ii-------ii-i-iii---------i-------------
Code Key          ********************************************************************
H1          (  1) .............G........................................T.............
H2          (  1) T.........TC..................................................A..T.A
H3          (  1) .TTT..G................................T..............T.............
H4          (  2) ....C...A.............................................T.............
H5          (  1) ....C.G.................G..................................G........
H6          (  1) ....C.GT..................AA........................................
H7          (  1) ....C.......G...T..........................G........................
H8          (  1) ....C..................A...................G....................C...
H9          (  1) ....C.........A..........................CA.A.......................
H10         (  1) ....C..........T..............G...A.......A.A.......................
H11         (  1) ....C...............G.....................A.A.......................
H12         (  1) ....C......................................G...................T....
H13         (  1) ....C.....................................A.A.......................
H14         (  1) ....C...............................................................
H15         (  1) ....C..............................................G........C.......
H16         (  1) ....C...............................................T...............
H17         (  1) .....GG...................................A.A.......................
H18         (  2) ......G...............................................T.............
H19         (  1) ......G..........G.............................T.....TT.............
H20         (  1) .........T...........A...T......A............AA...T....AAA..........
H21         (  1) .................................G....................T.............
H22         (  1) ..................CT......................A.A.......................
H23         (  1) .....................A...............TT.T...........................
H24         (  1) .....................A....................................G.........
H25         (  1) ......................C..............TT.T...........................
H26         (  1) ............................TT.......TT.T...........................
H27         (  1) ............................TT..........T...........................
H28         (  1) ...............................T..................................T.
H29         (  1) ..........................................A..................T......
H30         (  1) ...................................AC...........GC..................
H31         (  1) ....................................................................
```

Figure 5.21: Compatibility matrix of Occan

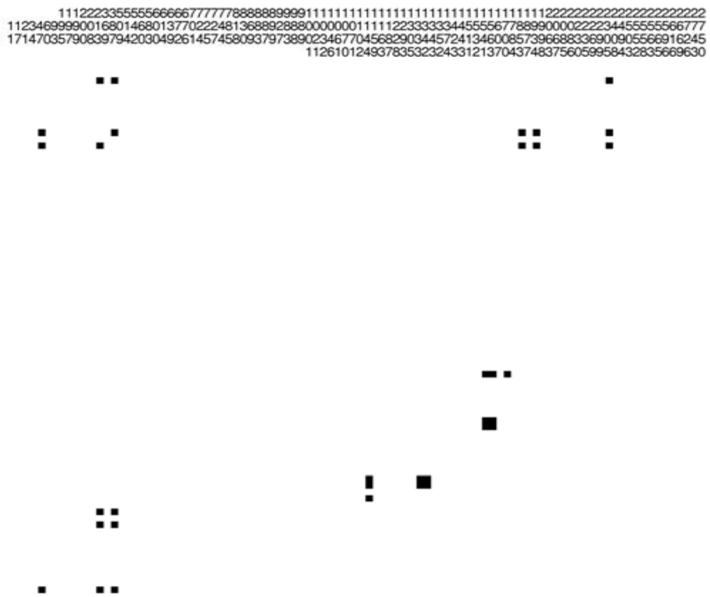
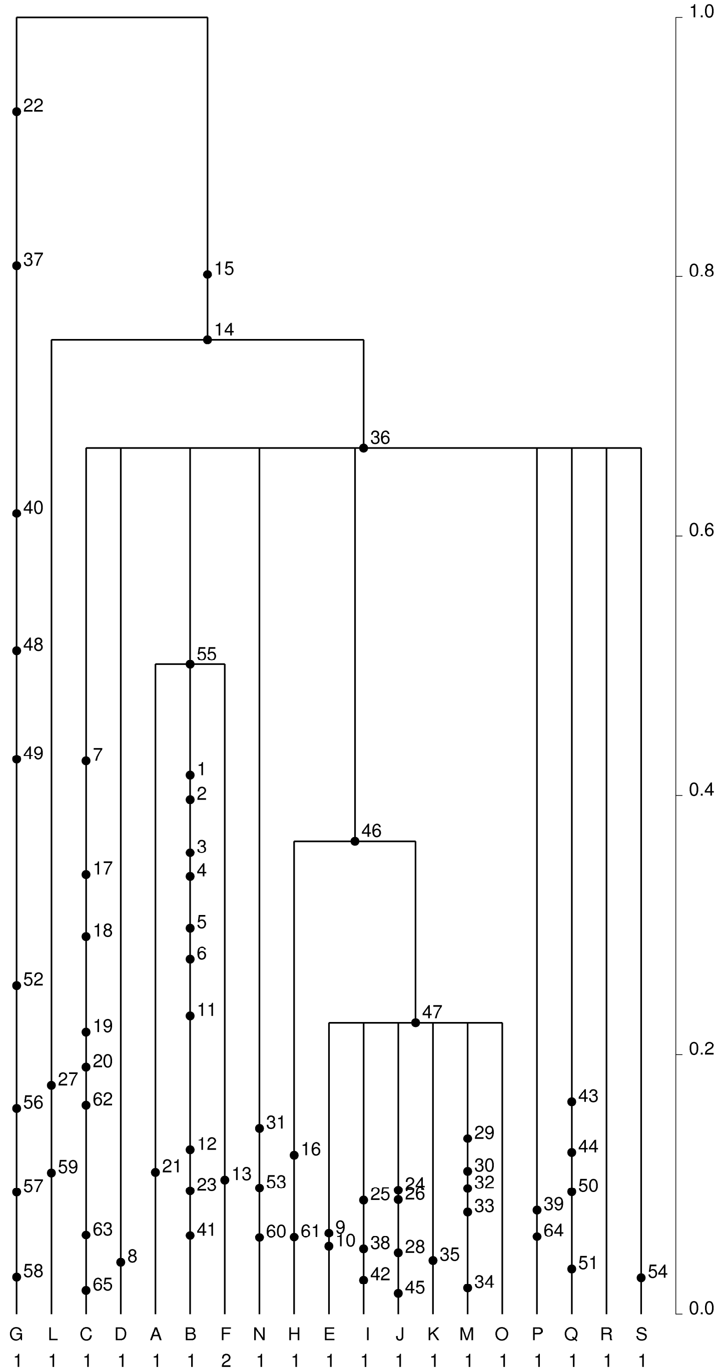


Figure 5.22: Compatibility matrix of Occan Clade4

```
                   11111111112222222222333333333344444444445555555555666666666
 sites    12345678901234567890123456789012345678901234567890123456789012345678
       1  ....................................................................
       2  ....................................................................
       3  ....................................................................
       4  ....................................................................
       5  ..........................................x.x.........x.............
       6  ....................................................................
       7  ....................................................................
       8  ....................................................................
       9  ....................................................................
      10  ....................................................................
      11  ....................................................................
      12  ....................................................................
      13  ....................................................................
      14  ....................................................................
      15  ....................................................................
      16  ....................................................................
      17  ....................................................................
      18  ....................................................................
      19  ....................................................................
      20  ....................................................................
      21  ....................................................................
      22  ....................................................................
      23  ....................................................................
      24  ....................................................................
      25  ....................................................................
      26  ....................................................................
      27  ....................................................................
      28  ....................................................................
      29  ....................................................................
      30  ....................................................................
      31  ....................................................................
      32  ....................................................................
      33  ....................................................................
      34  ....................................................................
      35  ....................................................................
      36  ....................................................................
      37  ....................................................................
      38  ....................................................................
      39  ....................................................................
      40  ....................................................................
      41  ....................................................................
      42  ....................................................................
      43  ....x...............................................................
      44  ....................................................................
      45  ....x...............................................................
      46  ....................................................................
      47  ....................................................................
      48  ....................................................................
      49  ....................................................................
      50  ....................................................................
      51  ....................................................................
      52  ....................................................................
      53  ....................................................................
      54  ....................................................................
      55  ....x...............................................................
      56  ....................................................................
      57  ....................................................................
      58  ....................................................................
      59  ....................................................................
      60  ....................................................................
      61  ....................................................................
      62  ....................................................................
      63  ....................................................................
      64  ....................................................................
      65  ....................................................................
      66  ....................................................................
      67  ....................................................................
      68  ....................................................................
```

Figure 5.23: Coalescent tree of Occan



Chapter 6

Conclusion

In an effort to expand our understanding of filamentous fungal secretion and evolution, we have delved into the genomes of *T. reesei* and *M. grisea*. Through the development of bioinformatics software, annotation and characterization of *T. reesei* EST and BAC libraries, and transposable element coalescent analysis within the genome of *M. grisea* we have enhanced the fungal communities understanding of the genetics of these two fungi.

NuclearBLAST is a bioinformatics tool which provides a user-friendly interface simplifying the process of BLAST analysis, result storage and data mining. This tool contains a BLAST request wizard to simplify the process of dataset analysis. Following the completion of BLAST analysis, NuclearBLAST stores the data in a MySQL database and makes the results viewable via an interactive web interface. Built into this tool are several methods allowing for advanced analyses using the results of BLAST. For example, one can identify the bi-directional best blast hit between two dataset or annotate a dataset with Gene Ontology terms via the results of BLAST to a previously annotated dataset.

Through the use of NuclearBLAST and other bioinformatics tools we were able to functionally annotate an EST dataset and identify numerous genes involved in the protein processing and secretion pathway of *T. reesei* previously unreported in this fungus. An EST dataset was created from mRNA collected under a variety of cellulase inducing conditions. Following sequencing, functional characterization of this dataset containing over 21000 5' sequences collapsed into roughly 5000 unique genes was able to annotate over 2700 of the unique sequence with Gene Ontology terms. These Gene Ontology terms were used to provide an overview of the functional categories present in the dataset. In the dataset, we were able to identify homologs to the majority of *A. niger's* Srg genes as well as a number of homologs to genes involved in protein folding and ornamentation pathways.

We constructed a BAC library containing 9,216 clones with an average insert size of 125 kb which provides a coverage of 28 genome equivalents. These BAC ends were sequenced and annotated with Gene Ontology terms using NuclearBLAST and other publicly available bioinformatics software. We compared the distribution of GO terms between the previously annotated EST dataset and the BAC end sequences and found a similar ratio of annotated genes in each category. Within the BAC end sequences we identified a number of genes not seen in previously sequenced EST datasets. These include genes involved in secretion and biomass degradation which the EST library was specifically targeted to select for showing the utility of survey sequencing genomes with high gene density. There was little evidence found for repetitive sequence in the *T.* reesei genome with the exception of several copies of an element with similarity to the *Podospora anserina* transposon, PAT. Hybridization of thirty-four genes involved in biomass degradation revealed five groups of co-located genes in the genome. The BAC clones were fingerprinted and analyzed using fingerprinted contigs (FPC) software. This analysis resulted in 334 contigs covering 28 megabases of the genome. The assembly of the BACs using FPC was verified by congruence with hybridization results.

The recent sequencing and annotation of *M. grisea* has uncovered the complete set of transposable elements within the genome. The geneaological and coalescent histories of these transposable elements can provide insight into the interplay between genome and mobile element. We analyzed five different transposable elements which are present in the genome in significant numbers using gene geneaologies and the coalescent. This analysis expanded traditional population genetics techniques into the realm of self-replicating genetic elements within a genome. We identified recombinational blocks within several of the transposable elements and were able to run coalescent analysis on each transposable element. The results of the coalescent analysis suggest that a catastrophic genomic event occurred in the past which allowed for the rapid expansion of all transposable elements studied. An event such as the loss of RIP directly through mutation or via a shift in the primary mode of reproduction to asexual, thus bypassing the pre-meiotic stage necessary for RIP activity, could have lead to such an expansion. The loss of genomic control of these transposable elements could represent an acceptance of the action of transposable elements as beneficial to the fungus. Transposable elements active in the genome can assist in the rapid evolution required for *M. grisea* to retain its pathogenicity on its host.

The recent explosion in genomic technologies has provided the filamentous fungal community with an incredible number of tool and resources. This dissertation represents a progression of genomic analysis over the past several years. This work starts with creation of bioinformatics tools to aide in the identification and analysis of an organism’s gene complement, continues with the structural characterization of a genome through BAC analysis and finishes with the evolutionary analysis of genomic elements. These past few years have greatly advance our knowledge of filamentous fungal genetics and evolution and have provided the starting point for a wide variety of future works.