

**“Comparative genomic analysis and functional annotation of
different strains of *Trichophyton rubrum*”**

A Major Project dissertation submitted

In partial fulfillment of the requirement for the degree of

Master of Technology

In

Bioinformatics

Submitted by

Sakshi

(DTU/13/M.TECH/365)

Delhi Technological University, Delhi, India

Under the supervision of

Dr. Asmita Das



Department of Biotechnology
Delhi Technological University
(Formerly Delhi College of Engineering)
Shahbad Daultpur, Main Bawana Road, Delhi-110042, INDIA



CERTIFICATE

This is to certify that the M. Tech. dissertation entitled “**Comparative genomic analysis and functional annotation of different strains of *Trichophyton rubrum***”, submitted by **SAKSHI (DTU/13/M.TECH/365)** in partial fulfillment of the requirement for the award of the degree of Master of Technology, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate’s own work carried out by her under my guidance.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.

Date:

Dr. ASMITA DAS

(Project Mentor)

Department of Bio-Technology
Delhi Technological University

Professor D. Kumar

(Head of Department)

Department of Bio-Technology
Delhi Technological University



IGIB
INSTITUTE OF GENOMICS
& INTEGRATIVE BIOLOGY
Genomics Knowledge Partner

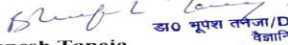
जीनोमिकी और समवेत जीव विज्ञान संस्थान
(वैज्ञानिक तथा औद्योगिक अनुसंधान परिषद)
विश्वविद्यालय परिसर, माल रोड, दिल्ली-110007
Institute of Genomics & Integrative Biology
(COUNCIL OF SCIENTIFIC & INDUSTRIAL RESEARCH)
DELHI UNIVERSITY CAMPUS
MALL ROAD, DELHI-110007, INDIA

To Whom It May Concern

Dated: 26-Jun-2015

This is to certify that the M. Tech. dissertation entitled "**Comparative genomic analysis and functional annotation of different strains of *Trichophyton rubrum***", submitted by **SAKSHI (DTU/13/M.TECH/365)** in partial fulfillment of the requirement for the award of the degree of Master of Engineering, Delhi Technological University (Formerly Delhi College of Engineering, University of Delhi), is an authentic record of the candidate's own work carried out by her from **5-Jan-2015 to 26-Jun-2015 at CSIR-IGIB**.

The information and data enclosed in this dissertation is original and has not been submitted elsewhere for honouring of any other degree.


Dr. Bhupesh Taneja
Senior Scientist
Genome Informatics & Structural Biology Lab
CSIR-IGIB
डा० भूपेश तर्नजा/Dr. BHUPESH TANEJA
वैज्ञानिक/Scientist
जीनोमिकी और समवेत जीव विज्ञान संस्थान
Institute of Genomics & Integrative Biology
माल रोड, दिल्ली-110007/Mall Road, Delhi-110007

TEL : 91 - 11 - 2766 7602, 2766 6156, 2766 6157, 2766 7439 • FAX : 91 - 11 - 2766 7471
TELEFAX : 91 - 11-2766 2407 - 08, 2766 2099 • E-mail : info@igib.res.in • Website : <http://www.igib.res.in>

DECLARATION BY THE CANDIDATE

I declare that the M. Tech. dissertation entitled “**Comparative genomic analysis and functional annotation of different strains of *Trichophyton rubrum*** ”is my own work conducted under the supervision of Dr. Asmita Das and Dr. Bhupesh Taneja.

I further declare to the best of my knowledge that the dissertation does not contain any part of work which has been submitted for the award of any degree in any university.

Sakshi

(2K13/BIO/15)

ACKNOWLEDGEMENT

It is indeed a great privilege for me to express, by way of this note of acknowledgement, for all the noble souls who have helped me in materializing this project.

I would like to thank and express my deep sense of gratitude towards **Dr. Asmita Das** to accept me as a student and allow me to work in CSIR- Institute of Genomic and Integrative Biology, New Delhi., under supervision of Senior Scientist Dr. Bhupesh Taneja. I am really grateful for her invaluable guidance and constant encouragement for the successful completion of my major project.

I would like to thank Head of the Department, Biotechnology, Professor D. Kumar for providing us the necessary opportunities for the completion of our project and other faculty members of my department for their invaluable help and guidance.

I would like to thank and express my deep sense of gratitude towards **Dr. Bhupesh Taneja** to accept me as a trainee and allowed me to work in his Structural Biology Lab, CSIR- Institute of Genomic and Integrative Biology, New Delhi, under his supervision. I am really grateful for his precious time, invaluable guidance and constant encouragement for the successful completion of my major project. The scientific environment provided in the laboratory is highly appreciated.

I would like to thank Ms. Chitra Latka for her persistent support and presence, important suggestions at every step and for her patience to respond to my queries and lab members for their support.

CONTENTS

TOPIC	PAGE NO.
<i>LIST OF FIGURES</i>	1
<i>LIST OF TABLES</i>	2
<i>LIST OF ABBREVIATIONS</i>	3
ABSTRACT	4
1. INTRODUCTION	5
2. REVIEW OF LITERATURE	
2.1 Dermatophytes	7
2.1.1 <i>Trichophyton rubrum</i>	8
2.2 Epidemiology and Ecological grouping of dermatophytes	10
2.2.1 Geophiles	10
2.2.2 Zoophiles	10
2.2.3 Anthropophiles	10
2.3 Antifungal resistance in dermatophytes	11
2.4 Whole genome sequencing	12
2.5 Comparative genomics	12
3. OBJECTIVES	14
4. METHODOLOGY	15
5. RESULTS	23
6. CONCLUSION	43
7. DISCUSSION AND FUTURE PERSPECTIVE	44
8. REFERENCES	46
9. APPENDIX	50

LIST OF FIGURES

Figure 1. *Trichophyton rubrum* a major dermatophyte species.

Figure 2. Dermatophyte infections of the skin.

Figure 3. Classification of protein coding genes of *Trichophyton rubrum* strains based on functions defined by KOG database.

Figure 4. Summary of *Trichophyton rubrum* strains genes assigned with CAZymes functional annotation.

Figure 5. Heatmap of peptidases families in different strains of *Trichophyton rubrum*.

LIST OF TABLES

Table 1. Total number of predicted protein coding genes in different strains of *Trichophyton rubrum*.

Table 2. Repetitive sequences in the genome assembly of different strains of *Trichophyton rubrum*.

Table 3. Carbohydrate active enzymes in different strains of *Trichophyton rubrum*.

Table 4. Number of lipase coding genes in different strains of *Trichophyton rubrum*.

Table 5: Number of genes coding for kinases in different strains of *Trichophyton rubrum*.

Table 6. Number of genes of *Trichophyton rubrum* strains coding for different peptidases.

Table 7. Number of genes of *Trichophyton rubrum* strains coding for different Cytochrome P450 family proteins.

Table 8. Number of pathogenicity related genes in different *Trichophyton rubrum* strains.

Table 9. Number of secretory peptidases, Carbohydrate active enzymes, lipase and kinases in different strains of *Trichophyton rubrum*.

Table 10. Number of unique genes predicted in different *Trichophyton rubrum* strains.

Table 11. Mismatch in different drug resistance genes of different strains of *T.rubrum*.

LIST OF ABBREVIATIONS

DNA: Deoxyribonucleic acid

WGS: Whole genome sequencing

NCBI: National Center for Biotechnology Information

Cazy: Carbohydrate-active-enzymes

CAT: Carbohydrate-active-enzymes Analysis Toolkit

dbCAN: A web server and database for Carbohydrate-active enzyme annotations

BLAST: Basic Local Alignment Search Tool

CYPED: Cytochrome P450 Engineering Database Database

PHI-base: Pathogen-Host Interaction database

KOG: Eukaryotic Orthologous Groups of proteins

WebMGA: Web server for metagenomic analysis

MDR: Multiple drug resistance

AA: Auxiliary activities

CBM: Carbohydrate binding modules

CE: Carbohydrate esterases

GH: Glycoside hydrolases

GT: Glycosyl transferases

PL: Polysaccharide lyases

SINs: Short interspersed elements

LINEs: Long interspersed elements

LTRs: Long terminal repeats

CD-HIT: Clustering database at high identity with tolerance

“Comparative genomic analysis and functional annotation of different strains of *Trichophyton rubrum*”

Sakshi

Delhi Technological University, Delhi, India

ABSTRACT

Dermatophyte species are a group of parasitic filamentous fungi that are known to be closely related and are primary causative agents of skin infections. *Trichophyton rubrum* is an anthropophile dermatophyte which is the most common agent for most of the skin infections like tinea pedis. To provide the genetic basis of fundamental and putatively pathogenicity-related traits of *Trichophyton rubrum* strains, we compared sets of pathogenicity-related proteins, such as secretory peptidases, lipases and proteins involved in skin infection. These dermatophytes are enriched in large number of enzymes i.e. Cytochrome P450 family which are associated with secondary metabolite synthesis. This comparative genomic analysis include finding gene families like Carbohydrate active enzymes, kinases etc. that have the ability to cause disease, genes that act as virulence factors and novel genes in different strains of *Trichophyton rubrum*.

In this study we predicted the large number of protein coding gene that encodes pathogenicity-related proteins in different strains of *Trichophyton rubrum*. We also performed multiple sequence alignment of four genes that show drug resistance to identify any amino acid replacement in drug resistance genes. This study could help in finding the genes responsible for dermatophytosis and in finding the reasons for drug resistance in dermatophyte species.

1. INTRODUCTION

The dermatophytes are a group of fungi that are known to be closely related, these dermatophyte species filamentous fungi and these fungi have the ability to cause infection of the humans or animals skin, nails and hair as this group of filamentous fungi i.e. dermatophyte species have the capacity to enter or invade keratinized tissue. These dermatophyte species secrete proteases, these secretory proteases like keratinases, subtilisins i.e. serine protease, are very important for dermatophyte virulence. Dermatophytes invade and degrade keratinized tissues by the action of all the proteases secreted from these dermatophytes (Monod, 2008; Burmester *et al.*, 2011; Bhatia *et al.*, 2014).

Dermatophyte species are parasitic fungi which are the primary causative agents of infection. Infection caused by these filamentous fungi is known as dermatophytosis. Dermatophytosis is a major public health concern in some geographic regions. These dermatophytic filamentous fungi causing infections belong to three genera i.e. *Trichophyton*, *Microsporum*, and *Epidermophyton*. *Microsporum gypseum*, *Microsporum langeroni*, *Trichophyton rubrum*, *Microsporum canis* and *Trichophyton verrucosum* are some dermatophytes. *Trichophyton rubrum* is a dermatophyte which is human skin fungal pathogen and this fungal species is the most common agent for causing skin infection in humans in the whole world. Dermatophytes are the fungi that cause a variety of skin diseases and infection caused by these fungi include athlete's foot i.e. feet infection, infection of the scalp, eyebrows and eyelashes, and nail infection etc. Dermatophytes are grouped in either anthropophilic fungi, geophilic or zoophilic fungi on the basis of their habitat. Among these filamentous fungi, *Trichophyton rubrum* which is an anthropophile is the most common agent for most of the superficial fungal infection like tinea pedis, tinea capitis, tinea corporis, tinea unguium, tinea cruris and tinea barbae in human keratinized tissue (Elewski B. 1998). Infections caused by dermatophyte species are communicable and infection in humans that are caused by group of filamentous fungi that belongs to anthropophile may be chronic and last for a long period of time. Dermatophyte species are the group of fungi that can cause fungal skin infection in healthy individuals who are immune-competent and these dermatophytes can cause deep infection in individuals who are immune-compromised. It has been seen that 30% to 70% of adults are asymptomatic carriers of these dermatophytic fungi. These fungal infections have a large possibility of relapse (Weitzman *et al.*, 1995; Simpanya 2000; Wang *et al.*, 2006; Martinez *et al.*, 2012; Latka *et al.*, 2015).

Epidemiology of dermatophyte species suggest that the infection caused by these fungi is different from place to place because of the difference in the climatic conditions (Aly, 1994; Bhatia *et al.*, 2014). Genome sequencing of dermatophytes help in identification of unique features in dermatophyte species provide their pathogenicity and help in new therapies development. Whole genome sequence comparison provides details how organisms are related to each other at the genetic level. Genome sequence comparison i.e. comparative genomics reveals features of dermatophyte species that differentiate this group of filamentous fungi from other fungi and from each other. It has been found that there is increase in

dermatophyte drug resistance incidences, this drug resistance of dermatophytes results in infection treatment failure and this prevents complete clearance of fungus (Mukherjee *et al.*, 2003; Martinez *et al.*, 2012).

2. REVIEW OF LITERATURE

2.1 Dermatophytes

Dermatophytes are the fungi belong to the group of filamentous fungi that infect skin, hairs and nails of humans and animals, they are highly specialized pathogenic fungi. These filamentous fungi are the most common cause infection i.e. superficial mycoses which is a fungal infection in humans and animals. Various environmental and physiological conditions contribute to the development of fungal infection. During infection caused by dermatophytes, these fungi infect host and fungi multiply within keratinized host structures like, the epidermal stratum corneum which is the outermost layer of the epidermis, this layer consists of dead cells, multiple protease are secreted by dermatophytes and these secreted proteases are known as virulence determinants (Burmester *et al.*, 2011). Infection caused by dermatophytes are treatable but there is high chance of infection may occur again i.e. reinfection. Dermatophytes have two important properties, one is keratinophilic means dermatophytes have affinity for keratin which is found in hair, skin and nails and other property of dermatophytes is keratinolytic (lysis of keratin), which means dermatophytic fungi digest keratin. Infection caused by dermatophytes i.e. dermatophytosis is generally cutaneous and this dermatophytosis is restricted to the nonliving layers i.e. outermost layers of epidermis. This restriction of dermatophytes is because of inability of the dermatophytes to penetrate the tissues that are deeper and organs of immunocompetent host means host body produce a normal immune response after exposure to an antigen. There is a range of dermatophytic infection i.e. infection caused by dermatophytic fungi, infection may range from mild infection to infection that is severe and this range of dermatophyte infection depend on the reactions of host to the fungus metabolic products, dermatophytes virulence, site of infection and various environmental factors. Dermatophytes belong to the family *Arthrodermataceae* i.e. filamentous fungi group that is closely related to dimorphic fungi.

Dermatophytes are the fungi that cause a variety of skin diseases and infection caused by these fungi include athlete's foot. Athlete's foot is clinically termed as tinea pedis. This infection caused by fungi also includes jock itch or crotch itch groin region dermatophyte infection which is clinically termed as tinea cruris (Diego *et al.*, 2012). Dermatophytes infection named according to the location of the infection, this involve the Latin term for anatomic location of the body where infection occurs and involve the tinea word before the Latin term like tinea capitis which is dermatophytosis of the scalp. There are some clinical terms that are used for dermatophytosis, these terms are tinea barbae for ringworm (infection) of the beard and mustache, tinea capitis is infection of the scalp, eyebrows and eyelashes, tinea corporis ringworm or infection of glabrous skin, tinea, tinea cruris i.e. groin infection caused by dermatophytes, tinea pedis for feet infection and tinea unguium for infection of nails. Infection caused by dermatophytes have significant geographic patterns like in developed countries tinea pedis (athlete's foot) is more common and in developing countries tinea capitis (scalp dermatophyte infection) is more common. Some fungal infections that

are more common in the world are endemic and these infections rarely cause death (Weitzman *et al.*, 1995). There are various species of fungi which are dermatophytic fungi and these dermatophytes are classified in three genera, *Trichophyton*, *Microsporum*, and *Epidermophyton*. Figure 1 shows *Trichophyton rubrum* a major dermatophyte species and Figure 2 shows some dermatophytic skin infection (Achterman *et al.*, 2012).

2.1.1 *Trichophyton rubrum*

There are various dermatophytic filamentous fungi which are known to cause various infections in human and animals. The most commonly observed dermatophyte species in the whole world is *Trichophyton rubrum*, an anthropophile fungal species. *Trichophyton rubrum* causes nail infection and athlete's foot. Chronic dermatophytosis and deep dermatophytosis i.e. deep infection in immune-compromised individuals caused by *Trichophyton rubrum*. A large number of proteins including subtilisins and metallo-endoproteases are secreted by *Trichophyton rubrum*, these secretory proteases are considered as potential virulence factors. Keratinized skin, hair or nails are degraded by *Trichophyton rubrum*, as large number of proteins which are virulence factors that include subtilisins and proteases or peptidases are secreted by *Trichophyton rubrum* (Wang *et al.*, 2006; Monod, 2008; Rivera *et al.*, 2012). Genome sequencing of *Trichophyton rubrum* provide pathogenic life style and help in development of new therapies, vaccines and diagnostics.

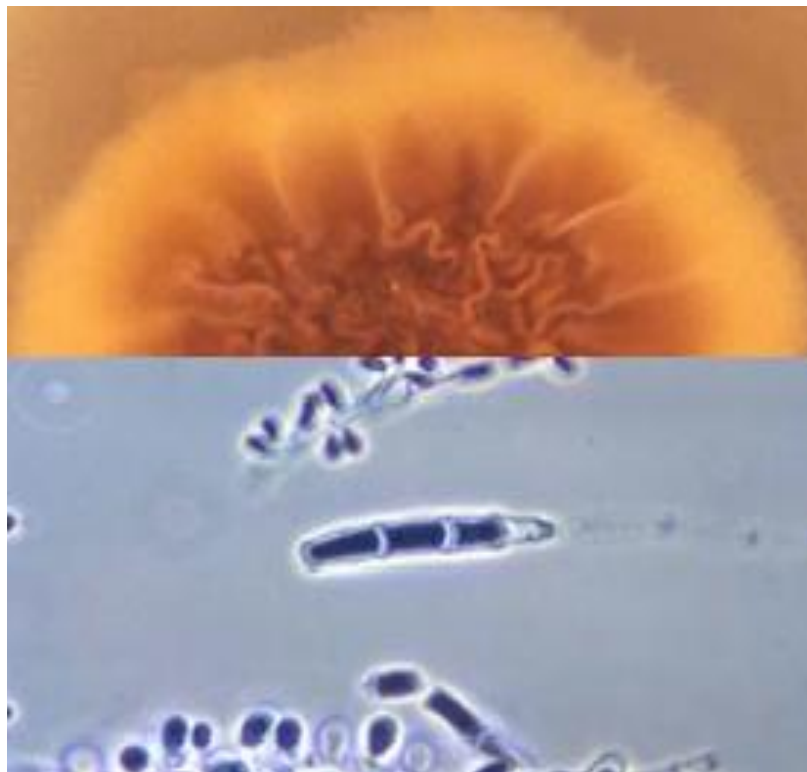


Figure1. *Trichophyton rubrum* a major dermatophyte species. This figure include a semicircle showing fungus growing on a agar plate and microscopic picture of the asexual spores i.e. macroconidia and microconidia (Achterman *et al.* 2012).

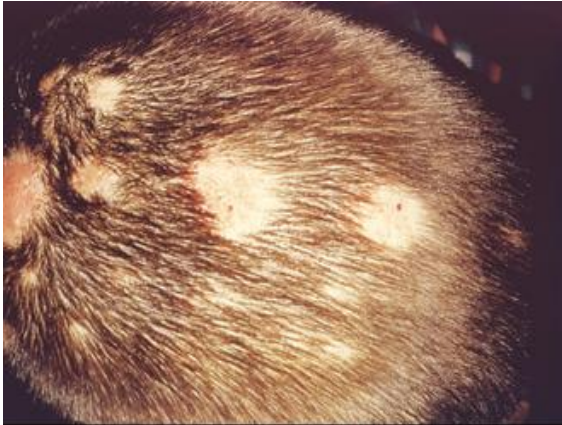


Figure 2A



Figure 2B



Figure 2C



Figure 2D



Figure 2E



Figure 2F

Figure 2. Dermatophyte infections of the skin. 2A: Hair (tinea capitis); 2B: Face (tinea capitis); 2C: Arm (tinea corporis); 2D: Groin (tinea cruris); 2E: Nails (onycomycosis) and 2F: Toe webbing and Foot (tinea pedis).

2.2 Epidemiology and Ecological grouping of dermatophytes

Dermatophytes are group of few filamentous fungi causing communicable diseases, this group of fungi produce keratinases (peoteases). Keratinized structures are degraded by keratinases and the superficial skin is invaded by dermatophytes (Monod, 2008; Burmester *et al.*, 2011; Bhatia *et al.*, 2014).

Dermatophytic fungi are grouped in either anthropophilic fungi which are human associated fungi, zoophilic fungi which are animal associated fungi or geophilic fungi which are soil dwelling fungi, these fungi are grouped on the basis of their habitat.

2.2.1 Geophiles

Geophiles are fungal species that are found in soil as saprophytes i.e. they use dead organic matter to obtain nutrients. Some of the geophiles have capacity of causing infection i.e. ringworm in some animal species and man. *M.icrosporium gypseum* and *Microsporium cookie* are examples of geophiles.

2.2.2 Zoophiles

Zoophiles are the fungal species that basically cause infection in animals i.e. animal pathogens, zoophilic dematophytes prefer a singal animal host or their host range is very limited. These dermatophytes do not grow actively as saprophytes but they survive in an inactive state i.e. dormant state on materials of animal that are contaminated. Some zoophiles are *Microsporium canis*, *Trichophyton verrucosum*, *Trichophyton mentagrophytes*, these species cause infection (ringworm) in animals and also cause human infection.

2.2.3 Anthropophiles

Anthropophiles are the fungal species that are human pathogens, but some anthropophile cause infection in animals, it has been found that *Trichophton rubrum* caused infection in a dog.

Microsporium audouinii, *Microsporium langeroni*, *Trichophton rubrum* are some anthropophile.

All the dermatophytes belong to three genera i.e. *Epidermophyton*, *Microsporium* and *Trichophyton* have different pathogenicity. These species have different invading capacity (Simpanya, M. 2000). Geophilic fungal species are thought to be ancestral to the pathogenic dermatophytes (Weitzman *et al.*, 1995; Simpanya 2000).

Anthropophilic fungal species which are human associated are responsible for large number of human infection , however fungal species belong to all three dermatophyte groups are associated with diseases. Infection in humans caused by anthropophile can be chronic that

may last for a long period of time, while infections that are caused by fungal species belong to zoophiles and geophiles are associated with self healing. Dermatophytic fungal taxa recognition is very important for clinical studies, this include identification of dermatophyte species that is related to epidemiological studies of dermatophytes.

Epidemiological studies of dermatophytes are important in different dermatophytosis and infection control. The major human infections are caused by anthropophilic species, pets infected with ringworm or infected with dermatophytes transmit this dermatophytic infection to their owner, cats having infection or carrying dermatophyte species are considered to transmit infection to humans in many eastern and southern European countries (Seebacher *et al.*, 2008).

Trichophyton tonsurans and *Microsporum canis* are important agents of tinea capitis (scalp dermatophyte infection) in North America, *Trichophyton tonsurans* caused a progressive and continent wide epidemic, is usually acquired from infected humans. *Microsporum canis* is limited human to human transfer, is acquired from infected animals i.e. cats or dogs. In many rural areas of the world and in some parts of Europe, South America *Microsporum canis* is the predominant agent of tinea capitis. In general Tinea capitis is a condition of scalp dermatophyte infection most commonly found in children. *Trichophyton violaceum* is endemic and common cause of dermatophytosis in certain parts of Eastern Europe, South America, Africa and Asia but not in North America. It has been found that *Trichophyton rubrum* an anthropophile, is the most common cause of various infection i.e. dermatophytosis in the whole world. Dermatophyte strain responsible for infection and the site of infection vary from country to country or region to region (Achterman *et al.*, 2012). It is found that there is decline in the incidence of tinea capitis in developed nations, but tinea pedis and tinea unguium are more common in developed nations (Aly, R. 1994). Tinea infections i.e. dermatophytosis are common worldwide but these infections are more common in tropic regions and geographical areas having higher humidity. In india hot and humid climate is there and this climate makes dermatophytic infection like fungal skin infection very common in India. The distribution and frequency of these fungi is different from place to place and this difference depends on the climatic conditions (Bhatia *et al.*, 2014).

Infection caused by dermatophytes are communicable these infection easily transmitted. There is increase in the incidences of drug resistance of fungal strains that result in fungal treatment failure and this prevents fungus complete clearance (Mukherjee *et al.*, 2003).

2.3 Antifungal Resistance in Dermatophytes

The rate of recurrence of dermatophyte infection is very high. Currently the reason for recurrence of these infections is not known, it could be due to insufficient clearing of the dermatophytic fungi or it could be the new infection. It has been found that to treat dermatophytic infection, over US\$500 million per year is spent worldwide (Achterman *et al.*, 2012). Commonly antifungals are used to treat tinea pedis and other dermatophytic

infections; this might lead to drug resistance in dermatophytes. It has been seen that drug resistance among dermatophyte species is rare, some dermatophytic fungi show drug resistance as it has been found that a single amino acid change in the target gene sequence is responsible for resistance to terbinafine in the clinical isolate from onychomycosis suffering patient (Mukherjee *et al.*, 2003).

2.4 Whole genome sequencing

A laboratory process that is used to determine an organism's genome complete DNA sequence i.e. organism's complete DNA make-up at a single time is known as WGS i.e. whole genome sequencing, this sequencing process is also known as full genome sequencing, entire genome sequencing or complete genome sequencing. Genetic variation of an individual or variation within the species and between the species is better understood by Whole genome sequencing process. An organism's genome sequence would represent the complete nucleotide base sequence for all chromosomes; this process is used for determining the ordered nucleotide sequences of entire genomes of organisms (Ekblom *et al.*, 2014). There are various methodologies for whole genome sequencing. Genome sequencing includes method or technology used for determination of the four bases (adenine, guanine, cytosine and thymine) order a DNA strand. It has applications in biological research and discovery, medical diagnosis, biotechnology, forensic biology and virology. Whole genome sequencing is used to identify the genetic basis of both normal and pathological cellular mechanisms. Genome sequences of pathogens provide understanding how pathogens cause diseases.

2.5 Comparative genomics

Genomics is the study of genome (entire genetic compliment of an organism) structure and function; it is the study of all genes that are present in an organism. Comparative genomic analysis is the genomic comparison study across species, this include structural genomics (genetic and physical mapping of genomes) and functional genomics (analysis of gene functions). Genetic features like DNA sequences, number of genes, gene order or regulatory sequences of different species or different strains of same species are compared in comparative genomics. This study is used to find basic similarities and differences between species and among species, and used to find evolutionary relationship of species. Whole genome sequence comparison provides details how organisms are related to each other at the genetic level.

Comparative genomic study of dermatophytic fungi reveals various features that differentiate this group of dermatophytic fungi from other fungi and from each other. Comparative genomic analysis of dermatophytes identified if there is any specific change in any functional category that is common to all genomes of dermatophytes, which would help in suggesting the candidate genes having role in dermatophytosis and help in the new therapies development (Martinez *et al.*, 2012). Different gene families in dermatophytes are

responsible for infection, dermatophytes secrete different proteases (subtilisin protease), keratinases, lipases, kinases and other gene families (carbohydrate active enzymes) that are responsible for keratinized tissue degradation. Secondary metabolites that are produced by these dermatophyte species are also responsible for infection.

Genome sequence analysis of dermatophytes explains that the number of proteases/proteinases which are necessary for keratin degradation increases in various dermatophytes. Comparative genomic studies show that there is an increased number of proteases in dermatophytes as compared to closely related fungi. The genome sequences of dermatophytes used with various genetic tools to study pathogenicity or virulence of dermatophytes. To find which gene products are important for virulence, like which genes are coding for proteinases, gene sequence information of dermatophytes can be used. This genomic study of dermatophytes will explain how dermatophytic fungi cause infection, how these dermatophyte species interact with human or animal keratinized tissues or cells, with genome sequence analysis of dermatophytes we can know the fungal gene products involved in infection and this could help in development of new better treatments. In this study we present the comparative genomic analysis of whole genomes of 11 different strains of *Trichophyton rubrum*, to find the genomic basis of phenotypic variation in these 11 different strains i.e. how these strains are similar or different from each other on their genetic basis. This comparative genomic analysis includes finding gene families like Cazy, kinases etc. that have the ability to cause disease, genes that act as virulence factors and novel genes in different strains of *T.rubrum*.

3. OBJECTIVES

1. Prediction of protein coding genes in different strains of *Trichophyton rubrum* and manual curation of predicted protein coding gene sequences against refseq *Trichophyton rubrum* CBS 118892.
2. Functional annotation of predicted genes of different strains of *Trichophyton rubrum*.
3. Prediction and analysis of protein families in different strains of *Trichophyton rubrum*.

4. METHODOLOGY

1. Data availability

There are 11 different strains of *Trichophyton rubrum* which were used for comparative genomics study and for functional annotation. All genome assemblies for *Trichophyton rubrum* strains were accessed through NCBI (“<http://www.ncbi.nlm.nih.gov/>”).

11 different strains of *Trichophyton rubrum* are :

CBS 118892, MR850, CBS 100081, CBS 288.86, CBS 289.86, CBS 202.88, MR1448, D6, MR1459, IGIB-SBL-C11 and CBS 735.88

2. Gene prediction and correction of protein coding gene sequences

Sequencing of *Trichophyton rubrum* IGIB-SBL-C11 strain was done in Structural Biology lab, IGIB. Sequences of other strains were downloaded from NCBI and these sequences were used for protein coding gene prediction. The protein coding genes were predicted using WebAUGUSTUS. WebAUGUSTUS is available at “<http://bioinf.unigreifswald.de/webaugustus>” (Hoff *et al.*, 2013). *Trichophyton rubrum* CBS 118892 strain is the refseq.

The sequences for protein coding genes of 10 different *Trichophyton rubrum* strains (MR850, CBS 100081, CBS 288.86, CBS 289.86, CBS 202.88, MR1448, D6, MR1459, IGIB-SBL-C11 and CBS 735.88) were corrected against protein coding gene sequences of refseq *Trichophyton rubrum* CBS 118892 strain by using blastp (protein-protein BLAST).

For blastp the refseq *Trichophyton rubrum* CBS 118892 strain protein coding gene sequences were converted into a blast database by using makeblastdb command.

```
makeblastdb -in CBS118892_database -dbtype prot -out CBS118892_db
```

Run blastp for all 10 strains of *Trichophyton rubrum* against refseq CBS 118892 strain with e value =0.00001. Run local blastp using blastp command.

```
blastp -db CBS118892_db -word_size 7 -query AOKX_new.fas -out AOKXvsCBS -evalue 0.00001 -outfmt '7 qseqid qlen sseqid pident length mismatch gapopen slen qstart qend sstart send evlue bitscore qcovs staxids sscinames stitle' -max_target_seqs 10
```

Alternative transcripts in other 10 strains were deleted by comparing protein coding genes of 10 strains with protein coding genes of refseq CBS118892. Now these corrected protein coding gene sequence FASTA files of all 10 strains were taken for further analysis.

3. Functional annotation of protein coding genes of *Trichophyton rubrum* strains

Total protein coding genes predicted from WebAUGUSTUS for all 11 strains of *T. rubrum* were annotated using Function annotation KOG a web server of WebMGA (web server for metagenomic analysis) server, available at “<http://weizhong-lab.ucsd.edu/metagenomic-analysis/server/kog/>” (Wu *et al.*, 2011). Function annotation by KOG is done by using RPSBLAST program on KOG database i.e. Eukaryotic Orthologous Groups of proteins.

The protein coding gene sequence FASTA file was uploaded with search parameters: e value = 0.00001.

4. Repeat content in the Genome of different strains of *Trichophyton rubrum*

Repetitive sequences in the Genome assembly of different strains of *T. rubrum* were predicted by online tool RepeatMasker. This online tool is available at “<http://www.repeatmasker.org/cgi-bin/WEBRepeatMasker>” (Tarailo-Graovac *et al.*, 2009).

DNA sequences in FASTA format for each strain of *T. rubrum* were downloaded from NCBI database and DNA sequence file of *T. rubrum* strain was uploaded on RepeatMasker Web Server. Search engine , speed/sensitivity parameter, DNA source species were selected and the species name was entered in the text box i.e. *Trichophyton rubrum* for prediction of repetitive sequences in genome assembly, this was done for all DNA sequences of 11 strains of *T. rubrum* for finding repeat content in all 11 strains.

5. Protein family classification

To find carbohydrate active enzymes, lipases, peptidases, kinases, cytochrome P450 family genes and pathogenicity related genes in different strains of *Trichophyton rubrum*

Genes predicted from WebAUGUSTUS were used for protein family classification.

A. Carbohydrate active enzymes

Carbohydrate active enzymes in all 11 different strains of *Trichophyton rubrum* were predicted using two web servers, CAT (CAZYmes Analysis Toolkit), which is available at “<http://mothra.ornl.gov/cgi-bin/cat/cat.cgi>” (Park *et al.*, 2010) and dbCAN which is a web server and database for Carbohydrate active enzyme annotation, which is available at “<http://csbl.bmb.uga.edu/dbCAN/blast.php>” (Yin *et al.*, 2012). Both CAT and dbCAN

perform blastp search of submitted protein sequences in FASTA format against Cazy database, a database for all carbohydrate active enzymes.

Protein coding gene sequences of *T. rubrum* strain in a FASTA file was uploaded on CAT web interface, we used sequence based annotation of CAT, selected e-value threshold 0.00001 for blastp search. This was done for all 11 different strains of *Trichophyton rubrum*.

Similarly protein coding gene sequences of *T. rubrum* strain in a FASTA file was uploaded on dbCAN web interface run blastp search. This was done for all 11 different strains of *Trichophyton rubrum* and all the positive hits obtained from blastp search using CAT and dbCAN were manually examined.

B. Lipases

The Lipase Engineering Database was used to predict genes that have function of lipases, esterases and related proteins in *Trichophyton rubrum* strains. Lipase Engineering Database is a database having all the information regarding sequences, structures and functions of lipases, different esterases and other proteins that are related to lipases or esterases.

Lipases, esterases and related proteins in *Trichophyton rubrum* strains were predicted by blastp search against Lipase Engineering Database “<http://www.led.uni-stuttgart.de/>” (Fischer *et al.*, 2003). For blastp search the Lipase Engineering Database was downloaded and converted into a blast database by using makeblastdb command.

```
makeblastdb -in lipase_database -dbtype prot -out lipase_db
```

Run blastp, *Trichophyton rubrum* strain as query against Lipase Engineering Database i.e. lipase_db (search database) with e value =0.00001.

Run blastp by using blastp command.

```
blastp -db lipase_db -word_size 7 -query CBS_new.fas -out CBSvsLipase -evaluate 0.00001 -outfmt '7 qseqid qlen sseqid pident length mismatch gapopen slen qstart qend sstart send evlue bitscore qcovs staxids sscinames stitle' -max_target_seqs 10
```

This was done for all 11 different strains of *Trichophyton rubrum*, all the positive hits obtained from blastp search were manually examined and each *Trichophyton rubrum* strain was used as a different query for search database i.e. lipase_db.

C. Kinases

For prediction of genes coding for kinases in all 11 strains of *Trichophyton rubrum* KinBase The kinase Database was used. KinBase is a database for kinase proteins of different species like human, mouse, Bakers yeast etc. KinBase is available at “<http://kinase.com/kinbase/>”.

Bakers yeast kinase protein FASTA file was used for prediction of kinases in different strains of *Trichophyton rubrum* by using blastp in which *Trichophyton rubrum* strain was taken as query and Bakers yeast kinase protein was taken as search database. For blastp search Bakers yeast kinase protein FASTA file was downloaded from KinBase and converted into a blast database using makeblastdb command for blastp search.

```
makeblastdb -in Bakers_Yeast_kinase_protein.fasta -dbtype prot -out kinase_db
```

Run blastp by using blastp command with e value =0.00001.

```
blastp -db kinase_db -word_size 7 -query CBS_new.fas -out CBSvsKinase -evaluate 0.00001 -outfmt '7 qseqid qlen sseqid pident length mismatch gapopen slen qstart qend sstart send evlue bitscore qcovs staxids sscinames stitle' -max_target_seqs 10
```

This was done for all 11 different strains of *Trichophyton rubrum*, all the positive hits obtained from blastp search were manually examined and each *T. rubrum* strain was used as a different query for search database i.e. kinase_db.

D. Peptidases

The MEROPS Database was used to predict genes coding for peptidases i.e proteases, proteinases and proteolytic enzymes in *Trichophyton rubrum* strains. MEROPS database is a database for peptidase available at “<http://merops.sanger.ac.uk/>” (Rawlings *et al.*, 2003). Genes coding for peptidases in different *Trichophyton rubrum* strains were predicted by blastp search against MEROPS database. For blastp search the MEROPS database was downloaded and converted into a blast database by using makeblastdb command.

```
makeblastdb -in merops_database -dbtype prot -out peptidase_db
```

Run blastp with *Trichophyton rubrum* strain as query against MEROPS database i.e. peptidase_db (search database) with e value =0.00001.

Run blastp by using blastp command

```
blastp -db peptidase_db -word_size 7 -query CBS_new.fas -out CBSvsPeptidase -evaluate 0.00001 -outfmt '7 qseqid qlen sseqid pident length mismatch gapopen slen qstart qend sstart send evlue bitscore qcovs staxids sscinames stitle' -max_target_seqs 10
```

This was done for all 11 different strains of *Trichophyton rubrum*, all the positive hits obtained from blastp search were manually examined and each *T. rubrum* strain was used as a different query for search database i.e. peptidase_db. A heatmap for all peptidase family was made using R.

E. Cytochrome P450 family proteins.

The CYPED i.e. Cytochrome P450 Engineering Database was used to predict Cytochrome P450 family genes in *T. rubrum* strains. CYPED is database for genes that are related to Cytochrome P450 family, CYPED is available at “<http://www.cyped.uni-stuttgart.de/>” (Fischer *et al.*, 2007).

CYPED Database was used for downloading cytochrome P450 family protein sequences. These downloaded cytochrome P450 family protein sequences were used for finding Cytochrome P450 family genes in *T. rubrum* strains by using blastp search. For blastp search Cytochrome P450 family protein sequences obtained from CYPED were converted into a blast database by using makeblastdb command.

```
makeblastdb -in all_cytP450_seq -dbtype prot -out CytP450_db
```

Run blastp, *Trichophyton rubrum* strain as query against CYPED i.e. CytP450_db (search database) with e value =0.00001

Run blastp by using blastp command

```
blastp -db CytP450_db -word_size 7 -query CBS_new.fas -out CBSvsCytP450 -evalue 0.00001 -outfmt '7 qseqid qlen sseqid pident length mismatch gapopen slen qstart qend sstart send evlue bitscore qcovs staxids sscinames stitle' -max_target_seqs 10
```

This was done for all 11 different strains of *Trichophyton rubrum*, all the positive hits obtained from blastp search were manually examined and each *T. rubrum* strain was used as a different query for search database i.e. CytP450_db.

F. Pathogenicity

Pathogenicity- related genes in different strains of *Trichophyton rubrum* were predicted by using PHI database i.e. Pathogen-host interaction database which is available at “<http://www.phi-base.org/>” (Winnenburg *et al.*, 2006).

Whole genome blast analysis against PHI database was done to find pathogenicity-related genes in *T. rubrum* strains.

PHI database was downloaded for blastp search. For blastp search for query *T. rubrum* strains PHI database was converted into a blast database by using makeblastdb command.

```
makeblastdb -in PHIdatabasce -dbtype prot -out PHI_db
```

Run blastp by using blastp command, *Trichophyton rubrum* strain as query against PHI database i.e. PHI_db (search database) with e value =0.00001

```
blastp -db PHI_db -word_size 7 -query CBS_new.fas -out CBSvsPHI -evaluate 0.00001 -  
outfmt '7 qseqid qlen sseqid pident length mismatch gapopen slen qstart qend sstart send  
evlue bitscore qcovs staxids sscinames stitle' -max_target_seqs 10
```

This was done for all 11 different strains of *Trichophyton rubrum*, all the positive hits obtained from blastp search were manually examined and each *T. rubrum* strain was used as a different query for search database i.e. PHI_db.

6. Secretory Peptidases, Carbohydrate active enzymes, Kinases, Lipases and Virulence genes in different strains of *Trichophyton rubrum*

Prediction of putative secretory lipases, peptidases, virulence genes, kinases, and carbohydrate active enzymes was done using SignalP server.

SignalP server is available at “<http://www.cbs.dtu.dk/services/SignalP/>” (Petersen *et al.*, 2011).

Gene sequences for the genes that code for lipases, peptidases, virulence genes, kinases and carbohydrate active enzymes were extracted from protein sequence fasta file using Fasta Sequence Extractor.

Fasta Sequence Extractor:
http://usersbirc.au.dk/biopv/php/fabox/fasta_extractor.php#formtools.

Protein sequence fasta file for the genes that code for lipases, peptidases, virulence genes, kinases and carbohydrate active enzymes were used to find secretory genes in different protein families.

Protein sequence fasta files for the genes that code for different protein families were uploaded on SignalP 4.1 Server with default parameters.

7. Unique genes in *T. rubrum* strains and functional annotation of unique genes

Unique genes in the genome of different strains of *T. rubrum* were predicted using CD-HIT-2D server (Clustering database at high identity with tolerance) “http://weizhong-lab.ucsd.edu/cdhit_suite/cgi-bin/index.cgi?cmd=cd-hit-2d” (Huang *et al.*, 2010), CD-HIT-2D server compares two protein dataset.

All vs all comparison of all gene sequences of all 11 strains of *T. rubrum* was done using CD-HIT-2D server with sequence identity cut-off 0.6 means 60% identity.

Blast2GO was used for functional annotation of unique (novel) gene sequences (Conesa *et al.*, 2005).

Unique gene sequences were uploaded on blast2GO and run Blast, after blast was completed interproscan, mapping and annotation was done.

8. Multiple sequence alignment of genes that show drug resistance in all 11 strains of *Trichophyton rubrum*

Squalene epoxidase gene, TratrD gene (gene responsible for multiple drug resistance), ABC transporter i.e. TERG_02186 and cytochrome P450 51 i.e. TERG_01703 are drug resistance genes in *Trichophyton rubrum* CBS 118892 (refseq).

Gene sequences for all these drug resistance genes in *Trichophyton rubrum* CBS 118892 (refseq) were obtained from NCBI database.

Protein sequence (Amino acid gene sequence) and corresponding nucleotide gene sequences (CDS) for all these drug resistance genes in *Trichophyton rubrum* CBS 118892 (refseq) were downloaded from NCBI database.

Protein sequences for refseq *Trichophyton rubrum* CBS 118892

Squalene epoxidase gene: >gi|327298303|ref|XP_003233845.1| squalene epoxidase [*Trichophyton rubrum* CBS 118892]

TratrD gene (multidrug resistance protein): >gi|327292416|ref|XP_003230907.1| multidrug resistance protein [*Trichophyton rubrum* CBS 118892]

TERG_02186 (ABC transporter) : >gi|327305547|ref|XP_003237465.1| ABC transporter [*Trichophyton rubrum* CBS 118892]

TERG_01703 (Cytochrome P450 51) : >gi|327304577|ref|XP_003236980.1| cytochrome P450 51 [*Trichophyton rubrum* CBS 118892]

Nucleotide gene sequences (CDS) for refseq *Trichophyton rubrum* CBS 118892

Squalene epoxidase gene: >gi|327298302|ref|XM_003233797.1| *Trichophyton rubrum* CBS 118892 squalene epoxidase (TERG_05717) mRNA, complete cds

TratrD gene (multidrug resistance protein): >gi|327292415|ref|XM_003230859.1| *Trichophyton rubrum* CBS 118892 multidrug resistance protein (TERG_08613) mRNA, complete cds

TERG_02186 (ABC transporter) : >gi|327305546|ref|XM_003237417.1| *Trichophyton rubrum* CBS 118892 ABC transporter (TERG_02186) mRNA, complete cds

TERG_01703 (Cytochrome P450 51) : >gi|327304576|ref|XM_003236932.1| *Trichophyton rubrum* CBS 118892 cytochrome P450 51 (TERG_01703) mRNA, complete cds

These drug resistance gene sequences in other 10 strains of *T. rubrum* were predicted by using blastp search against drug resistance gene sequences of refseq CBS 118892.

Multiple sequence alignment for protein sequences and nucleotide sequences of drug resistance genes was done using ClustalW (multiple sequence alignment tool) which is online available at “<http://www.ebi.ac.uk/Tools/msa/clustalw2/>” (Thompson *et al.*, 1994).

The protein sequence FASTA file containing drug resistance genes of all 11 strains was uploaded on ClustalW2 for multiple sequence alignment. This was done for all 4 drug resistance genes and for nucleotide sequences of drug resistance genes.

5. RESULTS

The total number of predicted protein coding gene different strains of *Trichophyton rubrum*.

Total protein coding genes in all 11 strains *Trichophyton rubrum* were predicted by using WebAUGUSTUS. *Trichophyton rubrum* CBS 118892 strain is the refseq, gene sequences of other strain were curated against CBS 118892 and alternate transcripts in all other 10 strains were deleted by comparing gene sequences of 10 strains with gene sequences of refseq CBS118892. Alternate transcripts having less query coverage and less percent identity were deleted. Total number of predicted protein coding genes was found different in all different *Trichophyton rubrum* strains.

<i>T.rubrum strain</i>	Total gene number
CBS 118892	8820
MR850	8331
CBS 100081	8313
CBS 288.86	8331
CBS 289.86	8292
CBS 202.88	8504
MR1448	8326
D6	8333
MR1459	8344
IGIB-SBL-CI1	8265
CBS 735.88	8308

Table 1. Total number of predicted protein coding genes in different strains of *Trichophyton rubrum*. Total number of protein coding genes in all different strains of *Trichophyton rubrum* predicted by WebAUGUSTUS and curated against CBS 118892.

Classification of the protein coding sequences of different strains of *T. rubrum* based on functions defined by KOG database

Different numbers of protein coding genes were predicted for all strains. Predicted protein coding gene sequences of all strains were grouped into different functional categories defined by clusters of Eukaryotic Orthologous Groups of proteins (KOG). Predicted protein coding gene sequences were grouped in functional categories based on their functions like Information Storage and Processing, Cellular Processes and Signaling, genes having functions of metabolism are grouped in Metabolism category, genes having multiple functions are grouped into Multiple families group, Poorly characterized (gene sequences containing conserved domains of uncharacterized function).

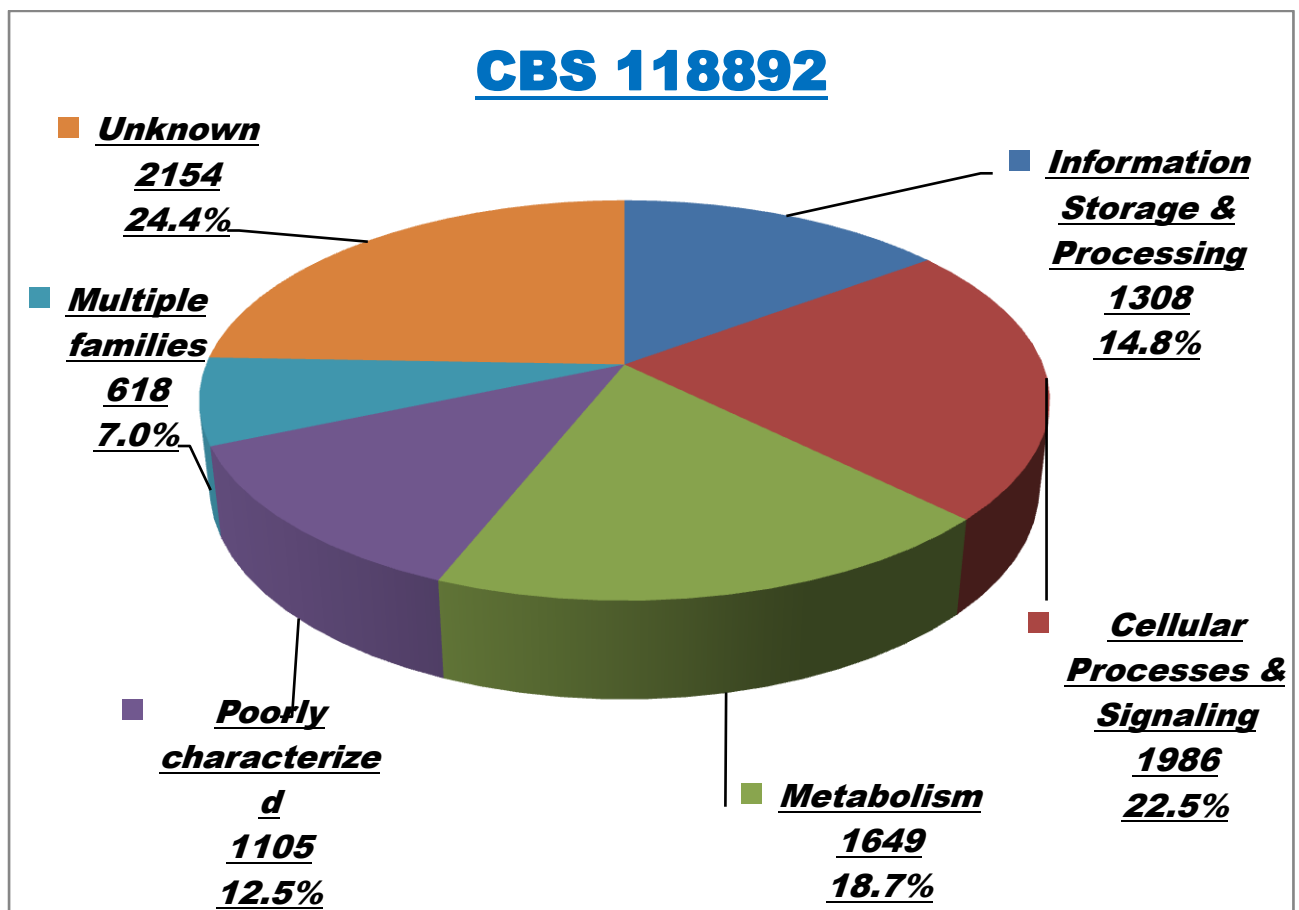


Figure 3A. Distribution of predicted protein coding gene sequences of *T. rubrum* strain CBS 118892

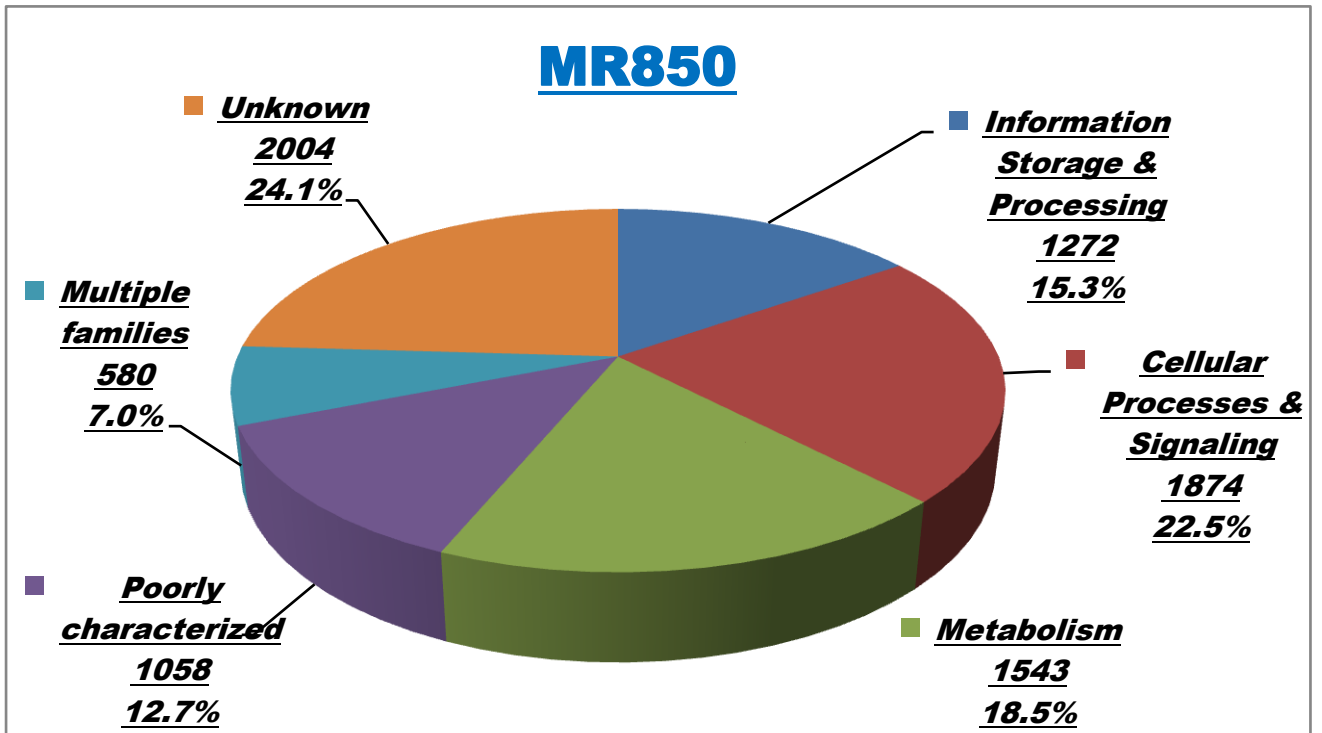


Figure3B. Distribution of predicted protein coding gene sequences of *T. rubrum* strain MR850.

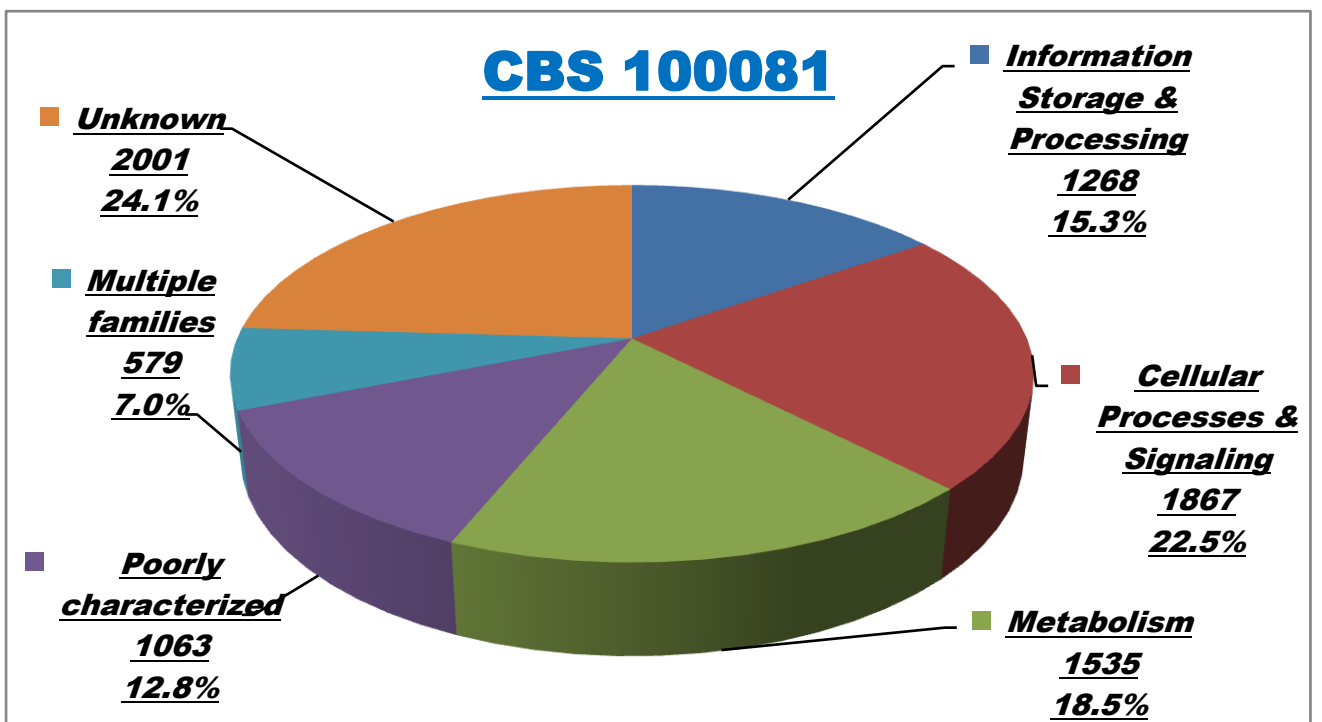


Figure3C. Distribution of predicted protein coding gene sequences of *T. rubrum* strain CBS 100081.

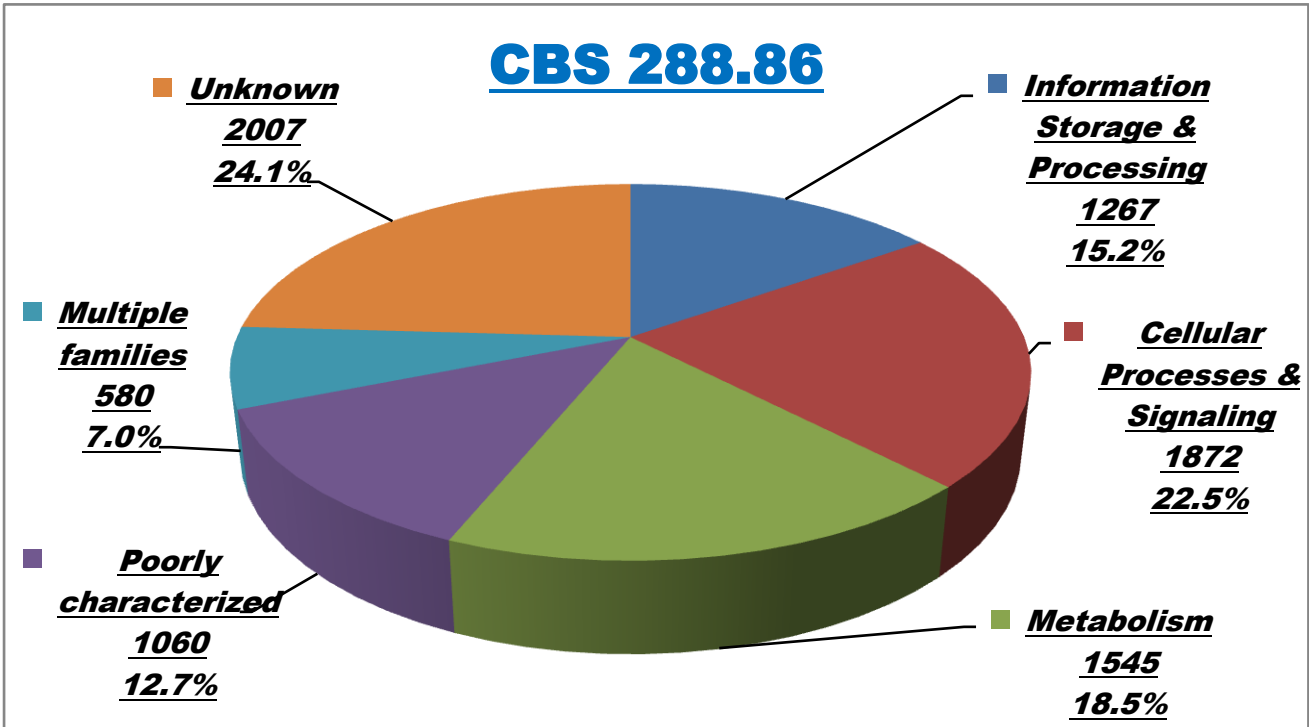


Figure3D. Distribution of predicted protein coding gene sequences of *T. rubrum* strain CBS 288.86.

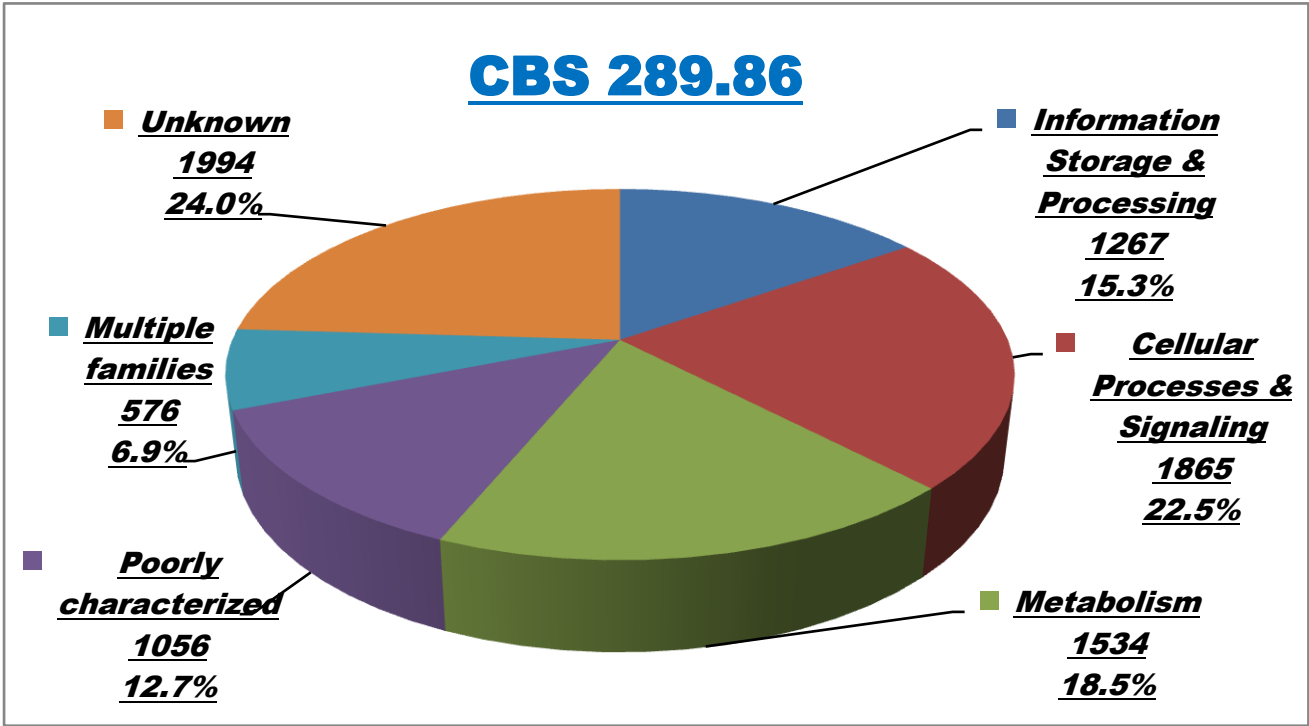


Figure3E. Distribution of predicted protein coding gene sequences of *T. rubrum* strain CBS 289.86.

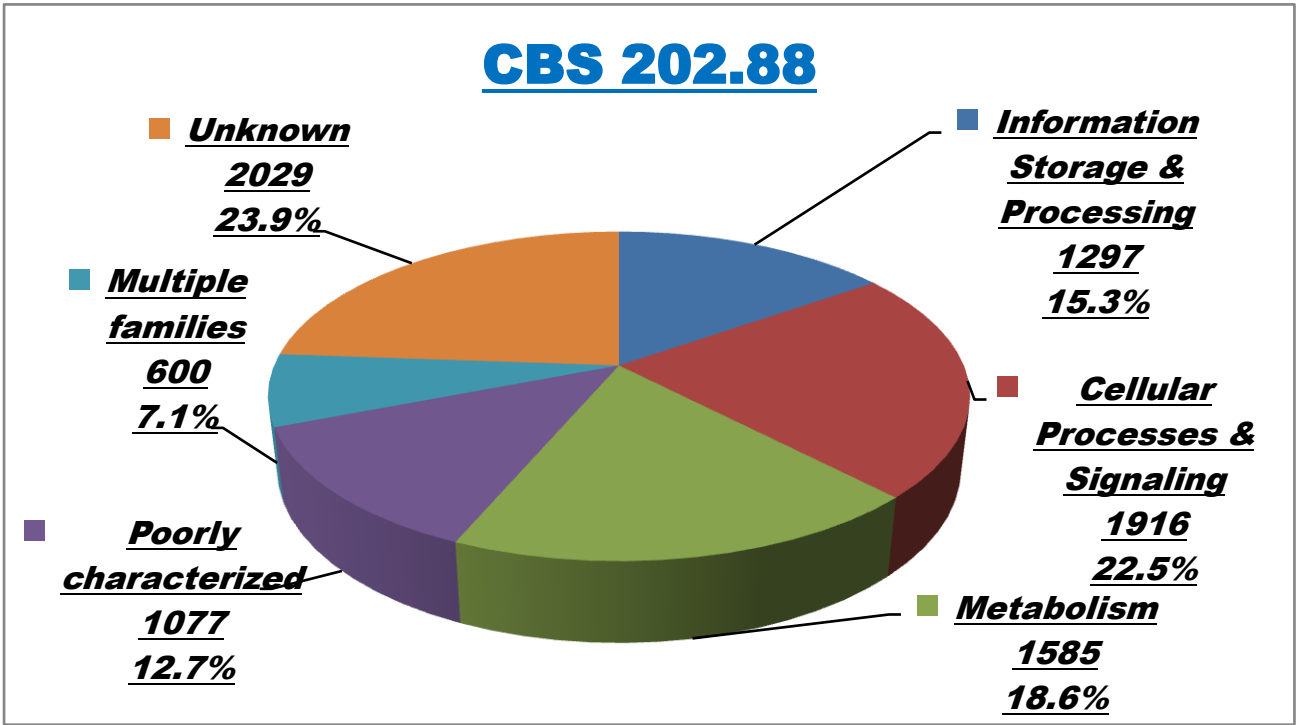


Figure3F. Distribution of predicted protein coding gene sequences of *T. rubrum* strain CBS 202.88

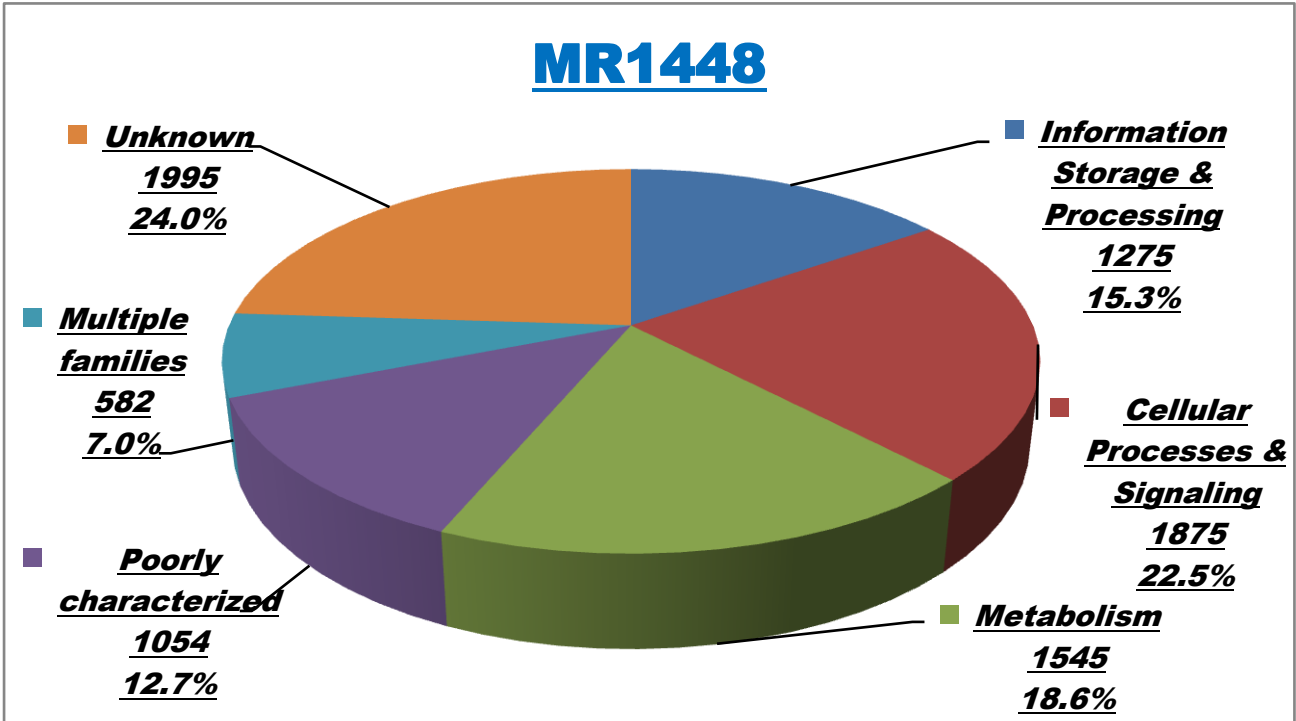


Figure3G. Distribution of predicted protein coding gene sequences of *T. rubrum* strain MR1448

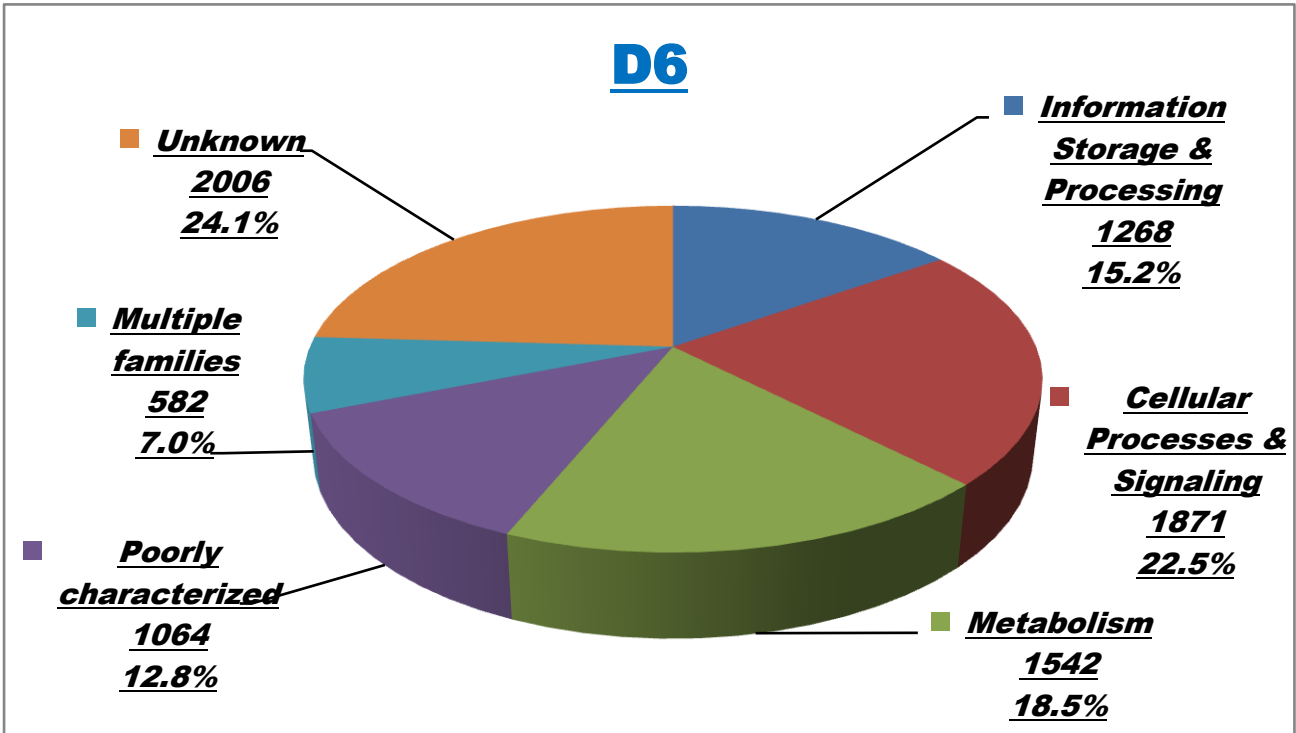


Figure3H. Distribution of predicted protein coding gene sequences of *T. rubrum* strain D6

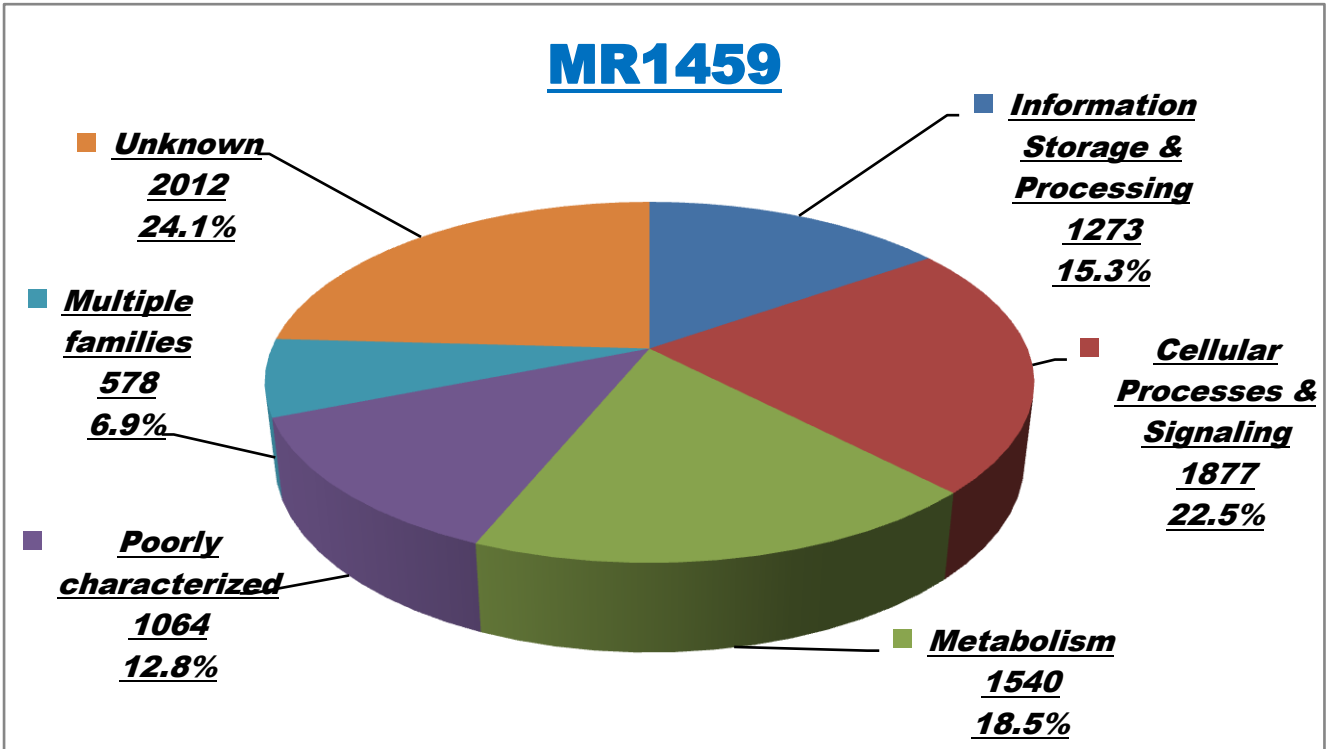


Figure3I. Distribution of predicted protein coding gene sequences of *T. rubrum* strain MR1459.

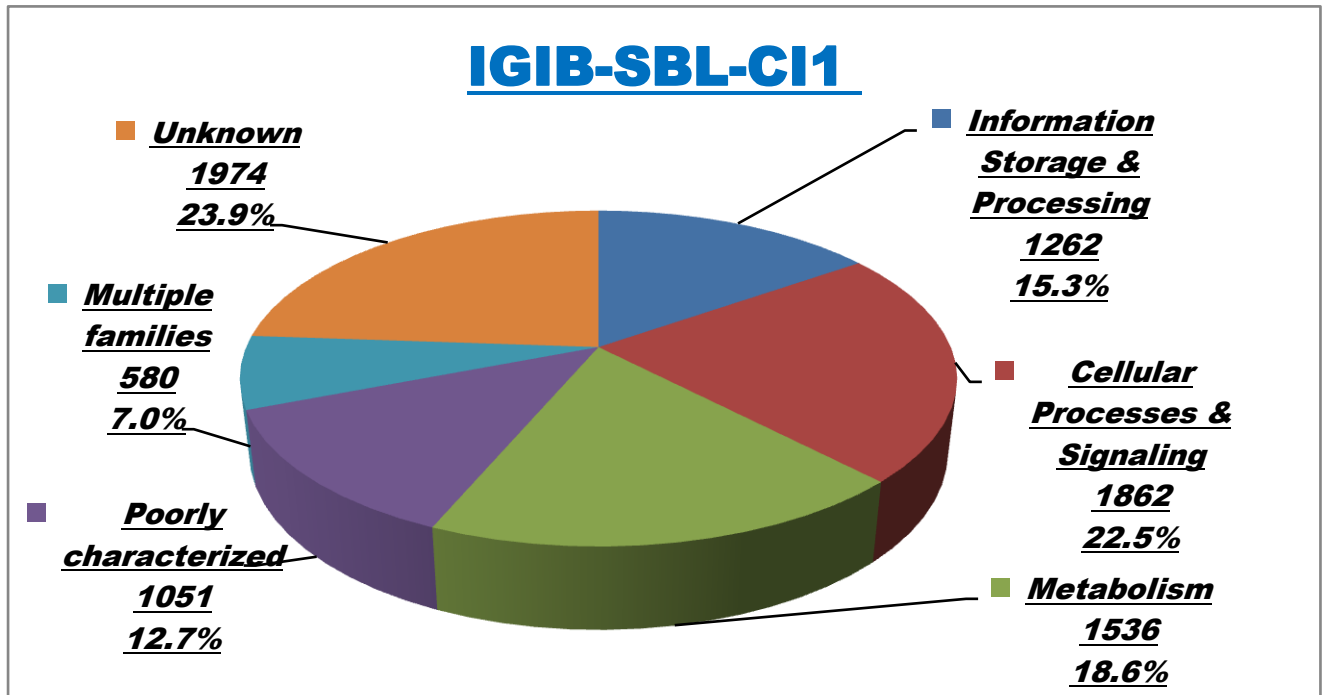


Figure3J. Distribution of predicted protein coding gene sequences of *T. rubrum* strain IGIB-SBL-CI1

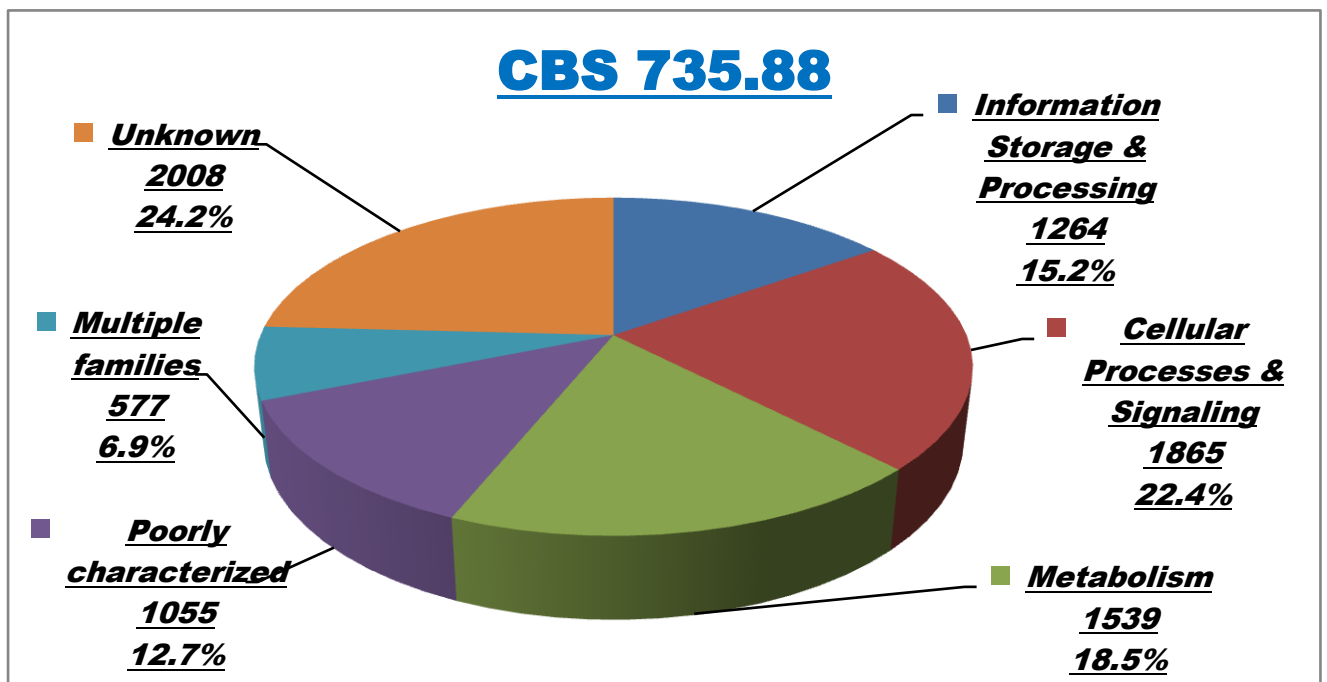


Figure3K. Distribution of predicted protein coding gene sequences of *T. rubrum* strain CBS 735.88

Figure 3. Classification of protein coding genes of *Trichophyton rubrum* strains based on functions defined by KOG database. Protein coding genes were classified in different functional categories defined by KOG.

Repeat content in all strains of *Trichophyton rubrum*

Repetitive sequences

Repetitive sequences in the Genome assembly of different strains of *T.rubrum* were predicted by using RepeatMasker online tool.

<i>T.rubrum</i> strain	GC content (%)	Retroelements (%)			DNA transposons (%)	simple repeats (%)	low complexity (%)	total repeats (%)
		SINEs	LINEs	LTRs				
CBS 118892	48.11	0	0	0	0	1.47	0.32	1.79
MR850	46.75	0	0	0	0	1.58	0.33	1.91
CBS 100081	46.78	0	0	0	0	1.58	0.33	1.91
CBS 288.86	46.77	0	0	0	0	1.58	0.33	1.91
CBS 289.86	46.78	0	0	0	0	1.58	0.34	1.92
CBS 202.88	47.73	0	0	0	0	1.62	34	1.96
MR1448	46.71	0	0	0	0	1.58	0.32	1.9
D6	46.82	0	0	0	0	1.59	0.34	1.93
MR1459	46.78	0	0	0	0	1.57	0.33	1.9
IGIB-SBL-CI1	47.78	0	0	0	0	1.58	0.31	1.93
CBS 735.88	47.56	0	0	0	0	1.56	0.34	1.9

Table 2. Repetitive sequences in the Genome assembly of different strains of *T.rubrum*. All the different strains of *Trichophyton rubrum* have similar GC content and similar repeats, predicted by using RepeatMasker.

Protein family classification

Carbohydrate active enzymes

Blast hits were obtained using both web servers CAT and dbCAN for prediction of carbohydrate active enzymes. Duplicate hits were removed from Blast hits obtained from CAT and dbCAN using conditional formatting.

Common hits for genes which were obtained in both blastp output of CAT and dbCAN were considered as carbohydrate active enzymes.

<i>T.rubrum strain</i>	AA	CBM	CE	GH	GT	PL	Total
CBS 118892	43	100	77	177	233	0	630
MR850	43	91	75	187	234	0	630
CBS 100081	36	129	87	212	256	0	720
CBS 288.86	42	91	77	185	228	0	623
CBS 289.86	35	123	84	216	257	0	715
CBS 202.88	36	126	85	210	253	0	710
MR1448	35	130	87	218	260	0	730
D6	35	126	84	212	256	0	713
MR1459	35	125	86	215	258	0	719
IGIB-SBL-CI1	35	127	87	217	253	0	719
CBS 735.88	36	123	81	217	255	0	712

Table 3. Carbohydrate active enzymes in different strains of *Trichophyton rubrum*. All the different strains of *Trichophyton rubrum* has large number of genes that are coding for Carbohydrate active enzymes and the number of genes coding for GH (Glycoside hydrolases) and GT (Glycosyl transferases) is higher than other Carbohydrate active enzymes in all strains.

Where AA, CBM, CE, GH, GT and PL are:

AA: Auxiliary activities (Redox enzyme that act in conjugation with CAZymes)

CBM: Carbohydrate binding modules (Important for the degradation of complex polysaccharides)

CE: Carbohydrate esterases (Hydrolysis of carbohydrate esters)

GH: Glycoside hydrolases (Hydrolysis of glycosidic bond)

GT: Glycosyl transferases (Formation of glycosidic bond)

PL: Polysaccharide lyases (Non-hydrolytic cleavage of glycosidic bond)

Auxiliary activities

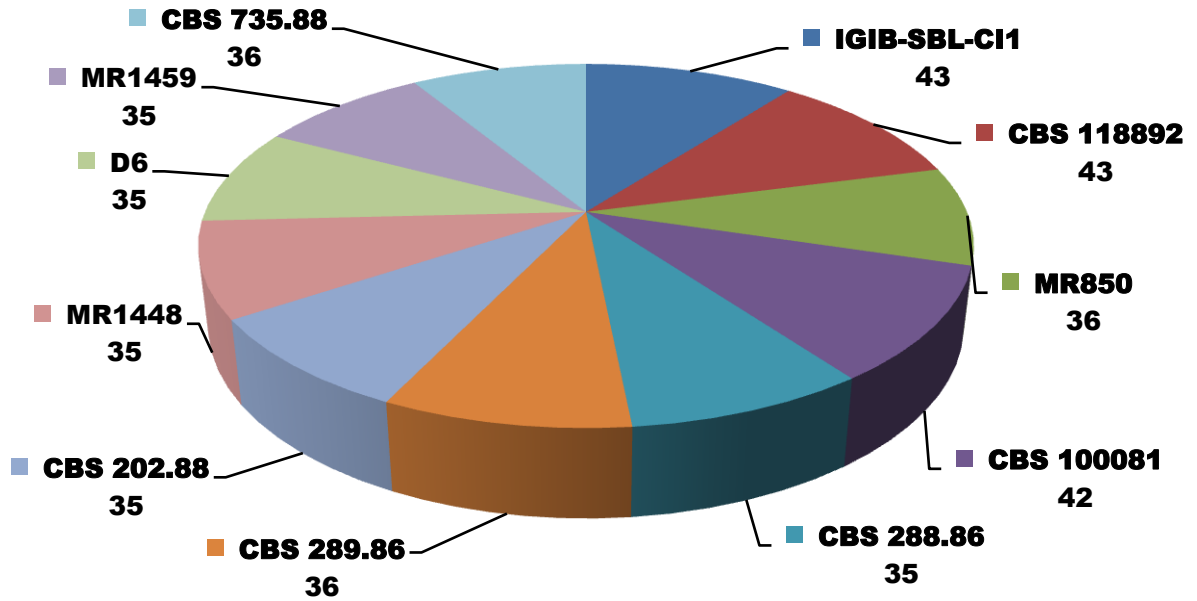


Figure 4A. Auxiliary activities

Carbohydrate binding modules

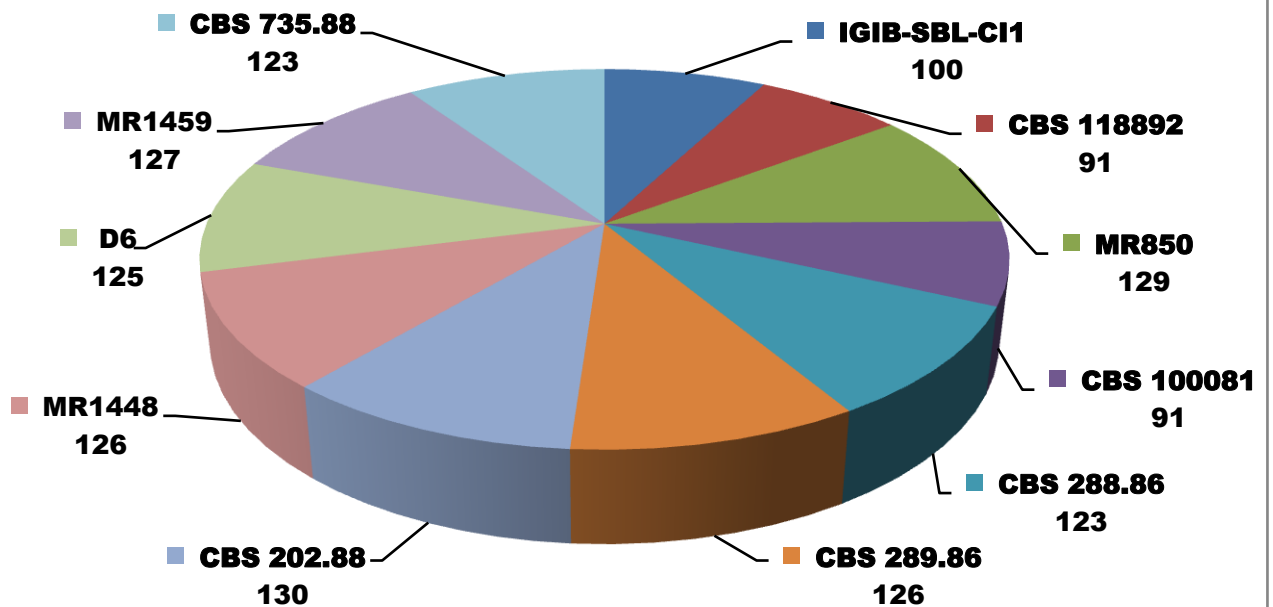


Figure 4B. Carbohydrate binding modules

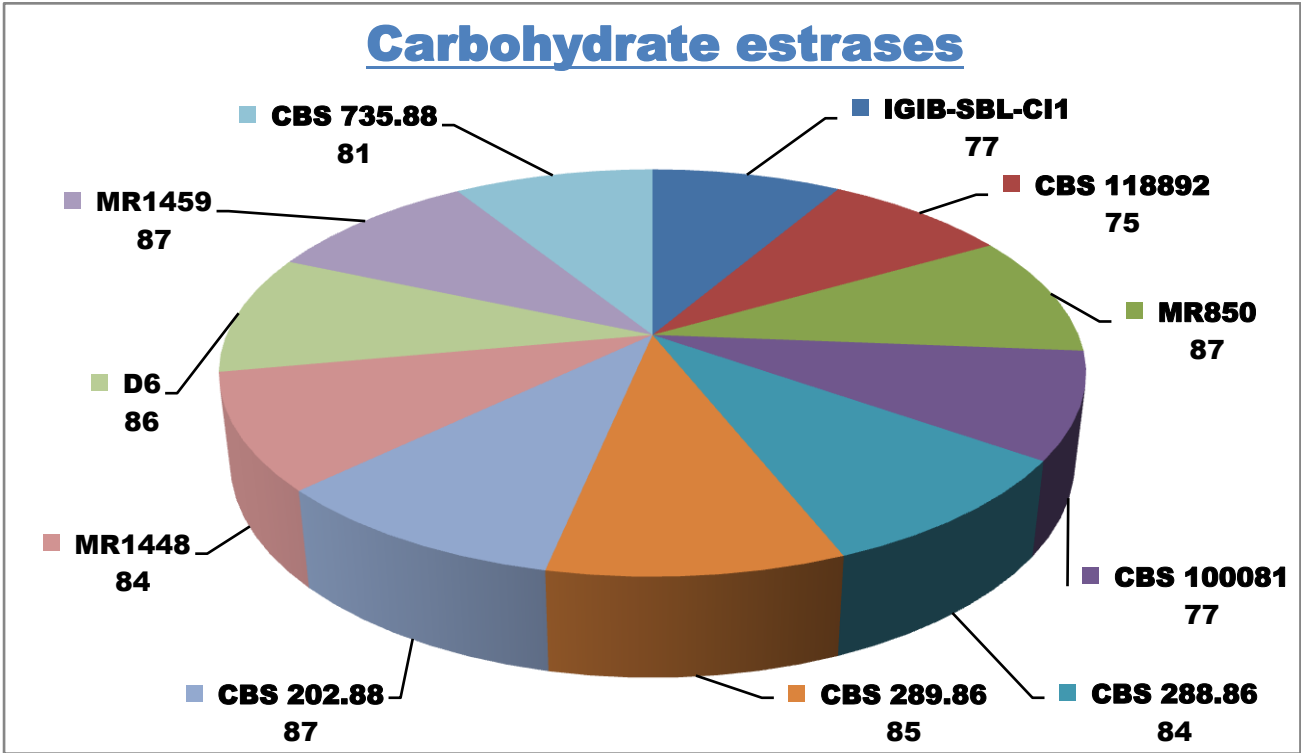


Figure 4C. Carbohydrate esterases

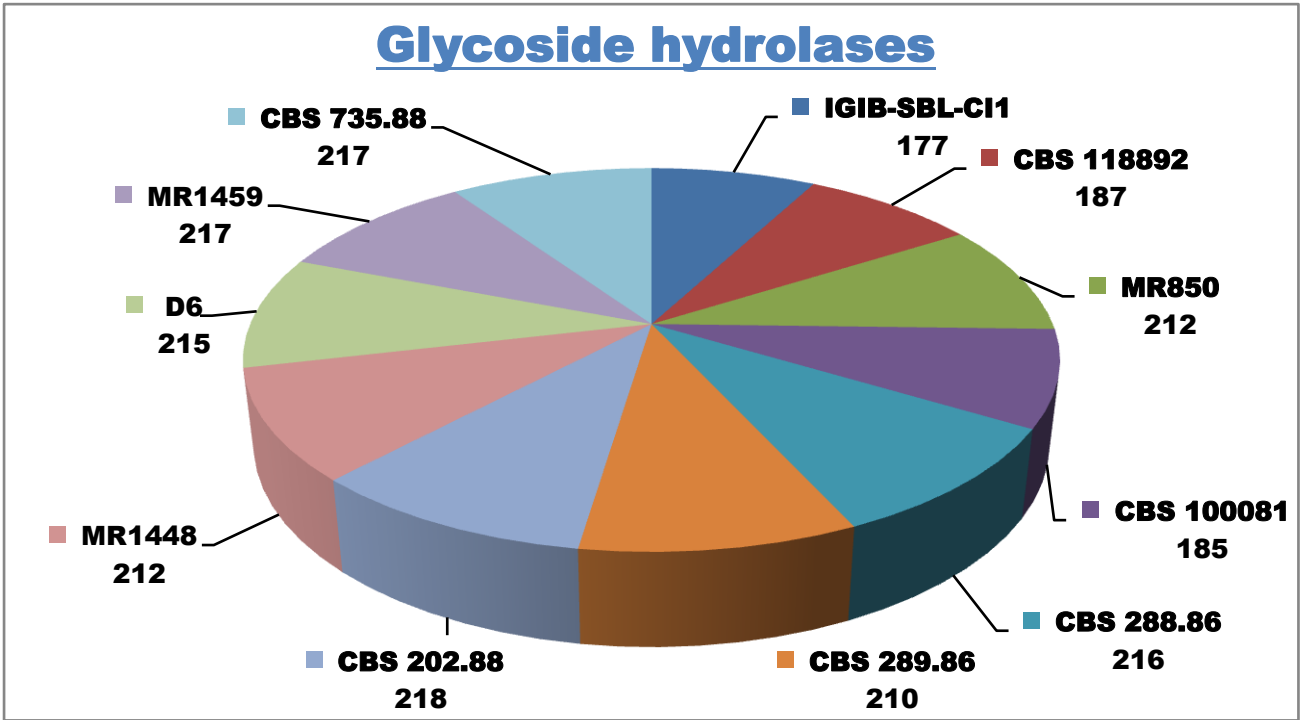


Figure 4D. Glycoside hydrolases

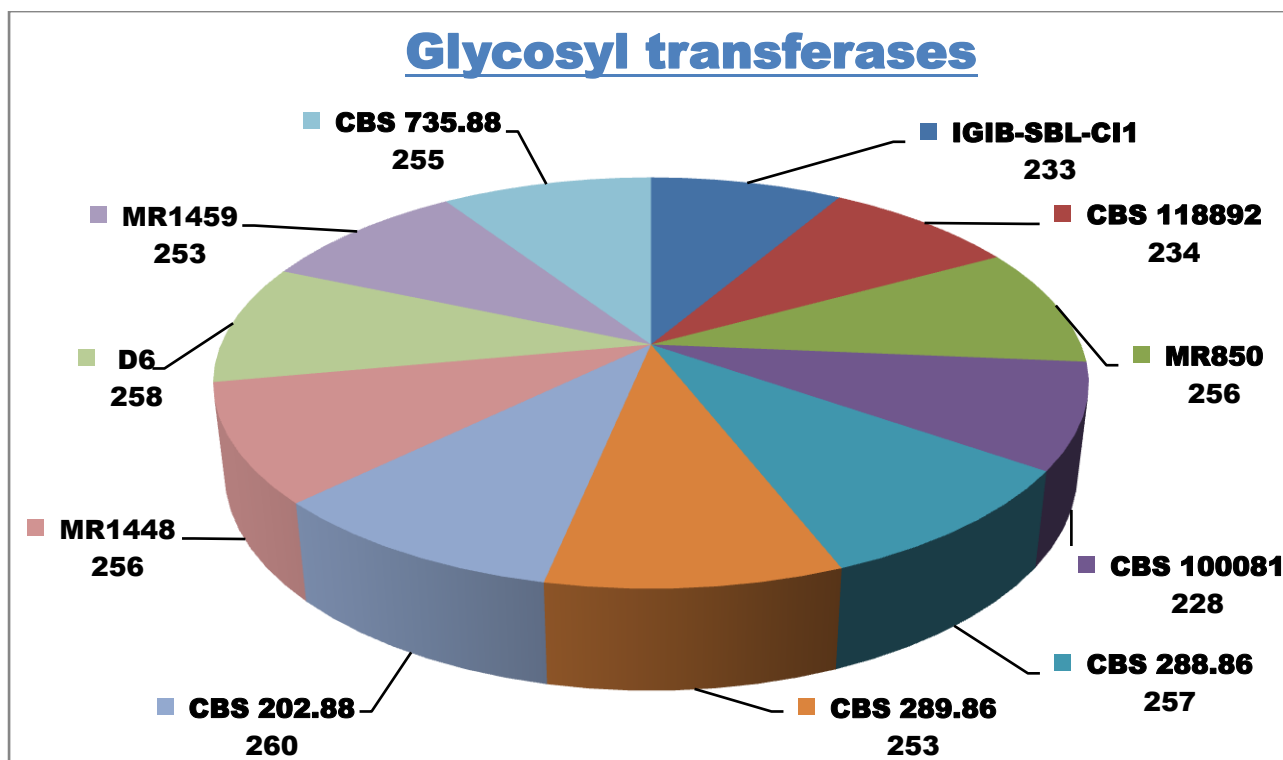


Figure 4E. Glycosyl transferases

Figure 4. Summary of *T.rubrum* strains genes assigned with CAZyme functional annotations. Figure 4A: Auxiliary activities, Figure 4B: Carbohydrate binding modules, Figure 4C: Carbohydrate esterases, Figure 4D: Glycoside hydrolases and Figure4E: Glycosyl transferases.

Lipases, Peptidases, Kinases and Cytochrome P450 family proteins.

Various Blast hits were obtained from blastp search for prediction of lipases, peptidases, kinases and Cytochrome P450 family proteins in different strains of *T.rubrum* by using Lipase Engineering Database, MEROPS database, KinBase and CYPED.

Duplicates were removed from Blast hits obtained by using conditional formatting and gene sequences with and above 80% query coverage and 60% sequence identity with search database (lipase_db, peptidase_db, kinase_db and CytP450_db) were considered as gene sequences coding for lipases, peptidases, kinases and Cytochrome P450 family proteins.

Lipases

Genes coding for different lipases, esterases and related proteins were predicted in different strains of *Trichophyton runrum*.

<i>T.rubrum strain</i>	Number of lipase coding genes
CBS 118892	22
MR850	22
CBS 100081	23
CBS 288.86	22
CBS 289.86	23
CBS 202.88	22
MR1448	22
D6	22
MR1459	23
IGIB-SBL-CI1	28
CBS 735.88	27

Table 4. Number of lipase coding genes in different strains of *Trichophyton rubrum*. All the different strains of *Trichophyton rubrum* has similar number of lipase coding genes.

Kinases : We predicted genes coding for different kinases in *T.rubrum* strains using KinBase.

<i>T.rubrum strain</i>	Number of kinase coding genes
CBS 118892	6
MR850	6
CBS 100081	5
CBS 288.86	5
CBS 289.86	4
CBS 202.88	7
MR1448	5
D6	5
MR1459	5
IGIB-SBL-CI1	5
CBS 735.88	5

Table 5. Number of genes coding for kinases in different strains of *Trichophyton rubrum*. All the different strains of *Trichophyton rubrum* has similar number of kinase coding genes.

Peptidases

Genes coding for different peptidases which were predicted in *T.rubrum* strains are

A: Aspartic Peptidases

C: Cysteine Peptidases

G: Glutamic Peptidases

M: Metallo Peptidases

N: Asparagine Peptide Lyases

S: Serine Peptidases

T: Threonine Peptidases

<i>T.rubrum</i> strain	A	C	G	M	N	S	T	Total
CBS 118892	13	75	0	106	2	100	24	320
MR850	12	72	0	105	2	100	22	313
CBS 100081	12	69	0	95	2	93	22	293
CBS 288.86	12	68	0	95	2	93	22	292
CBS 289.86	12	68	0	95	2	93	22	292
CBS 202.88	12	68	0	95	2	93	22	292
MR1448	13	69	0	98	2	92	22	296
D6	12	69	0	95	2	92	22	292
MR1459	12	69	0	95	2	93	22	293
IGIB-SBL-CI1	12	67	0	94	1	93	22	289
CBS 735.88	12	69	0	98	2	92	21	294

Table 6. Different peptidases in different strains of *Trichophyton rubrum*. All the different strains of *Trichophyton rubrum* has similar number of peptidase coding genes.

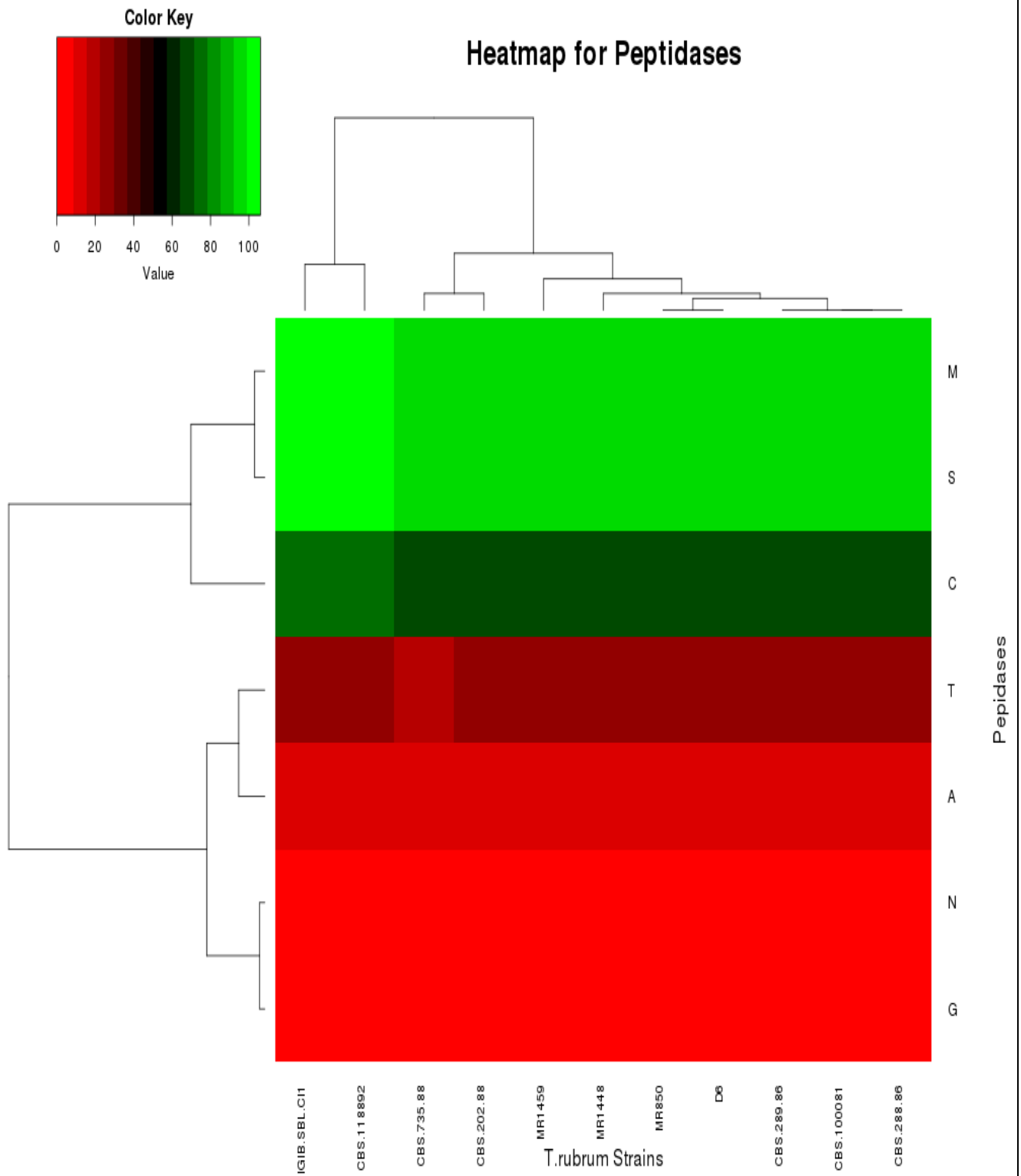


Figure 5. Heat map of peptidases families in different strains of *T. rubrum*.

Cytochrome P450 family proteins: Genes coding for different protein families of Cytochrome P450 family proteins were found in *T. rubrum* strains.

<i>T.rubrum</i> strain	Number of CytP 450 family coding genes
CBS 118892	13
MR850	12
CBS 100081	11
CBS 288.86	12
CBS 289.86	12
CBS 202.88	12
MR1448	12
D6	12
MR1459	12
IGIB-SBL-CI1	12
CBS 735.88	12

Table 7. Number of genes of *T.rubrum* strains coding for different Cytochrome P450 family proteins. All the different strains of *Trichophyton rubrum* has similar number of Cytochrome P450 family coding genes.

Pathogenicity

Pathogenicity related genes or virulence genes that can cause diseases were found from blastp search against PHI database for all strains of *T.rubrum*.

<i>T.rubrum</i> strain	Number of virulence genes
CBS 118892	192
MR850	192
CBS 100081	193
CBS 288.86	191
CBS 289.86	196
CBS 202.88	190
MR1448	191
D6	191
MR1459	191
IGIB-SBL-CI1	207
CBS 735.88	205

Table 8. Number of pathogenicity related genes in *T.rubrum* strains. All the different strains of *Trichophyton rubrum* has similar number of virulence genes predicted by using PHI database..

Secretory Lipases, Peptidases, Pathogenicity related genes, Kinases and Carbohydrate active enzymes

Secretory lipases, peptidases, virulence or pathogenicity related genes and carbohydrate active enzymes were predicted by using SignalP server.

<i>T.rubrum</i> strain	Secretory lipases	Secretory peptidases	Secretory virulence genes	Secretory CAZYmes	Kinases
CBS 118892	6	72	6	138	0
MR850	7	72	6	143	0
CBS 100081	7	72	7	144	0
CBS 288.86	7	72	6	141	0
CBS 289.86	7	72	6	144	0
CBS 202.88	7	72	6	142	0
MR1448	7	72	6	143	0
D6	7	72	6	144	0
MR1459	7	69	5	137	0
IGIB-SBL-CI1	9	77	7	217	0
CBS 735.88	9	76	81	217	0

Table 9. Number of Secretory, Peptidases, Virulence genes, CAZYmes , lipases and kinases and in *T.rubrum* strains. The number of Secretory lipases, Peptidases, Virulence genes and CAZYmes in *T.rubrum* strains are similar and any strain of *T.rubrum* do not have secretory kinases.

Unique genes

Unique genes in the genome of different strains of *T.rubrum* were predicted using CD-HIT-2D server. All vs all similarity search was done using CD-HIT-2D server.

By using conditional formatting unique genes that are present only in one single strain and absent in all other 10 strains were predicted. All unique genes present in all 11 different strains of *T.rubrum* were predicted.

Functional annotation of unique genes predicted in different strains of *T.rubrum*.

The functions for all unique genes predicted in all 11 different strains of *T.rubrum* were predicted using Blast2GO.

<i>T.rubrum</i> strain	Number of unique genes	Number of genes annotated by Blast2GO
CBS 118892	98	51
MR850	9	2
CBS 100081	12	2
CBS 288.86	9	1
CBS 289.86	9	2
CBS 202.88	26	13
MR1448	14	7
D6	7	1
MR1459	11	2
IGIB-SBL-CI1	62	32
CBS 735.88	129	59

Table 10. Number of unique genes predicted in different strains of *T.rubrum*. Different numbers of unique genes in all strains were annotated by Blast2GO.

Multiple sequence alignment of Drug resistance genes in all strains of *T. rubrum* .

Gene sequences for drug resistance genes in refseq CBS 118892 were obtained from NCBI and Drug resistance gene sequences in other strains of *T. rubrum* were find using Blastp search against refseq CBS 118892. Multiple sequence alignment was done using ClustalW (Multiple sequence alignment tool).

<i>T.rubrum</i> strain	Squalene epoxidase gene		TratrD gene (MDR)		TERG_01703 (Cytochrome P450 51)		TERG_02186 (ABC transporter)	
	Length	Mismatch	Length	Mismatch	Length	Mismatch	Length	Mismatch
CBS 118892	489	-	1331	-	527	-	1503	-
MR850	489	0	1331	0	527	0	1613	0
CBS 100081	489	0	1331	0	527	0	1613	0
CBS 288.86	489	0	1331	0	527	0	1613	0
CBS 289.86	489	0	1331	0	527	0	1613	0
CBS 202.88	489	0	1331	0	527	0	1613	0
MR1448	489	0	1331	0	527	0	320	0
D6	489	0	1331	0	527	0	1613	1
MR1459	489	0	1331	0	527	0	1613	0
IGIB-SBL-CI1	489	1	1331	0	527	0	1613	0
CBS 735.88	489	0	1331	2	527	0	1613	6

Table11. Mismatch in different drug resistance genes of different strains of *T.rubrum*. Multiple sequence alignment of four different drug resistance genes by CLUSTAL W.

From multiple sequence alignment of drug resistance genes (squalene epoxidase gene, TratrD gene, TERG_01703 (Cytochrome P450 51) and TERG_02186 (ABC transporter)) by using CLUSTAL W, we predicted that mismatch in some of the *T.rubrum* strains against refseq CBS 118892.1 mismatch in *T.rubrum* IGIB-SBL-CI1 strain for squalene epoxidase gene , 2 mismatch in *T.rubrum* CBS 735.88 strain for TratrD gene(MDR), no mismatch was found for TERG_01703 (Cytochrome P450 51) and for TERG_02186 (ABC transporter) gene 1 mismatch found in *T.rubrum* D6 and 6 mismatch in *T.rubrum* CBS 735.88. Mismatches that

were found in nucleotide sequences of different strains of *T.rubrum* code for same codon, eg. GTT and GTC both code for valine.

Squalene epoxidase gene

CBS 118892 **IGIB-SBL-CI1**

Leucine (L) Phenylalanine (F) (mismatch at position 393)

TratrD gene(MDR)

CBS 118892 **CBS 735.88**

Leucine (L) Serine (S) (mismatch at position 48)

Serine(S) Proline (P) (mismatch at position 858)

TERG_02186 (ABC transporter)

CBS 118892 **D6**

Valine (V) Alanine (A) (mismatch at position 1347)

CBS 118892 **CBS 735.88**

Proline (P) Alanine (A) (mismatch at position 202)

Valine (V) Alanine (A) (mismatch at position 829)

Valine (V) Alanine (A) (mismatch at position 840)

Valine (V) Isoleucine (I) (mismatch at position 884)

Threonine (T) Alanine (A) (mismatch at position 1266)

Isoleucine (I) Phenylalanine (F) (mismatch at position 1332)

6. CONCLUSION

Genome sequences of *Trichophyton rubrum* strains were used for comparative genomic analysis. Protein coding genes in different strains of *Trichophyton rubrum* were annotated in different functional categories using KOG database. These *Trichophyton rubrum* strains have a large number of genes encoding for proteins which are virulence factors that include peptidases or proteases, CAZymes, lipases, kinases, Cytochrome P450. These virulence factors in different strains of *T.rubrum* were identified by using bioinformatics approaches. Whole genomes of different strains were used to predict protein coding genes and these predict protein coding genes were used for identification of virulence factors. Prediction of protein families of virulence factors was done using blastp against databases of protein families of these different virulence factors i.e. PHI database, CAZY database, lipase database, peptidase database, kinase database and Cytochrome P450 database.

T.rubrum strains were found to encode large numbers of peptidases-encoding genes, these different strains have large number of carbohydrate active enzymes, virulence genes, lipases and it was predicted that these dermatophyte species have novel sets kinases that are fungus-specific. These strains have genes encoding Cytochrome P450 families and novel genes in all different strains were predicted. Multiple sequence alignment of four genes that show drug resistance i.e squalene epoxidase, TERG_02186 (ABC transporter), TERG_01703 (cytochrome P450 51) and TratrD gene (multidrug resistance gene) was done by using ClustalW.

This comparative genomic analysis of different *T.rubrum* strains help in identification of unique gene sequence features and gene families that are present in dermatophyte species that may cause fungal infection.

7. DISCUSSION AND FUTURE PERSPECTIVE

In this study we identified protein families that may be involved in causing dermatophytosis by *Trichophyton rubrum* strains i.e. virulence factors of *T.rubrum* strains. These virulence factors in these strains were identified by using bioinformatics approaches. Eleven whole genomes of different *Trichophyton rubrum* strains were used for this comparative study. By using WebAUGUSTUS for gene prediction and manual correction we found that the number of protein coding genes is different in all eleven strains. These predicted genes were used for identification of protein families. Total numbers of genes in all different strains of *Trichophyton rubrum* species were classified in different functional categories like Information Storage & Processing which include RNA processing and modification, chromatin structure and dynamics, translation, transcription, replication, cellular processes and signaling which include cell cycle control, signal transduction mechanism, cytoskeleton, nuclear structure, defense mechanism etc., metabolism which include energy production, nucleotide transport, secondary metabolites biosynthesis, amino acid transport etc, gene sequences having a conserved domain but uncharacterized function are grouped in poorly characterized category, gene sequences having multiple functions are grouped in multiple families and gene sequences which does not show any similarity with KOG database are grouped in category function unknown.

GC content and total repeats present in genome sequences of different strains were predicted and this identifies that total repeats and total GC content present in different strains are different.

We conducted a comparative analysis of different protein families like carbohydrate-active-enzymes, lipases, kinases, peptidases, cytochrome P450 and pathogenicity related genes. Analysis of various carbohydrate-active-enzymes (Auxiliary activities, Carbohydrate binding modules, Carbohydrate esterases, Glycoside hydrolases, Glycosyl transferases, Polysaccharide lyases) shows different number of CAZymes are present in different strains of *T.rubrum* responsible for hydrolysis of carbohydrates present in skin. It has been found that no gene is coding for Polysaccharide lyases responsible for non-hydrolytic cleavage of glycosidic bond. This study identifies that these dermatophyte strains of *T.rubrum* have number of genes that are coding for lipases, esterases and other related proteins that are known to be lipid degradation enzymes. All eleven genomes of *T.rubrum* strains were found to encode large numbers of peptidases-encoding genes. Different number of genes in all strains encoding for aspartic peptidases, cysteine peptidases, glutamic peptidases, metallo peptidases, asparagine peptidases and serine peptidases.

Genes of *Trichophyton rubrum* strains also encode for novel sets of kinases that are fungus-specific. These fungus specific kinases are known to involve in signaling and secondary metabolism of dermatophytes. All eleven strains have genes that encode kinases like AGC, CAMK, CMGC, MAPK, CDK and CK2. AGC is a group of protein kinase includes PKA, PKG and PKC. CAMK is calmodulin/calcium regulated kinases, MAPK is Mitogen Activated Protein Kinase, CMGC kinases include key kinases like MAPK growth and stress-response kinases, CDK (Cyclin Dependent Kinases) and CK2 i.e. Cell Kinase 2

(Casein Kinase 2). These predicted fungus specific kinases may also contribute to pathogenicity of *T.rubrum*.

These strains have genes encoding Cytochrome P450 familie i.e. CYP584, CYP51, CYP551,CYP584,CYP505,CYP58.CYP65,CYP55 and CYP503, these cytochrome P450 family related genes are associated in secondary metabolite synthesis which are known as virulence factors in dermatophytes. Pathogenicity-related genes were also predicted in all eleven strains of *T.rubrum*.Secretory analysis by using SignalP shows that there is high number of secretory peptidases and carbohydrate active enzymes. Secretome analysis identifies secretory lipases and secretory pathogenicity related genes, this study show that *T.rubrum* strains do not have secretory kinases. From this study we found that genome of these *T.rubrum* strains have high number of genes encoding for peptidases, lipases and CAZymes, this suggest that these genes enhance pathogenicity of dermatophytes i.e. *T.rubrum* strains and involve in skin infection.

Comparative genomic analysis identifies different number of novel genes in *T.rubrum* strains, novel genes were annotated by Blast2GO. These predicted novel genes may also contribute to the pathogenicity of dermatophytes i.e. *T.rubrum* strains.

In this study we also performed multiple sequence alignment of four genes that show drug resistance i.e squalene epoxidase, TERG_02186 (ABC transporter), TERG_01703 (cytochrome P450 51) and TratrD gene (multidrug resistance gene). This multiple sequence alignment study identifies some mismatches between refseq drug resistance and other *T.rubrum* strains. This analysis suggests that the replacement of one amino acid by other amino acid in these genes may contribute to drug resistance of dermatophytes *T.rubrum* strains.

The availibility of genome sequences of different dermatophytes help in identification of unique gene sequence features present in dermatophyte species that may cause fungal infection. Identification and characterization of virulence associated genes, genes involved in secondary metabolite biosynthesis and secretory degradative enzymes in dermatophytes may help in development of new therapies, this may help in understanding host dermatophyte interaction and role of these virulence related genes in pathogenicity of dermatophytes. This could and also help in finding the reasons for drug resistance in dermatophyte species that prevents dermatophyte species from complete clearance and result in treatment failure.

8. REFERENCES

- Achterman, RR. and White, TC (2012). A foot in the door for dermatophyte research. *PLoS Pathog.* **8**: e1002564.
- Achterman, RR. & White TC (2012). Dermatophyte Virulence Factors: Identifying and Analyzing Genes That May Contribute to Chronic or Acute Skin Infections. *International Journal of Microbiology*. Article ID 358305.
- Alfoldi, J; and Lindblad-Toh, K (2013) Comparative genomics as a tool to understand evolution and disease. *Genome Research*. **23**:1063–1068.
- Altschul,SF; Gish,W; Miller, W; Myers, EW. and Lipman, DJ (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**; 403–410.
- Aly R (1994) Ecology and epidemiology of dermatophyte infections. *J Am Acad Dermatol*. S21-5.
- Ameen,M (2010). Epidemiology of superficial fungal infections. *Clin Dermatol*. **28**: 197-201.
- Apodaca, G; and McKerrow, JH (1990). Expression of proteolytic activity by cultures of *Trichophyton rubrum*. *J. Med. Vet. Mycol.* **28**:159–171.
- Baldwin,TK; Winnenburg,R; Urban,M; Rawlings,C; Koehler,J. and Hammond-Kosack,KE (2006). The pathogen-host interactions database (PHI-base) provides insights into generic and novel themes of pathogenicity. *Mol. Plant Microbe Interact.* **19**: 1451–1462.
- Bhatia, VK. and Sharma, PC (2014). Epidemiological studies on Dermatophytosis in human patients in Himachal Pradesh, India. *SpringerPlus*. **3**:134.
- Burmester, A; Shelest, E; Glöckner, G; Heddergott, C; *et al.* (2011). Comparative and functional genomics provide insights into the pathogenicity of dermatophytic fungi. *Genome Biol.* **12**: R7.
- Conesa, A; Götz, S; García-Gómez, JM; Tero, J; Talón, M. and Robles ,M (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. **21** (18): 3674-3676.
- Eklom, R. and Wolf, BW (2014). A field guide to whole-genome sequencing, assembly and annotation. Department of Evolutionary Biology. 18D.
- Elewski, BE (1998). Onychomycosis: pathogenesis, diagnosis, and management. *ClinMicrobiol Rev.* **11**: 415–429.

Fischer, M; Knoll, M; Sirim, D; *et al.* (2007). The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics/btm*. Vol. **23**: 268.

Fischer, M. and Jurgen, P (2003). The Lipase Engineering Database: a navigation and analysis tool for proteins. *Nucl. Acid. Res.* **31**: 319-321.

Hane, JK; Anderson, JP; Williams, AH; Sperschneider, J. and Singh, KB (2014). Genome Sequencing and Comparative Genomics of the Broad Host-Range Pathogen *Rhizoctonia solani* AG8. *PLoS Genet.* 10(5): e1004281.

Havlickova, B; Czaika, VA. and Friedrich, M (2008). Epidemiological trends in skin mycoses worldwide. *Mycoses.* **51** (Suppl 4): 2-15.

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

Huang, Y; Niu, B; Gao, Y; *et al.* (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics.* **26**:680-682.

Latka, C; Dey, SS; Mahajan, S; Prabu, R; *et al.* (2015). Sequence of a clinical isolate of dermatophyte, *Trichophyton rubrum* from India. *FEMS Microbiology.* **8**:362.

Lee, WJ; Kim, SL; Jang, YH; *et al.* (2014). Increasing Prevalence of *Trichophyton rubrum* Identified through an Analysis of 115,846 Cases over the Last 37 Years. *J Korean Med Sci.* **30**: 639-643.

Ma, LJ; Borkovich, KA; Coleman, JJ; *et al.* (2010). Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature*. Vol 464.

Martinez, DA; Oliver, BG; Gräser, Y; *et al.* (2012). Comparative genome analysis of *Trichophyton rubrum* and related dermatophytes reveals candidate genes involved in infection. *MBio.* **3**: e00259–00212.

Martinez-Rossi, NM; Peres, NTA. and Rossi, A (2008). Antifungal Resistance Mechanisms in Dermatophytes. *Mycopathologia.* **166**:369–383.

Monod, M (2008). Secreted proteases from dermatophytes. *Mycopathologia.* **166**: 285–294.

Mukherjee, PK; Leidich, SD; Isham, N; Leitner, I; Ryder, NS. and Ghannoum, MA (2003). Clinical *Trichophyton rubrum* strain exhibiting primary resistance to terbinafine. *Antimicrob Agents Chemother* **47**: 82–86.

Ohm, RA; Feau, N; Henrissat, B; Schoch, CL; *et al.* (2012). Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen *Dothideomycetes* Fungi. *PLoS Pathog.* **8**(12): e1003037.

Osborne, CS; Leitner, I; Favre, B. and Ryder, NS (2005). Amino acid substitution in *Trichophyton rubrum* squalene epoxidase associated with resistance to terbinafine. *Antimicrob Agents Chemother.* **49** : 2840-2844.

Park, BH; Karpinets, TV; Syed, MH; Leuze, MR. and Edward, CU (2010). CAZymes Analysis Toolkit (CAT): Web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *Glycobiology* vol. 20 no. 12 pp. 1574–1584.

Petersen, TN; Brunak, S; Heijne, G. and Nielsen, H (2011). SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods.* **8**: 785–786.

Rawlings, ND; Morton, FR; Kok, CY; Kong, J. and Barrett, AJ (2008). MEROPS: the peptidase database. *Nucleic Acids Res.* **36**: D320–D325.

Seebacher, C; Bouchara, JP. and Mignon, B (2008). Updates on the Epidemiology of Dermatophyte Infections. *Mycopathologia.* **166**:335–352.

Simpanya, MF. (2000). Dermatophytes: Their taxonomy, ecology and pathogenicity. *Revista Iberoamericana de Micología.* E-48080.

Stein, LD; Bao, Z; Blasiar, D; *et al.* (2003). The Genome Sequence of *Caenorhabditis briggsae*: A Platform for Comparative Genomics. *PLoS Biology.* Vol 1.

Tarailo-Graovac, M. and Chen, N (2009). Using RepeatMasker to Identify Repetitive Elements in Genomic Sequences. *Current Protocols in Bioinformatics.* 25:4.10:4.10.1–4.10.14.

Tatusov, RL; Fedorova, ND; Jackson, JD; Jacobs, JR; *et al.* (2003). The COG database: an updated version includes eukaryotes. *BMC Bioinformatics.* **4** : 41.

Teixeira, MM; Almeida, LGP; Kubitschek-Barreira, P; *et al.* (2014). Comparative genomics of the major fungal agents of human and animal Sporotrichosis: *Sporothrix schenckii* and *Sporothrix brasiliensis* *BMC Genomics.* **15**:943.

Thompson, JD; Higgins, DG; and Gibson, TJ (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **22**: 4673–4680.

Wang, L; Ma, L; Leng, W; *et al.* (2006). Analysis of the dermatophyte *Trichophyton rubrum* expressed sequence tags. *BMC Genomics*. **7**:255.

Weitzman, I. and Summerbell, RC (1995). The dermatophytes. *Clin Microbiol Rev.* **8**: 240–259.

Winnenburg, R; Baldwin, TK; Urban, M; *et al.* (2006). PHI-base: a new database for pathogen host interactions. *Nucl. Acid. Res.* **34**:D459–D464.

Wu, S; Zhu, Z; Fu, L; Niu, B. and Li, W (2011). WebMGA: a customizable web server for fast metagenomic sequence analysis. *BMC Genomics*. **12**: 444.

Wu, Y; Yang, J; Yang, F; *et al.* (2009). Recent dermatophyte divergence revealed by comparative and phylogenetic analysis of mitochondrial genomes. *BMC Genomics*. **10**:238.

Yang, J; Chen, L; Wang, L; *et al.* (2007). TrED: the *Trichophyton rubrum* Expression Database. *BMC Genomics*. **8**:250.

Yang, J; Wang, L; Ji, X; Feng, Y; Li, X; *et al.* (2011). Genomic and Proteomic Analyses of the Fungus *Arthrobotrys oligospora* Provide Insights into Nematode-Trap Formation. *PLoS Pathog.* **7**(9): e1002179.

Yin, Y; Mao, X; Yang, X; Chen, X; Mao, F. and Xu, Y (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation . *Nucleic Acids Research*. Vol. 40 W445–W451.

Zhao, Z; Liu, H; Wang, C. and Xu, JR (2013). Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics*. **14**:274.

9. APPENDIX

KOG Category CBS 118892		
Information Storage & Processing		1308
A	RNA processing and modification	260
B	Chromatin structure and dynamics	98
J	Translation, ribosomal structure and biogenesis	369
K	Transcription	363
L	Replication, recombination and repair	218
Cellular Processes & Signaling		1986
D	Cell cycle control, cell division, chromosome partitioning	192
M	Cell wall/membrane/envelope biogenesis	74
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	522
T	Signal transduction mechanisms	485
U	Intracellular trafficking, secretion, and vesicular transport	423
V	Defense mechanisms	35
W	Extracellular structures	8
Y	Nuclear structure	30
Z	Cytoskeleton	213
Metabolism		1649
C	Energy production and conversion	278
E	Amino acid transport and metabolism	278
F	Nucleotide transport and metabolism	83
G	Carbohydrate transport and metabolism	227
H	Coenzyme transport and metabolism	112
I	Lipid transport and metabolism	319
P	Inorganic ion transport and metabolism	141
Q	Secondary metabolites biosynthesis, transport and catabolism	211
Poorly characterized		1105
R	General function prediction only	778
S	Function unknown	327
Multiple families		617
X	multiple functions	1
Unknown		2154
Total		8820

KOG Category MR850		
Information Storage & Processing		1272
A	RNA processing and modification	255
B	Chromatin structure and dynamics	95
J	Translation, ribosomal structure and biogenesis	352
K	Transcription	362
L	Replication, recombination and repair	208
Cellular Processes & Signaling		1874
D	Cell cycle control, cell division, chromosome partitioning	182
M	Cell wall/membrane/envelope biogenesis	73
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	496
T	Signal transduction mechanisms	454
U	Intracellular trafficking, secretion, and vesicular transport	398
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	28
Z	Cytoskeleton	198
Metabolism		1543
C	Energy production and conversion	260
E	Amino acid transport and metabolism	259
F	Nucleotide transport and metabolism	76
G	Carbohydrate transport and metabolism	218
H	Coenzyme transport and metabolism	102
I	Lipid transport and metabolism	298
P	Inorganic ion transport and metabolism	132
Q	Secondary metabolites biosynthesis, transport and catabolism	198
Poorly characterized		1058
R	General function prediction only	743
S	Function unknown	315
Multiple families		579
X	multiple functions	1
Unknown		2004
Total		8331

KOG Category CBS 100081		
Information Storage & Processing		1268
A	RNA processing and modification	253
B	Chromatin structure and dynamics	96
J	Translation, ribosomal structure and biogenesis	349
K	Transcription	363
L	Replication, recombination and repair	207
Cellular Processes & Signaling		1867
D	Cell cycle control, cell division, chromosome partitioning	182
M	Cell wall/membrane/envelope biogenesis	72
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	493
T	Signal transduction mechanisms	455
U	Intracellular trafficking, secretion, and vesicular transport	397
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	27
Z	Cytoskeleton	196
Metabolism		1535
C	Energy production and conversion	259
E	Amino acid transport and metabolism	257
F	Nucleotide transport and metabolism	75
G	Carbohydrate transport and metabolism	216
H	Coenzyme transport and metabolism	101
I	Lipid transport and metabolism	297
P	Inorganic ion transport and metabolism	133
Q	Secondary metabolites biosynthesis, transport and catabolism	197
Poorly characterized		1063
R	General function prediction only	745
S	Function unknown	318
Multiple families		578
X	multiple functions	1
Unknown		2001
Total		8313

KOG Category CBS 288.86		
Information Storage & Processing		1267
A	RNA processing and modification	252
B	Chromatin structure and dynamics	96
J	Translation, ribosomal structure and biogenesis	351
K	Transcription	361
L	Replication, recombination and repair	207
Cellular Processes & Signaling		1872
D	Cell cycle control, cell division, chromosome partitioning	182
M	Cell wall/membrane/envelope biogenesis	73
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	495
T	Signal transduction mechanisms	452
U	Intracellular trafficking, secretion, and vesicular transport	400
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	26
Z	Cytoskeleton	199
Metabolism		1545
C	Energy production and conversion	259
E	Amino acid transport and metabolism	258
F	Nucleotide transport and metabolism	77
G	Carbohydrate transport and metabolism	216
H	Coenzyme transport and metabolism	104
I	Lipid transport and metabolism	300
P	Inorganic ion transport and metabolism	133
Q	Secondary metabolites biosynthesis, transport and catabolism	198
Poorly characterized		1060
R	General function prediction only	745
S	Function unknown	315
Multiple families		579
X	multiple functions	1
Unknown		2007
Total		8331

KOG Category CBS 289.86		
Information Storage & Processing		1267
A	RNA processing and modification	253
B	Chromatin structure and dynamics	95
J	Translation, ribosomal structure and biogenesis	350
K	Transcription	361
L	Replication, recombination and repair	208
Cellular Processes & Signaling		1865
D	Cell cycle control, cell division, chromosome partitioning	180
M	Cell wall/membrane/envelope biogenesis	73
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	495
T	Signal transduction mechanisms	449
U	Intracellular trafficking, secretion, and vesicular transport	400
V	Defense mechanisms	32
W	Extracellular structures	8
Y	Nuclear structure	27
Z	Cytoskeleton	197
Metabolism		1534
C	Energy production and conversion	261
E	Amino acid transport and metabolism	256
F	Nucleotide transport and metabolism	76
G	Carbohydrate transport and metabolism	214
H	Coenzyme transport and metabolism	104
I	Lipid transport and metabolism	295
P	Inorganic ion transport and metabolism	132
Q	Secondary metabolites biosynthesis, transport and catabolism	196
Poorly characterized		1056
R	General function prediction only	741
S	Function unknown	315
Multiple families		575
X	multiple functions	1
Unknown		1994
Total		8292

KOG Category CBS 202.88		
Information Storage & Processing		1297
A	RNA processing and modification	256
B	Chromatin structure and dynamics	100
J	Translation, ribosomal structure and biogenesis	356
K	Transcription	376
L	Replication, recombination and repair	209
Cellular Processes & Signaling		1916
D	Cell cycle control, cell division, chromosome partitioning	187
M	Cell wall/membrane/envelope biogenesis	77
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	502
T	Signal transduction mechanisms	464
U	Intracellular trafficking, secretion, and vesicular transport	407
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	28
Z	Cytoskeleton	206
Metabolism		1585
C	Energy production and conversion	272
E	Amino acid transport and metabolism	266
F	Nucleotide transport and metabolism	79
G	Carbohydrate transport and metabolism	222
H	Coenzyme transport and metabolism	111
I	Lipid transport and metabolism	301
P	Inorganic ion transport and metabolism	135
Q	Secondary metabolites biosynthesis, transport and catabolism	199
Poorly characterized		1077
R	General function prediction only	755
S	Function unknown	322
Multiple families		599
X	multiple functions	1
Unknown		2029
Total		8504

KOG Category MR1448		
Information Storage & Processing		1275
A	RNA processing and modification	254
B	Chromatin structure and dynamics	95
J	Translation, ribosomal structure and biogenesis	352
K	Transcription	366
L	Replication, recombination and repair	208
Cellular Processes & Signaling		1875
D	Cell cycle control, cell division, chromosome partitioning	182
M	Cell wall/membrane/envelope biogenesis	73
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	498
T	Signal transduction mechanisms	454
U	Intracellular trafficking, secretion, and vesicular transport	398
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	27
Z	Cytoskeleton	198
Metabolism		1545
C	Energy production and conversion	262
E	Amino acid transport and metabolism	258
F	Nucleotide transport and metabolism	78
G	Carbohydrate transport and metabolism	218
H	Coenzyme transport and metabolism	103
I	Lipid transport and metabolism	298
P	Inorganic ion transport and metabolism	133
Q	Secondary metabolites biosynthesis, transport and catabolism	195
Poorly characterized		1054
R	General function prediction only	740
S	Function unknown	314
Multiple families		581
X	multiple functions	1
Unknown		1995
Total		8326

KOG Category D6		
Information Storage & Processing		1268
A	RNA processing and modification	253
B	Chromatin structure and dynamics	95
J	Translation, ribosomal structure and biogenesis	350
K	Transcription	362
L	Replication, recombination and repair	208
Cellular Processes & Signaling		1871
D	Cell cycle control, cell division, chromosome partitioning	183
M	Cell wall/membrane/envelope biogenesis	74
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	492
T	Signal transduction mechanisms	452
U	Intracellular trafficking, secretion, and vesicular transport	398
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	27
Z	Cytoskeleton	200
Metabolism		1542
C	Energy production and conversion	260
E	Amino acid transport and metabolism	258
F	Nucleotide transport and metabolism	76
G	Carbohydrate transport and metabolism	216
H	Coenzyme transport and metabolism	102
I	Lipid transport and metabolism	298
P	Inorganic ion transport and metabolism	133
Q	Secondary metabolites biosynthesis, transport and catabolism	199
Poorly characterized		1064
R	General function prediction only	748
S	Function unknown	316
Multiple families		581
X	multiple functions	1
Unknown		2006
Total		8333

KOG Category MR1459		
Information Storage & Processing		1273
A	RNA processing and modification	253
B	Chromatin structure and dynamics	94
J	Translation, ribosomal structure and biogenesis	353
K	Transcription	365
L	Replication, recombination and repair	208
Cellular Processes & Signaling		1877
D	Cell cycle control, cell division, chromosome partitioning	183
M	Cell wall/membrane/envelope biogenesis	74
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	497
T	Signal transduction mechanisms	452
U	Intracellular trafficking, secretion, and vesicular transport	401
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	27
Z	Cytoskeleton	198
Metabolism		1540
C	Energy production and conversion	260
E	Amino acid transport and metabolism	257
F	Nucleotide transport and metabolism	77
G	Carbohydrate transport and metabolism	216
H	Coenzyme transport and metabolism	102
I	Lipid transport and metabolism	298
P	Inorganic ion transport and metabolism	132
Q	Secondary metabolites biosynthesis, transport and catabolism	198
Poorly characterized		1064
R	General function prediction only	748
S	Function unknown	316
Multiple families		577
X	multiple functions	1
Unknown		2012
Total		8344

KOG Category IGIB-SBL-C11		
Information Storage & Processing		1262
A	RNA processing and modification	250
B	Chromatin structure and dynamics	95
J	Translation, ribosomal structure and biogenesis	348
K	Transcription	361
L	Replication, recombination and repair	208
Cellular Processes & Signaling		1862
D	Cell cycle control, cell division, chromosome partitioning	180
M	Cell wall/membrane/envelope biogenesis	73
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	495
T	Signal transduction mechanisms	448
U	Intracellular trafficking, secretion, and vesicular transport	397
V	Defense mechanisms	33
W	Extracellular structures	8
Y	Nuclear structure	27
Z	Cytoskeleton	197
Metabolism		1536
C	Energy production and conversion	261
E	Amino acid transport and metabolism	257
F	Nucleotide transport and metabolism	76
G	Carbohydrate transport and metabolism	214
H	Coenzyme transport and metabolism	103
I	Lipid transport and metabolism	294
P	Inorganic ion transport and metabolism	133
Q	Secondary metabolites biosynthesis, transport and catabolism	198
Poorly characterized		1051
R	General function prediction only	738
S	Function unknown	313
Multiple families		579
X	multiple functions	1
Unknown		1974
Total		8265

KOG Category CBS 735.88		
Information Storage & Processing		1264
A	RNA processing and modification	250
B	Chromatin structure and dynamics	97
J	Translation, ribosomal structure and biogenesis	349
K	Transcription	361
L	Replication, recombination and repair	207
Cellular Processes & Signaling		1865
D	Cell cycle control, cell division, chromosome partitioning	183
M	Cell wall/membrane/envelope biogenesis	74
N	Cell motility	4
O	Posttranslational modification, protein turnover, chaperones	493
T	Signal transduction mechanisms	452
U	Intracellular trafficking, secretion, and vesicular transport	394
V	Defense mechanisms	32
W	Extracellular structures	9
Y	Nuclear structure	26
Z	Cytoskeleton	198
Metabolism		1539
C	Energy production and conversion	261
E	Amino acid transport and metabolism	257
F	Nucleotide transport and metabolism	77
G	Carbohydrate transport and metabolism	216
H	Coenzyme transport and metabolism	103
I	Lipid transport and metabolism	298
P	Inorganic ion transport and metabolism	131
Q	Secondary metabolites biosynthesis, transport and catabolism	196
Poorly characterized		1055
R	General function prediction only	738
S	Function unknown	317
Multiple families		576
X	multiple functions	1
Unknown		2008
Total		8308

R code for heatmap

Set working directory

Set variable

```
data= read.csv("peptidase_families", sep=",")
```

Prepare data

```
row.names(data)=data$peptidases
```

```
data=data[,2:12]
```

Prepare matrix

```
data_matrix=data.matrix(data)
```

```
library("gplots")
```

Heatmap code

```
data_heatmap=heatmap.2(data_matrix, trace="none", density.info="none", symm=FALSE,  
cexRow=0.9, cexCol=0.75, col="redgreen")
```