MISS HAILIN ZHENG (Orcid ID : 0000-0003-0247-1853)

DR LIU WEI-DA (Orcid ID : 0000-0003-1936-3736)

Article type : Original Article

Assembly and analysis of the whole genome of *Arthroderma uncinatum* strain T10, compared with *Microsporum canis* and *Trichophyton rubrum*

Hailin Zheng[1,3], Oliver Blechert[1], Huan Mei[1], Liyu Ge[1], Jia Liu[1], Ye Tao[4], Dongmei Li[5], G.S. de Hoog[6], Weida Liu[1,2,3] *

1 Department of Medical Mycology, Institute of Dermatology, Chinese Academy of Medical Science and Peking Union Medical College, Nanjing, 210042, China

2 Center for Global Health, School of Public Health, Nanjing Medical University, Nanjing, 211166, China

3 Jiangsu Key Laboratory of Molecular Biology for Skin Diseases and STIs, Nanjing, Jiangsu, China

4 Shanghai Biozeron Biotechnology Co., Ltd, Shanghai, China

5 Department of Microbiology & Immunology, Georgetown University Medical Center, Washington, D.C. 20057, United States

6 Center of Expertise in Mycology of Radboud University Medical Center / Canisius Wilhelmina Hospital, Nijmegen, The Netherlands

Running title: Analysis of the whole genome of *A. uncinatum*

Corresponding Author

* Weida Liu:

Tel: +086 025 85478983. Fax: +086-025-85414477. E-mail: liumyco@hotmail.com.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process, which may lead to differences between this version and the Version of Record. Please cite this article as doi: 10.1111/MYC.13079

ABSTRACT

Background: *Arthroderma uncinatum* is a geophilic dermatophyte that occasionally causes superficial infections in humans leading to skin diseases.

Objectives: To better understand the ecology and potential pathogenicty of *A. uncinatum*, we analyzed its whole genome. We compared *A. uncinatum* with the genome of the zoophilic dermatophyte *Microsporum canis* and with the anthropophilic species *Trichophyton rubrum*. The compared species differ significantly in the frequence of human infection.

Methods: We reported the genome sequence of strain T10 of *A. uncinatum* based on SMRT (Single Molecule Real Time) technology (PacBio).

Results: We obtained a near-complete 23.56 Mb genome; with 7,153 predicted gene models and ~20 % repetitive sequences. We subsequently determined the specific genetic differences between *A. uncinatum, M. canis* and *T. rubrum.* The functional enrichment analysis suggests that *A. uncinatum* is particularly enriched in specific virulence genes. This suggests that the ancestral condition in dermatophytes is with high virulence, which has decreased in the course of evolution to enhance coexistence with animal or human hosts.

Keywords: dermatophytes, *Arthroderma uncinatum*, *Microsporum canis*, *Trichophyton rubrum*, skin infection

INTRODUCTION

Dermatophytes commonly cause superficial infections on skin, hair and nails due to their preference for keratin-rich substrates[1, 2]. This group of filamentous fungi is classified as anthropophilic, zoophilic and geophilic according to its natural habitat. Infected animals sometimes act as vectors to transmit zoophilic fungi to human hosts, which mostly leads to acute, highly inflammatory mycoses. Geophilic fungi reside in soil around the burrows of specific terrestrial mammals[3, 4]. Geophilic fungi are regarded as ecological ancestors of anthropophilic fungi, which corresponds with molecular phylogeny. According to the new taxonomy, dermatophyte species are grouped into seven genera: *Trichophyton*, *Epidermophyton*, *Nannizzia*, *Paraphyton*, *Lophophyton*, *Microsporum*, and *Arthroderma*. *Arthroderma* was highly diverse, containing well-resolved geophilic species. In older literature, the name *Arthroderma* was reserved for the sexual morph produced after heterothallic mating, but it is currently applied to the entire fungus. *Arthroderma uncinatum* only rarely infects humans[1, 5, 6]. The documented pathogenic role of this species has been limited to only a few cases of tinea corporis until today[5, 7, 8]. The incidence of infection by some other geophiles can be slightly higher[9], depending on the predominance of the species within a particular geographical region, and can be influenced by as population movement, socioeconomic circumstances and levels of scrutiny, in addition to intrinsic factors of virulence. We still lack comprehensive understanding of the health implications of geophilic dermatophytes on the human host.

Microsporum canis is a zoophilic dermatophyte which, when transmitted from pet animals to humans, causes multifocal alopecia, scaling, and circular lesions [10]. Its transmission occurs through direct contact with diseased or subclinically infected animals, mainly cats, via arthrospores[11]. Human-to-human infection has been frequently recorded in small, self-limited outbreaks. Asymptomatic animals are considered to be spreaders of the disease in about 50% of the infections in humans[12]. So the ecological niche of zoophilic dermatophyte is between geophilic and anthrophilic. *T. rubrum* is a typical representative of anthrophilic fungi, accounting for as many as 69.5 % of all dermatophytoses in humans[13, 14], and presents with tinea capitis, tinea corporis, tinea inguinalis, tinea manuum, tinea unguium, and tinea pedis. In addition to these

superficial infections, *T. rubrum* may also be responsible for some deep dermal infections in immunocompromised patients, which has promoted some mechanism-based studies of its pathogenesis[15-17], including the genome sequencing of *T. rubrum* recently published by our group. The availability of sequence data from other dermatophytes may assist us in outlining the evolutionary relationship among clonal species and characterize their individual pathogenic attributes. Today, long fragment reads (LFR) technique is employed in many studies for the sequencing of animals [18], plants[19], insects[20] and microorganisms[21, 22] . With LFR, whole genome sequences of several dermatophytes and subsequent genome analysis have identified some genes that may be associated with the pathogenesis of infection[15].

In the present study, we have used the PacBio technique to sequence the *A. uncinatum* genome. The assembly of the genome shows a high degree of integrity and accuracy. The comparison of the whole genome sequence of *A. uncinatum* with those of *M. canis* and *T. rubrum*, together with functional enrichment analysis, suggest that *A. uncinatum* is enriched for several specific genes involved in pathogenicity, such as the secondary metabolite biosynthetic process, the toxin metabolic process, the toxin biosynthetic process, and the mycotoxin metabolic process. In contrast, *M. canis* and *T. rubrum*, when compared with *A. uncinatum*, lost several specific genes.

MATERIALS AND METHODS

Sample collection

Arthroderma uncinatum strain T10 (CBS 119779) was isolated from a 60-year-old male patient with recurrent onychomycoses, in Italy. The strain was cultured on potato dextrose agar (PDA) for 7–14 days at 27 ℃ prior to mycelial collection. The strain was stored at –80 ℃ in the Collection Center of Pathogenic Microorganisms (Medical Fungi sub-center), Chinese Academy of Medical Sciences, Nanjing, China.

DNA Extraction and Sequencing

Genomic DNA was extracted from *A. uncinatum* strain T10 by using QIAamp DNA Mini Kit (Qiagen, Hilden, Germany) according to manufacturer's instructions. Then the DNA quantification was performed by qubit 3.0 (>30 ng/μL, ≥ 2 μg, $OD_{260/280}$ = 1.8~2.0; Life Invitrogen). The qualified DNA was used for Pacbio platform sequencing and construction of a small fragment library for Illumina sequencing.

For Pacific Biosciences sequencing, a 20 kb whole genome shotgun library was generated before sequencing on a Pacific Biosciences RSII instrument using standard methods. An aliquot of 8 μg DNA was spun in a Covaris g-TUBE (Covaris, MA) at 6,000 RPM for 60 seconds using an Eppendorf 5424 centrifuge (Eppendorf, NY). Then the DNA fragments were purified, end-paired and ligated with SMRT bell sequencing adapters as per manufacturer's recommendations (Pacific Biosciences, CA). Resulting sequence library was purified three times using 0.45 x volumes of Agencour tAMPureXP beads (Beckman Coulter Genomics, MA) according to manufacturer's recommendations.

We also performed Illumina pair-end sequencing (PE150, Hiseq X Ten) for the assembly correction. At least 2 μg genomic DNA was used for library construction of each sample. Paired-end libraries with insert sizes of ~400 bp were prepared by Illumina's standard genomic DNA library method (TruSeq® Nano DNA Library Prep). The quantified Illumina paired-end library was used for PE150 sequencing with Illumina Hiseq X Ten platform. All the experiments and sequencing were performed at Biozeron Co. (Shanghai, China).

Genome Assembly and Gene Annotation

In total, ~10.77 Gb PacBio data with an average length of 8.465 kb per read and 26 Mb Illumina reads (PE150, Illumina) were generated. Reads with adaptors with a percentage of unknown nucleotides of >5 %, a percentage of low-quality bases (base quality < 20) of more than 20 %, and Q30 of clean data 97.59 % or a percentage of N more than 10 %, were filtered out using the software Trimmomatic v0.39, leaving 25.9 Mb clean reads with nearly 3.8 Gb clean bases.

After filtering polymerase reads with lengths shorter than 100 bp or quality lower than 0.8 and deletion of adaptor sequences, clean subreads were obtained. The *A. uncinatum* genome was assembled using the PacBio data of strain T10 by Canu (v1.8, https://github.com/marbl/canu, parameter settings: genome size = 23 m , min read length = 2000, min overlap length = 500)[23]. Reads were mapped to the genome by BWA mem mode (v0.7.17, parameter settings: -k 32, others are defaulted). Then, three rounds of Pilon (v1.22, https://github.com/broadinstitute/pilon, parameter settings: fix SNPs, indels)[24] were run performing the sequence correction with short reads from the Illumina platform. Finally, the genomic integrity and accuracy were estimated by BUSCO (v3.0.2)[25]. The sequence of *A. uncinatum* T10 was submitted to NCBI (ID: PRJNA526381).

We followed *ab initio* prediction methods to obtain the gene models for *A. uncinatum*. Gene models were identified using Augustus (v3.2.3, http://bioinf.uni-greifswald.de/augustus/binaries/)[26]. Then all the gene models were blasted (E-value < 1e-5) against non-redundant (NR in NCBI) database, SwissProt (http://uniprot.org), KEGG (http://www.genome.jp/kegg/)[27], and COG (http://www.ncbi.nlm.nih.gov/COG)[28] to analyze functional annotation by blastp module. The terms of gene ontology (GO) classification were performed by Blast2GO[29, 30]. In addition, tRNA were identified using the tRNAscan-SE (v1.23), http://lowelab.ucsc.edu/tRNAscan-SE)[29] and rRNA[1], using the RNAmmer (v1.2, http://www.cbs.dtu.dk/services/RNAmmer/).

Identification of clusters of orthologous genes

Clusters of orthologous genes between species were identified using OrthoFinder (https://github.com/davidemms/OrthoFinder, v2.2.7). All other parameters were used in the default settings. Strain-specific gene modules were extracted from OrthoFinder output by PERL scripts provided in OrthoFinder software package.

Phylogenetic analysis of orthologous genes

In this study, seven dermatophytes (*A. uncinatum* T10, *Microsporum canis* CBS 113480, *Nannizzia gypsea* CBS 118893, *Trichophyton mentagrophytes* TIMM 2789, *T. rubrum* CBS 139224, *T. violaceum* CMCC(F)T3l, *T. benhamiae* CBS 112371) were compared for phylogenetic analysis, with *Aspergillus fumigatus* Z5 as outgroup. Orthologous gene families were identified by ORTHOMCL v2.0 (reciprocal all-by-all BLASTP analysis) with an E-value of 10^{-5}. Multiple alignments were generated with the MUSCLE v3.8.31, and the alignments were examined visually. Maximum-likelihood (ML) method was performed for phylogenetic analyses using PhyML v3.0 using model GTR + I + G and with 500 bootstrap replicates.

Mitochondrial genome assembly and annotation

The mitochondrial genomes of *A. uncinatum* T10 were assembled combining the illumina and Pacbio data using SPAdes v3.13.1 (http://cab.spbu.ru/software/spades/, parameter setting: pacbio m 250). Circularization of mitochondrial genome sequences was done manually. After removing the consensus sequences (>200 bp overlap in *cox1* gene), a mitochondrial genome was assembled as a single circular dsDNA.

The initial mitochondrial genome annotations were done using MFANNO (http://megasun.bch.umontreal.ca/, parameter setting: default) and were manually curated. Annotation of tRNA genes was improved using tRNA scan-SE v2.0.3, annotation of intron-exon boundaries was improved by comparing to the reference genome of *T. rubrum* (FJ385026). Intron encoded proteins were identified using NCBI's ORF Finder (https://www.ncbi.nlm.nih.gov/orffinder/, parameter setting: minimal ORF length = 75 bp, genetic code = 4) and annotated by BLASTP (v2.7.1) with NCBI mitochondrial genome database

(ftp://ftp.ncbi.nlm.nih.gov/refseq/release/mitochondrion/).

Adhesin prediction

For adhesin prediction, three softwares were used based on earlier reports, including SignalP4.0 server (http://www.cbs.dtu.dk/services/SignalP) for the signal peptide prediction, Big-PI predictor (http://mendel.imp.ac.at/sat/gpi/gpi_server.html) for GPI anchor sites, and TMHMM Server v2.0 ((http://www.cbs.dtu.dk/services/TMHMM//) for the transmembrane helix. In order to provide best combinations, total results were evaluated.

Secondary metabolite biosynthesis

In this study, we predicted several potential secondary metabolite biosynthesis gene clusters by the ANTIbiotic and Secondary Metabolite Analysis Shell (ANTISMASH v4.1.0, Weber et al. 2015).

RESULTS

Sequence analysis and *de novo* assembly

A total of 4.01 GB bases of *A. uncinatum* T10 were generated by HiseqX Ten with PE150 mode. The Q30 value of raw reads was 92.63 %. After data QC and filtering, nearly 3.85 Gb NGS clean bases were used in the assembly (Table 1a). In addition, 1,272,620 PacBio reads (N50 = 11,535 bp, totally 10.77 Gb bases) were obtained (Table 1b). We applied the entire subreads longer than 500 bp for subsequent genome assembly and analysis.

A chromosome-level assembly of *A. uncinatum* strain T10 (CBS 119779) was obtained with a 23.56 Mb genome (Contig N50 = 6.24 Mb) consisting of 5 non-redundant contigs. In order to assess the completeness of the assembly and the prediction of genes, we analyze the genome by Benchmarking Universal Single-Copy Orthologs (BUSCOs). We found that ~97.3 % were complete and single-copy, with only ~1.6 % BUSCOs missing. The average length of these contigs was more than 4.71 Mb, with the longest chromosome being 6.63 Mb and the shortest 2.04 Mb. The G+C content of entire genome was 48.59 % and the ambiguous bases accounted for less than 0.01 %. We predicted 7,153 genes (Table 2), covering a total of 11,843,178 bp of the genome with a mean length of 1,656 bp per gene. The GC content was 51.82 % in coding-gene regions (CDS) that accounted for 50.26 % of the genome and GC content of 46.03 % in a total of 11,720,078 bp inter-genic regions that constituted for 49.74 % of the whole genome.

In addition, the mitochondrial genome of *A. uncinatum* T10 was successfully assembled into a single circular DNA molecule with a genome size of 30,449 Kb and a GC content of 23.52 %. The mitochondrial genes encoded for 15 proteins, a small and large subunit of the ribosome (*rns* and *rnl*, respectively) and 25 tRNAs.

Functional annotation

We performed the functional annotation and KEGG pathway analysis for 7,153 genes (Table 2). On the basis of sequence homology using BLASTP software, the gene sequences were searched against the NR, the KEGG and the clusters of orthologous groups (COG) database (E-value < 1e-5). The terms of gene ontology (GO) classification were carried out with Blast2GO

program. The statistics of functional annotation of all the genes in public databases is shown in Table 3.

A total of 7,034 (98.34 % of all genes) were matched once or more than once in these databases (Table 3), for which 98.29 % match efficiency was observed in NR databases, 80.33 % in COG databases, 61.53 % in Swiss-Prot databases, 69.51 % in GO database and 40.95 % in KEGG databases. However, genes of more than 2.4 Kb in size had only 34.6 % efficiency when matched to the databases mentioned above.

With NCBI NR protein database, 7,031 genes were annotated. Only 5,746 genes were positively corresponding to orthologs in COG database (Fig. 1A). Among all 25 categories of COG, O (Posttranslational modification, protein turnover, chaperones); J (Translation, ribosomal structure and biogenesis); E (amino acid transport and metabolism) and G (carbohydrate transport and metabolism) were then screened for genes which are more likely related to pathogenesis.

An alignment between genes and KEGG database was performed to obtain a deeper understanding of the biological pathways operating in *A. uncinatum* T10. A total of 2,929 genes were included within 311 pathways, including 671 genes that relate to metabolic pathways (ko 01100) and 289 genes involved in biosynthesis of secondary metabolites (ko 01110). Other genes were more similar to microbial metabolism in diverse environment (ko 01120, 161 genes) and biosynthesis of amino acids (ko 01230, 116 genes) (Fig. 1B).

GO assignments were also used to classify the functions of the predicted genes of *A. uncinatum* T10. Based on sequence homology, 4,972 genes were categorized into 3 functional groups (biological process, cellular component and molecular function) (Fig. 1C). High assignments within the cellular component were cell (2,386, 47.99 %) and cell part (2,374, 47.75 %), followed by the organelle term (1,894, 38.09 %). The dominant terms in the biological process category were metabolic process (2,845, 57.22 %), cellular process (2,836, 57.04 %) and single-organism process (2,243, 45.11 %). Within the molecular function category, binding (2,145, 43.14 %) and catalytic activity (2,515, 50.58 %) were highly represented.

Comparative genomics

In order to maintain consistency of comparisons among species, published genomes were analyzed for gene prediction here again. Earlier studies (Zhan et al. 2017; Wu et al. 2009; de Hoog et al. 2017) reported on evolutionary relationships of *A. uncinatum* to zoo- and anthropophilic dermatophytes based on the ITS region or the mitochondrial genome. We also constructed a phylogenetic tree of several common dermatophytes based on some single-copy genes (Fig. 2). Relationships between *A. uncinatum* and *M. canis* (CBS 113480) was closer than with *Nannizzia* and *Trichophyton* fungus; we therefore performed a genomic comparison between *A. uncinatum* T10 and *M. canis* (NCBI Accession ID: GCA_000151145.1). Genome sizes were found to be closely similar: *A. uncinatum* was 23.56 Mb, while *M. canis* was 23.26 Mb. Based on the BUSCOs analysis, both had high genome integrity (*A. uncinatum* ~97.3 %, *M. canis* ~96.8 %). The total gene number of the two species was very close, with 7,153 gene families in *A. uncinatum*, and 8,765 in *M. canis*; 6,269 were homologous including 6,036 single-copy genes. The number of unique gene families was 708 in *A. uncinatum*, and 2,225 in *M. canis*. The most represented GO term of these specific genes in both species belonged to biological process. In *A. uncinatum*, the majority of unique genes were related to secondary metabolic process, secondary metabolite biosynthetic process, toxin metabolic process, toxin biosynthetic process, mycotoxin metabolic process, mycotoxin biosynthetic process, and others (Table 4a). In *M. canis,* these unique genes were enriched in only two GO terms: transcription factor catabolic process, and negative regulation of transcription by transcription factor catabolism (Table 4b).

Based on the analysis of these specific-genes in the KEGG Database, no biological pathway was enriched in *A. uncinatum*. The only pathway enriched in *M. canis* was focused on the ribosome term (corrected P-value 0.01211).

Trichophyon rubrum is the most commonly observed human dermatophyte, and geophilic dermatophytes are regarded as genetic ancestors of anthropophilic dermatophytes. To better understand the potential pathogenic impact of *A. uncinatum* T10, we compared its genome with *T. rubrum* CBS 139224 (NCBI Accession ID: GCA_001651445.1). *Arthroderma uncinatum* had a slightly larger genome size (23.56 Mb) with less scaffolds (5) when compared to strain *T. rubrum* of 22.30 Mb (19 scaffolds). N50 of T10 genome was much longer than *T. rubrum.* Analysis of

BUSCOs of *T. rubrum* (~98.7 %), both strains were revealed to have large genomic integrity and quality.

As the genomes of *A. uncinatum* and *T. rubrum* CBS 139224 were sequenced using a combination of PacBio RS and Illumina platforms, comparative genomic analysis between the two species is warranted. After statistical analysis, a total of 8,643 gene families were found in the analyzed strains, 6,069 of which concerned homologous gene families. CBS 139224 of *T. rubrum* had more specific genes (1,665) than *A. uncinatum* T10 (909). Differences in functions and pathways were noted. In *A. uncinatum*, the function of specific-genes were mainly related to 5 GO terms, i.e. the monopolin complex, the regulation of clathrin-dependent endocytosis, the positive regulation of clathrin-dependent endocytosis, the regulation of protein localization by the Cvt pathway and the isoprenoid metabolic process (Table 4c). Analysis of biological pathways of the strain-specific genes in the KEGG database showed that none was enriched in *A. uncinatum*, whereas the ribosome (corrected P-value, ~0.02712) was the only one enriched in *T. rubrum*.

Comparison of mitochondrial genomes

The single circular contigs of mitochondrial genomes of *A. uncinatum, M. canis* and *T. rubrum* are shown in Fig. 3. The genes in all three mitochondrial genomes are in the same orientation and order, suggesting a high degree of conservation. Like in *M. canis* and *T. rubrum*, *nad6*, *cox3* and *atp8* are located between the *rns* and the *rnl* in *A. uncinatum.* The most striking difference is that strain *T. rubrum* has longer ORF insertions, such as *orf752* between *nad2* and *cob*, *orf277* downstream the *cox1, orf174* between *cob* and *nad3*, and *orf531* between *nad4* and *atp6* in the mitochondrial genome of *T. rubrum* The *rnl* gene contains a group I intron that codes for *rps3* in *A. uncinatum*, while it codes for *rps5* in *M. canis*. Some tRNA-coding genes are found downstream *rps3* in *T. rubrum* and *rps5* in *M. canis*, which were absent from *A. uncinatum*. The mitochondrial genome of T10 contains a ribozymal gene *rnpB* that is lacking in *T. rubrum*.

The mitochondrial genome size of *A. uncinatum* has been published to be 28,530 bp[2], which was smaller than the 30,449 bp in the present study. Comparing homologous genomic regions between these two mitochondrial genomes (Fig. 4), it was noticed that there is an expansion of the

genome from ~25 Kb to ~27 Kb with nearly 1.9 Kb in T10, which was missed in the previous reports. This part was annotated as *orf531*, a hypothetical protein in *A. uncinatum* T10. The difference between these two mitochondrial genomes might be caused application of by different strains of *A. uncinatum*.

Adhesion

Three web servers or platforms with fungal adhesin predictors were employed for analysis of putative adhesion-like proteins, which were chosen with the following parameters: SignalP 4.0 positive; TMHMM 2.0 zero; Big-PI Predictor positive. A total of 23 adhesins were present in *A. uncinatum* T10, as listed in Table S1. This number is compatible with that found in other dermatophytes. Most *Trichophyton* species have 20 to 26 adhesins, while some other dermatophytes have less; for example, *Nannizzia gypseum* CBS 118893 carried only 5 adhesins [1].

Nearly half of the adhesins in *A. uncinatum* T10 were hypothetical and uncharacterized proteins. The extracellular serine-threonine rich protein existed in both *A. uncinatum* T10 and *M. canis* CBS 113480. In *A. uncinatum*, the adhesins were mainly GPI anchored serine-rich proteins, i.e. ecm33, acid phosphatase, chitinase3, extracellular matrix protein, Cu-Zn superoxide dismutase, lysophospholipase Plb2 and class III chitinase.

In previous studies, 21 and 20 adhesins were predicted in *T. rubrum* CBS 139224 and CBS 118892, respectively. They are serine carboxypeptidase, glycolipid-anchored surface proteins, polysaccharide deacetylase family proteins, aspartyl protease and mannosyl-oligosaccharide 1,2-alpha-mannosidase. Similar to *T. rubrum* CBS 139224, most of the adhesins in *A. uncinatum* T10 wre hypothetical proteins as well. However, we noticed that *A. uncinatum* strain some specific adhesins different from *T. rubrum*, such as acid phosphatase, extracellular serine-threonine rich protein, extracellular matrix protein, class V chitinase and chitinase 3, Cu-Zn superoxide dismutase, lysophospholipase Plb2, ecm33 and GPI anchored serine-rich protein.

Secondary metabolism

Secondary metabolites were compared between *A. uncinatum* T10 and *T. rubrum* CBS

139224 (Table S2). In total, 15 metabolite clusters were present in *A. uncinatum,* but only 9 clusters were found in *T. rubrum*. Five metabolite clusters were common to both fungi, including a cluster "Citrinin", which showed only 12 % identity of T10 of *A. uncinatum* with the reference strain. Of 10 unique clusters in strain T10, three were conserved in other fungi. These clusters were identified as pyranonigrin E, HC-toxin and clavaric acid. The existence of Fusaridione A in two scaffolds are not certain since identity rates were both 12 % when compared to the reference strain. Compared with *A. uncinatum* T10, *T. rubrum* had only four unique clusters, of which hexadehydro-astechrome (HAS) biosynthetic gene cluster and neosartoricin biosynthetic gene cluster showed high similarity with the reference genome (87 % and 100 %, respectively). Two other clusters, the penicillin biosynthetic cluster and the huperzine gene cluster, showed only 18 % and 7 % identity with reference sequences.

DISCUSSION

Research on geophilic fungi is relatively limited. These fungi cause only accidental infections on the human host; they have been hypothesized as ancestral dermatophytes which have not yet experienced evolutionary selection to optimize host/pathogen interactions. Due to their limited adaptation to the habitat of human skin, geophilic species generally cause acute, severe diseases[31, 32]. In this respect, zoophilic fungi are similar, as their frequency of interacting with human hosts is mostly lower than for anthropophilic species. Anthropophilic fungi have developed more subtle interactions with the human immune system [33].

Arthroderma uncinatum resides in soil without a known connection to any host animal; occasional human skin infections have been reported[34]. In our study we have shown a near-complete *de novo* assembly of the genome of *A. uncinatum* T10 using data from Illumina Hiseq and Pacbio platforms. Combined with short read and long read technologies, we obtained excellent continuity and completeness of the assembly, with contig N50 reaching 11,535 bp. Such a high quality genome can provide a good reference for further research in the evolution of dermatophytes, and ultimately also in clinical diagnosis and treatment of skin infections. Given the ancestral position of *A. uncinatum*, we compared the species with *M. canis* and *T. rubrum* as representatives of zoophilic and anthropophilic dermatophytes, respectively, to analyze and explore the differences of some functional genes in an evolutionary perspective.

The number of unique gene families was much larger in *M. canis* than in *A. uncinatum*. The most common GO term of these specific genes in both strains belonged to biological process. However, compared to *M. canis*, *A. uncinatum* was enriched e.g. in genes were related to toxin metabolic process, toxin biosynthetic process, mycotoxin metabolic process, and mycotoxin biosynthetic process. *Arthroderma uncinatum* also had a large number of genes related to virulence. This suggests that the ancestral condition in dermatophytes is with high virulence, which has decreased in the course of evolution to enhance coexistence with animal or human hosts. In *M. canis,* the unique genes were enriched in only two GO terms, i.e. the transcription factor catabolic process and the negative regulation of transcription by transcription factor catabolism. Our study further showed that the genomes of *T. rubrum* and *A. uncinatum* were similar, but that

the genome of *T. rubrum* was slightly smaller, the anthropophilic species possibly having lost some genes that triggered immune response.

Mitochondria are generally judged to be descendants of endosymbiotic alpha-proteobacteria and are considered to be of monophyletic origin. As vital physiological processes and basic adaptive strategies do not always correlate with trees derived from ribosomal sequences, mitochondrial DNA (mtDNA) sequences have become a popular tool for phylogenetic studies[35, 36]. In our study, the genes in both genomes are in the same orientation and order and show complete conservation, indicating that *A. uncinatum* and *T. rubrum* are closely related. This is consistent with earlier studies [2]. In adhesion factors and secondary metabolism, some additional deviant proteins and pathways between *T. rubrum* and *A. uncinatum* were found, which may play a role in host selection. These aspects should be studied further in the future.

In our earlier studies, we found that growth of *A. uncinatum* was significantly affected by temperature, the species being unable to grow at 34 °C or above. The optimum temperature for *A. uncinatum* is 25 °C, which is more suitable for survival in the natural environment. However, *T. rubrum* still thrives at 35 °C and above, enabling the fungus to cope with the temperature of human skin. *Nannizzia gypsea,* which is often considered as a geophile but has also the ability to grow at 35 °C, is found on human hosts much more frequently than *A. uncinatum*[34]. It may be speculated that under the influence of global warming, *A. uncinatum* may become more thermotolerant, enhancing its infective ability of humans.

This is the first reported description of the whole genome of the geophilic fungus *A. uncinatum*. Comparing the whole genome of *M. canis* and *T. rubrum* to that of *A. uncinatum*, some discrepancies of these two fungi were found in genes, proteins and pathways. Through these differences, we believe it will be possible to more closely study the dermatophytes and also better analyze the relationship between anthropophilic, zoophilic and geophilic fungi. This should lead to better understanding of the relationship of pathogenic dermatophytes and the human host.

ACKNOWLEDGEMENT

This work was financially supported by The National Natural Science Foundation of China (No. 81972949)， The Scientific and Technological Innovation Projects of Medicine and Health of Chinese Academy of Medical Sciences (No. 2016-I2M-3-021), The Basical Scientific Research Fund Projects of Chinese Academy of Medical Sciences (No. 2018PT31013)， The Basic Scientific Research Fund Projects of Chinese Academy of Medical Sciences (Young Teacher foundation, No. 3332018121) and National Science and Technology Major Project (No. 2018ZX10734404).

COMPLIANCE WITH ETHICAL STANDARDS

Not applicable

CONFLICT OF INTEREST

The authors declare no competing financial interest.

STATEMENT OF AUTHOR CONTRIBUTIONS

Hailin Zheng and Weida Liu conceived the ideas; Huan Mei, Liyu Ge, Jia Liu collected the data; Oliver Blechert, Ye Tao, Hailin Zheng and G.S. de Hoog analysed the data; Hailin Zheng, Dongmei Li and G.S. de Hoog led the writing.

REFERENCES

1 Zhan P, Dukik K, Li D, et al. Phylogeny of dermatophytes with genomic character evaluation of clinically distinct Trichophyton rubrum and T. violaceum. *Studies in Mycology* 2018;89: 153-75.

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

3 Kachuei R, Emami M, Naeimi B, Diba K. Isolation of keratinophilic fungi from soil in Isfahan province, Iran. *Journal de Mycologie Médicale* 2012;22: 8-13.

4 Ciesielska A, Bohacz J, Korniłłowicz-Kowalska T, Stączek P. Microsatellite-Primed PCR for Intra-species Genetic Relatedness in Trichophyton ajelloi Strains Isolated in Poland from Various Soil Samples. *Microbes and Environments* 2014;29: 178-83.

5 de Hoog GS, Dukik K, Monod M, et al. Toward a Novel Multilocus Phylogenetic Taxonomy for the Dermatophytes. *Mycopathologia* 2017;182: 5-31.

6 Kosanke S, Hamann L, Kupsch C, Moreno Garcia S, Chopra A, Graser Y. Unequal distribution of the mating type (MAT) locus idiomorphs in dermatophyte species. *Fungal Genetics and Biology* 2018;118: 45-53.

7 Alvarez Dp Fau - de Bracalenti BJ, de Bracalenti BJ. Dermatophytosis caused by Trichophyton ajelloi. *Mycopathologia* 1982;77: 27-9.

8 Presbury Dg Fau - Young CN. Trichophyton ajelloi isolated from a child. *Sabouraudia* 1978;16: 233-5.

9 Bassiri-Jahromi S. Epidemiological trends in zoophilic and geophilic fungi in Iran. *clinical and expermental dermatology* 2013;38: 13-9.

10 Aneke CI, Otranto D, Cafarchia C. Therapy and Antifungal Susceptibility Profile of Microsporum canis. *J Fungi (Basel)* 2018;4: 1-14.

11 Arrese JE, Martalo O Fau - Pierard-Franchimont C, Pierard-Franchimont C Fau - Pierard GE, Pierard GE. Urban and rural mycozoonoses. *Rev Med Liege* 2000;55: 998-1002.

12 Mancianti F, Nardoni S Fau - Corazza M, Corazza M Fau - D'Achille P, D'Achille P Fau - Ponticelli C, Ponticelli C. Environmental detection of Microsporum canis arthrospores in the households of infected cats and dogs. *J Feline Med Surg* 2003;5: 323-8.

13 Weitzman I, Summerbell RC. The dermatophytes. *Clinical microbiology reviews* 1995;8: 240-59.

14 Chen BK, Friedlander SF. Tinea capitis update: a continuing conflict with an old adversary. *Current opinion in*

pediatrics 2001;13: 331-5.

15 Martinez DA, Oliver BG, Graser Y, et al. Comparative genome analysis of Trichophyton rubrum and related dermatophytes reveals candidate genes involved in infection. *mBio* 2012;3: e00259-12.

16 Persinoti GA-O, Martinez DA-O, Li W, et al. Whole-Genome Analysis Illustrates Global Clonal Population Structure of the Ubiquitous Dermatophyte Pathogen Trichophyton rubrum. *Genetics* 2018;208: 1657-69.

17 Zhan P, de Hoog S, Liu W. Draft Genome Sequences of Trichophyton rubrum CMCC(F)T1i and Trichophyton violaceum CMCC(F)T3l by Illumina 2000 and Pacific Biosciences. LID - e00920-17 [pii] LID - 10.1128/genomeA.00920-17 [doi]. *Genome announcements* 2017;5: 1-2.

18 Bickhart DM, Rosen BD, Koren S, et al. Single-molecule sequencing and conformational capture enable de novo mammalian reference genomes. *bioRxiv p* 2016: 1-31.

19 Li MA-Ohoo, Zhang D, Gao Q, et al. Genome structure and evolution of Antirrhinum majus L. *Nature plants* 2019;5: 174-83.

20 Matthews BJ, Dudchenko O, Kingan SB, et al. Improved reference genome of Aedes aegypti informs arbovirus vector control. *nature* 2018;563: 501-7.

21 Tedersoo L, Tooming-Klunderud A, Anslan S. PacBio metabarcoding of Fungi and other eukaryotes: errors, biases and perspectives. *the new phytologist* 2018;217: 1370-85.

22 Jason L Weirather MdC, Yunhao Wang,Paolo Piazza,Vittorio Sebastiano,Xiu-Jie Wang,David Buck,Kin Fai Au. Comprehensive comparison of Pacific Biosciences and Oxford Nanopore Technologies and their applications to transcriptome analysis. *F1000Research* 2017.

23 Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 2017;27: 722-36.

24 Bruce J. Walker TA, Terrance Shea, Margaret Priest, Amr Abouelliel,, Sharadha Sakthikumar CAC, Qiandong Zeng, Jennifer Wortman, Sarah K. Young,, Earl AM. Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *plos one* 2014;9: 1-14.

25 Simao FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 2015;31: 3210-2.

26 Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* 2011;27: 757-63.

27 Kanehisa M, Goto S Fau - Sato Y, Sato Y Fau - Furumichi M, Furumichi M Fau - Tanabe M, Tanabe M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* 2012;40: 109-14.

28 Roman L. Tatusov MYG, Darren A. Natale and Eugene V. Koonin. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* 2000;28: 33-6.

29 Eddy TMLaSR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research* 1997;25: 955-64.

30 Shubo Jin HF, Qiao Zhou, Shengming Sun, Sufei Jiang, Yiwei Xiong,, Yongsheng Gong HQ, Wenyi Zhang. Transcriptome Analysis of Androgenic Gland for Discovery of Novel Genes from the Oriental River Prawn, Macrobrachium nipponense, Using Illumina Hiseq 2000. *plos one* 2013;8: 1-13.

31 Souza BD, Sartori DS, Andrade C, Weisheimer E, Kiszewski AE. Dermatophytosis caused by Microsporum gypseum in infants: report of four cases and review of the literature. *An Bras Dermatol* 2016;91: 823-5.

32 Lee WJ, Park JH, Kim JY, et al. Low But Continuous Occurrence of Microsporum gypseum Infection in the Study on 198 Cases in South Korea from 1979 to 2016. *Ann Dermatol* 2018;30: 427-31.

33 White TC, KF, TLDJ, et al. Fungi on the skin dermatophytes and Malassezia. *Cold Spring Harbor Perspectives in Medicine* 2014;1: 1-16.

34 de Hoog GS GJ, Genéˊ J, Ahmed SA, Al-Hatmi AMS, Figueras MJ, Vitale RG. Atlas of Clinical Fungi，4th e-edition 2019.

35 Andersson SG, Karlberg O, Canback B, Kurland CG. On the origin of mitochondria: a genomics perspective. *Philos Trans R Soc Lond B Biol Sci* 2003;358: 165-77.

36 Tambor JH, Guedes Rf Fau - Nobrega MP, Nobrega Mp Fau - Nobrega FG, Nobrega FG. The complete DNA sequence of the mitochondrial genome of the dermatophyte fungus Epidermophyton floccosum. *Current genetics* 2006;49: 302-8.

Tables and legends

Table 1a. Statistics of Illumina sequencing. The aver coverage is defined as the average depth that the clean bases covered the genome.

Statistics of Illumina data	T10 strain（CBS119779）
Raw reads	26,722,734
Raw bases(bp)	4,008,410,100
Clean reads	25,895,797
Clean bases(bp)	3,849,216,310
Average coverage	163

Table 1b. Statistics of Pacbio sequencing. N50 here is a measure to describe the reads quality of Pacbio sequencing data that are fragmented in reads of different length. The N50 is defined as the minimum reads length needed to cover 50 % of the total sequencing bases.

Statistics of PacBio data	T10 strain (CBS119779)
Total reads num	1,272,620
Total bases (bp)	10,772,527,788
Largest length (bp)	67,012
N50 length (bp)	11,535
N90 length (bp)	5,552
Average length (bp)	8,465

Table 2. Statistics of T10 and other two species assembly. Gene internal length means the total length for the intergenic regions at the whole genome level.

Statistics of predicted gene	***A. uncinatum* T10 strain**	***M. canis* CBS 113480**	***T. rubrum* CBS139224**
Genome size	23,563,256	23,263,091	22,301,977
Gene number:	7,153	8,765	8,173
Gene total length(bp):	11,843,178	12,802,803	12,251,022
Gene's GC content:	51.82%	50.74%	50.96%
% of genome(Genes):	50.26%	55.03%	54.93%
Gene average lengt(bp):	1,656	1,461	1,499
Intergenic region length(bp):	11,720,078	10,460,288	10,050,955
Intergenic's GC content:	46.03%	43.47%	45.15%
% of genome(internnic):	49.74%	44.97%	45.07%

Table 3. Summary statistics of functional annotation of *A. uncinatum* T10 strain genes in public databases. NR: the non-redundant protein database in NCBI. GO: Gene Ontology, the framework for the model of biology which defines concepts/classes used to describe gene function. COG: Clusters of Orthologous Groups is an attempt on a phylogenetic classification of the proteins (https://www.ncbi.nlm.nih.gov/COG/). KEGG: KEGG is a database resource for understanding high-level functions and utilities of the biological system. SWISS: the manually annotated and reviewed section of the UniProt Knowledgebase.

DataBase_name	**Total_genes**	**annotated_genes**	**percent**
NR	7153	7031	0.9829
GO	7153	4972	0.6951
COG	7153	5746	0.8033

KEGG	7153	2929	0.4095
SWSS	7153	4401	0.6153
In all DB	7153	2475	0.346
AT least one DB	7153	7034	0.9834

Table 4 Enrichment of GO term of unique genes.

Table 4a: Gene Ontology (GO) terms enrichment analysis of T10 strain unique genes compared with *M. canis* CBS 113480.

ID	Description	Ratio in study	Ratio In pop	P value (p_fdr)	GO Type
GO:0019748	secondary metabolic process	55/708	261/7153	0	biological_process
GO:0044550	secondary metabolite biosynthetic process	42/708	156/7153	0	biological_process
GO:0009404	toxin metabolic process	41/708	168/7153	0	biological_process
GO:0009403	toxin biosynthetic process	30/708	84/7153	0	biological_process
GO:0043385	mycotoxin metabolic process	26/708	62/7153	0	biological_process
GO:0043386	mycotoxin biosynthetic process	24/708	58/7153	0	biological_process
GO:1901378	organic heteropentacyclic compound biosynthetic process	13/708	34/7153	0.008	biological_process
GO:1901376	organic	13/708	35/7153	0.018	biological_process

	heteropentacyclic compound metabolic process				
GO:0017000	antibiotic biosynthetic process	13/708	36/7153	0.024	biological_process
GO:0006720	isoprenoid metabolic process	13/708	38/7153	0.05	biological_process
GO:0004659	prenyltransferase activity	11/708	27/7153	0.026	molecular_function
GO:0033551	monopolin complex	8/708	12/7153	0.002	cellular_component
GO:2000369	regulation of clathrin-dependent endocytosis	8/708	12/7153	0.002	biological_process
GO:2000370	positive regulation of clathrin-dependent endocytosis	8/708	12/7153	0.002	biological_process
GO:2001159	regulation of protein localization by the Cvt pathway	8/708	12/7153	0.002	biological_process
GO:0038083	peptidyl-tyrosine autophosphorylation	8/708	14/7153	0.018	biological_process
GO:0003964	RNA-directed DNA polymerase activity	6/708	8/7153	0.026	molecular_function

Table 4b: Gene Ontology (GO) terms enrichment analysis of *M. canis* CBS 113480 unique genes compared with *A.uncinatum* T10.

ID	**Description**	**Ratio in**	**Ratio In pop**	**P value (p_fdr)**	**GO Type**

		study			
GO:0036369	transcription factor catabolic process	17/2225	23/8765	0.002	biological_process
GO:0010620	negative regulation of transcription by transcription factor catabolism	13/2225	17/8765	0.02	biological_process

Table 4c. Gene Ontology (GO) terms enrichment analysis of T10 unique genes compared with *T. rubrum* Tli CBS 139224 .

ID	Description	Ratio in study	Ratio In pop	P value (p_fdr)	GO Type
GO:0033551	monopolin complex	8/909	12/7153	0.034	cellular_component
GO:2000369	regulation of clathrin-dependent endocytosis	8/909	12/7153	0.034	biological_process
GO:2000370	positive regulation of clathrin-dependent endocytosis	8/909	12/7153	0.034	biological_process
GO:2001159	regulation of protein localization by the Cvt pathway	8/909	12/7153	0.034	biological_process
GO:0006720	isoprenoid metabolic process	15/909	38/7153	0.05	biological_process

ID: The coding of the gene in the GO database; Description: The functional description of the genes;

Ratio in study: Number of the unique genes in this GO/ total number of the unique genes of strains; Ratio in pop: Total number of this GO/Total number of genes;

P value: P-value (p_fdr) of the significant level in the enrichment analysis;

GO Type: The type of the GO function.

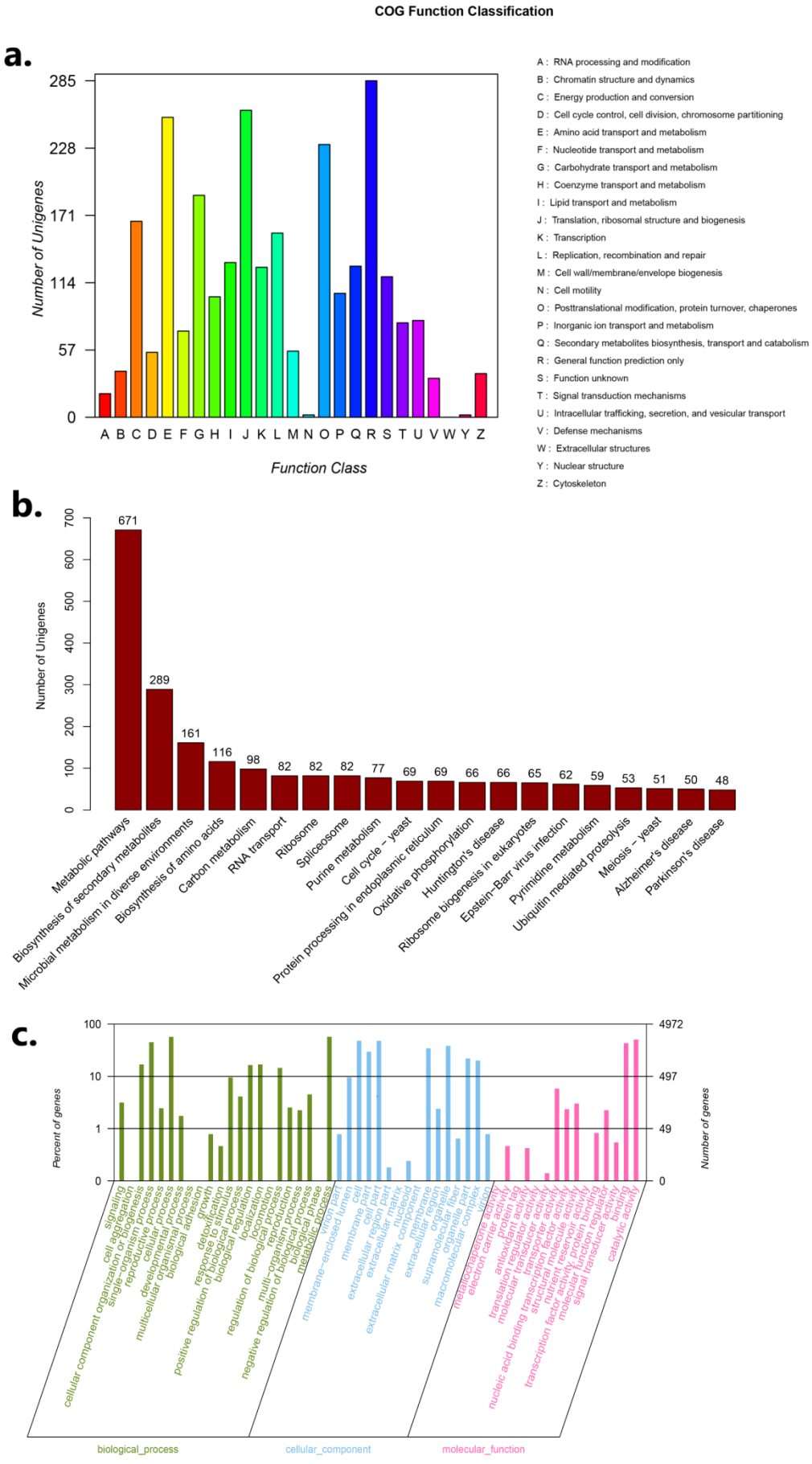
Figures and legends

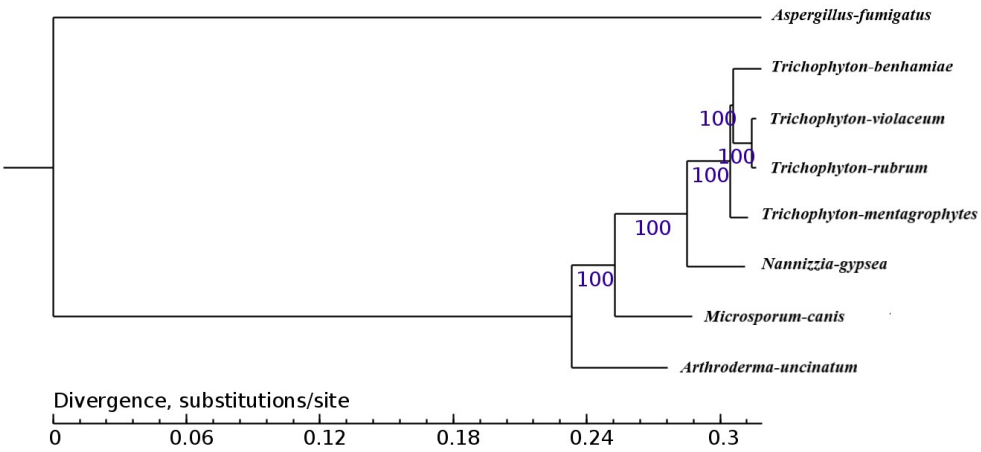
Figure1 Function annotation of *A. uncinatum* T10 strain. 1a. Cluster of orthologous groups (COG) classification of putative proteins. Y-axis indicates number of genes in a category. X-axis indicates 25 functional COG categories.1b. KEGG Classification of the genes. 7153 genes were assigned to 311 KEGG pathways. The top 20 most abundant KEGG pathways are shown. 1c. Functional categories of the annotated genes, broadly separated into 'biological process', 'cellular component' and 'molecular function' based on Gene Ontology (http://geneontology.org/). Left y-axis indicates percentage of a specific category of GO annotated genes in each main category. Right y-axis indicates number of genes in a category.

Figure 2. Maximum likelihood phylogenetic tree based on 4030 single-copy homologous genes among common seven dermatophyte strains showing the position of *A. uncinatum* T10 in the dermatophyte. *Aspergillus fumigatus* Z5 here is as an outgroup. The scale bar corresponds to 0.06 nucleotide substitutions per 5 sites.

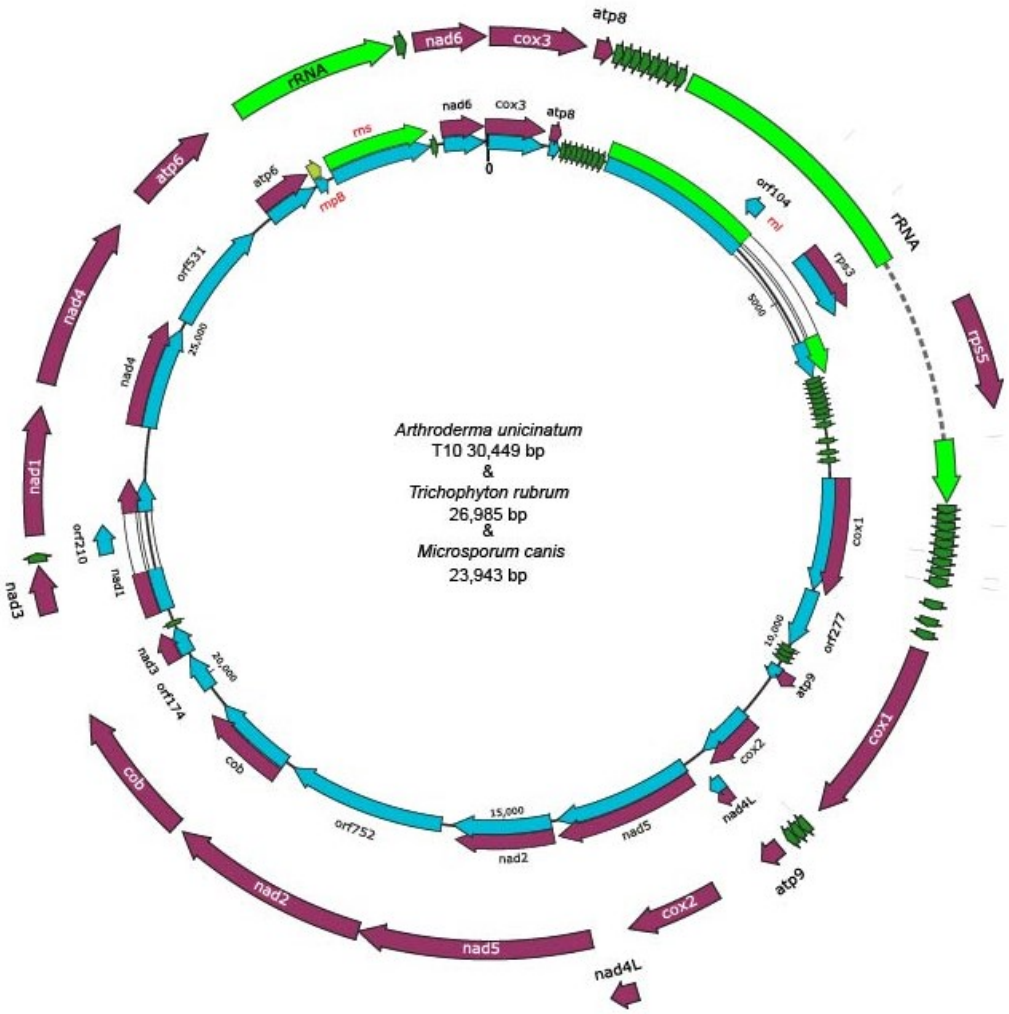
Figure 3. Mitochondrial genomes of *M.canis, A. uncinatum* T10 strain and *T. rubrum.* Outer circle is the mitochondrial genome of *M.canis,* middle circle is of T10 strain, inner circle is of *T. rubrum*. Dark green arrows： tRNA coding genes; Light green arrow: rRNA coding sequence. Blue and purple arrows: genes. ORFs are shown with blue arrows without corresponding purple arrows.

Figure 4. The linear genome comparisons of two *A. uncinatum* mitochondrial genome. The abscissa is the location of the genomes. The upper orange rectangle was the mitochondrial genome of NC-012828 reported before[2]. The bottom one shows the mitochondrial genome of T10 strain. The Pink parallelograms indicate the locations of homologous genomic regions between them.

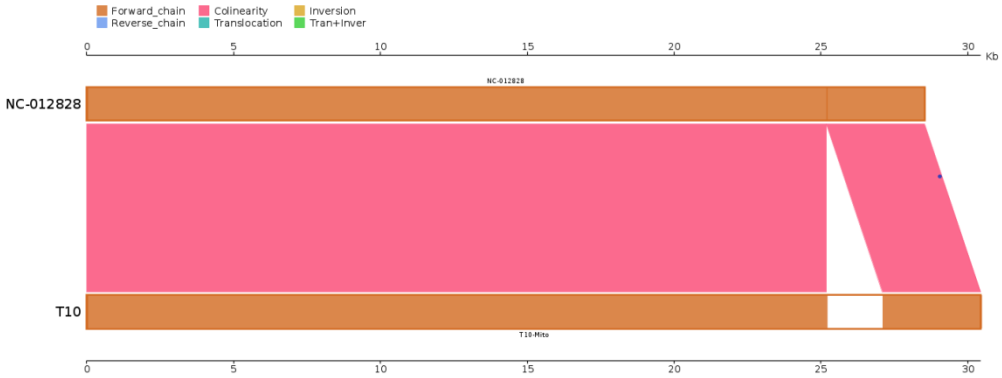




myc_13079_f2.jpg



myc_13079_f3.jpg



myc_13079_f4.tif