Mycosphere 7(8) 1124-1146(2016) www.mycosphere.org ISSN 2077 7019

Article

Doi 10.5943/mycosphere/si/2c/5

Whole genome re-sequencing and transcriptome analysis of the Stylosanthes Anthracnose pathogen *Colletotrichum gloeosporioides* reveal its characteristics

Huang HP [1, 2], Ma S[3], Huang JH[2], Zheng JL [2], Yi KX[2,*]

[1] *College of Agronomy, Hainan University, Haikou, China.*
[2] *Environment and Plant Protection Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China.*
[3] *Institute of Tropical Bioscience and Biotechnology, Chinese Academy of Tropical Agricultural Sciences, Haikou, China.*

Huang HP, Ma S, Huang JH, Zheng JL, Yi KX 2016 – Whole genome re-sequencing and transcriptome analysis of the Stylosanthes Anthracnose pathogen *Colletotrichum gloeosporioides* reveal its characteristics. Mycosphere 7(8) 1124–1146, Doi 10.5943/mycosphere/si/2c/5

Abstract

Stylosanthes anthracnose is an important disease caused by *Colletotrichum gloeosporioides*, which affects the production and utilization of *Stylosanthes* species in the world. Despite their importance, little progress has been made in understanding Stylosanthes anthracnose pathogens due to their insufficient genomic and transcriptomic data. In the current study, whole genome re-sequencing of seven *C. gloeosporioides* strains was carried out together with next generation RNA sequencing of *C. gloeosporioides* strain CH008. Whole genome re-sequencing was conducted with reference to the genome of highly virulent *C. gloeosporioides* strain CH008. It resulted in 10.06 G of effective data, with a mean proportion of 67.40% genes mapped to the reference strain and an average genome coverage of 79.71%. A total of 2,149,030 single nucleotide polymorphisms, 178,620 insertions and deletions, and 654 structural variations were discovered on the genomes of the seven tested strains. Transcriptomic analysis produced approximately 30.5 million high quality reads assembled into 33,529 contigs and further integrated into 25,376 unigenes. All assembled sequences were annotated against public databases. A total of 12,398 unigenes were assigned to 61 sub-categories of Gene Ontology terms. Among them, 3,717 unigenes were identified and mapped onto 264 pathways against Kyoto Encyclopaedia of Genes and Genomes pathway database. This study provides the first comprehensive transcriptomic resource available for *C. gloeosporioides.* The genome assembly and the unigenes identified in the study will benefit further studies on functional genes involved in pathogenicity. Simple sequence repeat markers developed in this study will facilitate studies on marker-assisted genomic and pathogenic diversities of *Colletotrichum gloeosporioides*.

Key words – *de novo* transcriptome assembly – gene annotation – pathogenicity – pathway analysis– RNA sequencing – SSR

Submitted 1 October 2016, Accepted 27 November 2016, Published online 3 December 2016

Corresponding Author: Yi KX – e-mail – Yikex@126.com

Introduction

Stylosanthes guianensis (Aublet) Sw., which is native to Africa, Southern and Central America, is an important forage legume with the largest planting area and widest application in tropical and subtropical regions of the world (Michalk et al. 1993, Kelemu et al. 1996, Munaut et al. 2001, Jiang et al. 2005). It is commonly used as ruminant feed, as green manure and as cover crops for soil protection and water conservation in plantations and orchards (Yi 2001). Cultivation and production of this forage legume has been threatened by Stylosanthes anthracnose (Kelemu et al. 1996, Munaut et al. 2001, Chen et al. 2014), a disease caused mainly by the pathogen *Colletotrichum gloeosporioides* (Penz.) Penz. & Sacc. and occasionally by *Colletotrichum truncatum* (Schwein.) Andrus & W.D. Moore and *Colletotrichum lindemuthianum* (Sacc. & Magnus) Briosi & Cavara (Feng et al. 1994). The pathogen forms appressoria on the leaf surface, penetrates into the cuticle, colonizes and initiates infections resulting in host cell necrosis leading to blight-like symptoms. It also induces chlorosis, abscission of herbage leaves, wilts of stems and petioles, abortion of flowers, and in severe cases, failure in seed production and finally whole plant death (Manners et al. 1992, 2000, Stephenson et al. 2000, Chakraborty et al. 2002). *Colletotrichum gloeosporioides* biotypes with high genetic variability and extensive pathogenic variations have been identified from Australia, Brazil, China and Colombia (Kelemu et al. 1999, Munaut et al. 2001, 2002, Yi 2001, Chakraborty et al. 2002, Jiang et al. 2005) Despite their importance, only few molecular resources such as a transfer DNA insert mutant library (Chen et al. 2014) and few identified pathogenicity genes (Stephenson et al. 2000), are available for this pathogen.

Next-generation sequencing technologies provide cost effective, rapid, powerful tools for genomic and transcriptomic studies to identify and annotate genes in non model organisms (Maher et al. 2009, Zhou et al. 2015, Verbruggen et al. 2015). Whole-genome re-sequencing can be used to study genetic variation between individuals (Li et al. 2009), as well as to illustrate the pathogenic mechanisms of biological processes at the molecular level. Whole genome analyses conducted on plant pathogens, such as *Botrytis cinera*, *Fusarium graminearum*, *Magnaporthe oryzae*, *Metarhizium anisopliae*, *Mycosphaerella graminicola*, *Neurospora crassa*, *Sclerotinia sclerotiorum*, *Ustilaginoidea virens*, *Valsa mali* and *Valsa pyri*, facilitate in establishing the foundation for functional genomics studies (Galagan et al. 2003, Dean et al. 2005, Amselem et al. 2011, Goodwin et al. 2011, Ohm et al. 2012, Xue et al. 2012, Zhao et al. 2013, Yin et al. 2015). Transcriptomic analyses of these pathogenic fungi have led to the development of molecular markers, identification and characterization of innate immune system pathways and the discovery of many interesting genes (Wicht et al. 2014, Chen et al. 2015, Verbruggen et al. 2015). These genomic and transcriptomic analyses have been used to understand the host-pathogen interactions by providing insights into the mechanisms underlying disease development, innate defense and gene to gene resistance (Zhang et al. 2014, Yap et al. 2015, Xing et al. 2016). Since *Colletotrichum* species represent a group of highly devastating pathogens of many crops, species such as *C. orbiculare* and *C. gloeosporioides* have been subjected to comparative genomic and transcriptomic studies revealing their distinct stage-specific gene expressions including potential pathogenicity genes (Robinson et al. 1998, Kim et al. 2000, Abang et al. 2009, O'Connell et al. 2012, Yoshino et al. 2012). In the current study, whole genome re-sequencing and transcriptomic analyses of *C. gloeosporioides* were conducted using an Illumina high-throughput sequencing system. We have designed 1,505 simple sequence repeat (SSR) markers, identified 3,242 simple sequence repeat (SSR) markers, and assembled 30.5 million raw high quality sequencing reads into 25,376 unigenes. Results obtained from this study are vital for obtaining a comprehensive understanding on the variations of highly pathogenic *C. gloeosporioides* at the genomic level and also provides foundation for molecular and genetic research on characterization of different biotypes of *C. gloeosporioides* and identification of key genes involved in their pathogenic processes.

Materials & Methods

Microorganisms & culture conditions

All the *C. gloeosporioides* strains used in this study were obtained from the culture collection of the Environment and Plant Protection Institute, Chinese Academy of Tropical Agricultural Sciences, Haikou, China. For the whole genome re-sequencing, seven *C. gloeosporioides* strains were used and the *C. gloeosporioides* strain CH008 was used as the reference strain (Table 1). For the transcriptomic analysis, biotype B of the highly pathogenic *C. gloeosporioides* strain CH008 was used. Single spore cultures of all pathogenic strains were maintained on potato dextrose agar (PDA) for 3 to 4 days.

Table 1 Data on *Colletotrichum gloeosporioides* strains used in the current study

Strain ID	Collection date	Collection site	Host	Host strain/species
CH008	1997.12.09	Dongfang City, Hainan Province	*S. guianensis*	CIAT184
CH010	1997.12.09	Dongfang City, Hainan Province	*S. guianensis*	722
CH013	1997.12.09	Dongfang City, Hainan Province	*S. guianensis*	722
CH020	1997.05.01	Dianbai County, Guangdong Province	*S. guianensis*	CIAT184
CH036	1997.12.03	Changjiang Li Autonomous County, Hainan province	*S. scabra*	Seca
CH100	1999.12.12	Dongfang City, Hainan Province	*S. guianensis*	CIAT184
CH247	2001.12.01	Danzhou City, Hainan Province	*S. scabra*	RRR94-96
CH450	2001.11.23	Qianjiang District, Guangxi Province	*S. guianensis*	CIAT184

Pathogenicity assessment of the strains used for the whole genome re-sequencing

Six weeks after sowing and when the plants reached the height of 40–50 cm, *Colletotrichum gloeosporioides* pathogens were inoculated onto *Stylosanthes* plants. *Stylosanthes* plants with similar leaf numbers and stem diameters were selected and tagged. Fungal spores were collected by scraping the mycelium and conidiomata on the medium, and then, they were suspended in sterilized water to make a spore suspension with a concentration of 10^6 spores/ml. Spore suspensions were applied immediately as a spray onto the plants. For each strain, three inoculation repetitions were conducted with six plants in each repetition. After inoculation, the plates with raised seedling were moved into a basin containing water and covered with an organic transparent plastic cover to ensure saturated humidity and constant temperature for 48 h. Then, the plates were removed from the water basin and seedlings were cultivated. Fifteen days after inoculation, the degree of the disease was graded from 0 to 7, Disease severity was expressed as disease index (DI) according Powell (Powell et al. 1971).

$$\mathrm{DI} = \sum \frac{(\text{Number of stylo plants at each grade } \times \text{ corresponding grade value})}{(\text{Total stylo plant number } \times \ 7)} \times 100$$

Genomic DNA extraction, library construction & high throughput sequencing

For the genomic DNA extraction, fungi grown on PDA were transferred to 150 ml of potato dextrose liquid medium and incubated in a rotary shaker at 140 rpm for 3 days at 25–28 °C. Then, mycelia were collected by vacuum filtration through sterilized gauze, followed by washing with deionized water. Filtered mycelia were vacuum freeze dried, and stored at -20 °C. Genomic DNA was extracted using Fungal Genome DNA Extraction Kit D2300 (Solarbio, China) by following manufacturer's instructions. DNA quality was determined using a combination of Nano Drop ND1000 (Thermo Fisher Scientific Inc., USA) and an Agilent Bioanalyzer 2100 (Agilent Technologies, USA). These total genomic DNA samples were stored at -20 °C until further processing.

High quality genomic DNA was fragmented into short segments by sonication. DNA segments were purified, end repaired, joined with A- at the 3' end, and with an adaptor. The processed segments of appropriate length were screened by electrophoresis, and amplified using PCR to form a sequencing library. The concentration of the DNA library was determined by Qubit 2.0 Fluorometer (Invitrogen Inc., USA), and the size of the DNA library was determined by Agilent 2100 Bioanalyzer (Agilent Technologies, USA). The whole genome re-sequencing was performed using an Illumina HiSeq 2500 system (Biomarker Technologies Co. Ltd., China).

Read mapping, SNP detection & annotation

Paired-end sequencing libraries were constructed for *Colletotrichum gloeosporioides*. The quality-score distribution of each library was checked. If a sequence base had a quality score less than Q20, which indicates an accuracy of 99% for the base call, the base call was changed to "N". Reads with fewer than 90 bp or with Ns at more than 10% of their total base positions were removed. Remaining reads, with an average length of 126 bp, were aligned to the reference genome of *Colletotrichum gloeosporioides* strain CH008 using the Burrows-Wheeler Aligner (BWA0.7.7, UK) with default parameters (Li & Durbin 2010).

Alignments were further processed by Sequence alignment/map (SAM) tool (version 0.1.19) (Li et al. 2009) used for converting SAM files into binary SAM (BAM) file formats, for sorting and for indexing purposes. Local re-alignment and re-calibrations were performed using the Genome Analysis Toolkit (GATK 3.1) frame work (Mckenna et al. 2010). Initial SNP discovery was performed using multi-sample SNP-calling procedure in the GATK package. To reduce the false discovery rate, filtering was conducted using the following criteria: Phred scaled polymorphism probability (QUAL) <30.0, variant confidence normalized by depth (QD) < 2.0, mapping quality (MQ) < 40.0, strand bias (FS) > 60.0, HaplotypeScore> 13.0, MQRankSum< −12.5, and ReadPos Rank-Sum < −8.0. Detailed description of the terms can be found at the GATK website (https://www.broadinstitute.org/gatk/guide/). All filtered SNPs were assigned using snpEff version 4.0 (Cingolani et al. 2012).

Phylogenetic analysis of the re-sequenced genomes

Maximum likelihood analysis was conducted based on the SNPs present in the intergenic regions that presumably were not subjected to selective pressures. A phylogenetic tree was constructed for the eight *Colletotrichum gloeosporioides* genomes, including seven re-sequenced genomes and the reference genome, with default parameters of RAxML (Stamatakis 2014). Repeated sequences and genomic regions were excluded to clarify the relationship between strains.

Total RNA extraction & cDNA synthesis for transcriptomic analysis

Fungal mycelia were transferred and incubated in potato dextrose liquid medium using the same culture conditions as described previously in the section 2.3. The collected mycelia were harvested by filtration through Mira-cloth (Calbiochem, USA), immediately frozen and stored in liquid nitrogen at -80 °C until further processing. Total RNA was extracted from stored mycelia using an AxyPrep total RNA extraction kit (Axygen, USA) and then treated with DNase I according to the manufacturer's instructions. Quality of extracted total RNA was verified using a NanoDrop ND1000 (Thermo Fisher Scientific Inc., USA) and Agilent Bioanalyzer 2100 (Agilent Technologies, USA).

Construction of cDNA library, Illumina sequencing and de novo transcriptome assembly

Attached magnetic beads were utilized to purify mRNA from the total RNA. After purification, mRNA was fragmented into small pieces using heat treatment in the presence of Mg^{2+} ions, and the cleaved RNA fragments were used as templates to synthesize first-strand cDNA using reverse transcriptase and random hexamer primers. Followed by second-strand cDNA synthesis using DNA polymerase I (Thermo-Scientific, USA) and RNase H (New England Bio labs, USA), they were subjected to paired-end adapter ligation. The products were then amplified to generate a cDNA library and sequenced on an Illumina HiSeq 2500 platform (Illumina Inc., USA) to generate 2 × 100 bp paired-end reads.

Prior to transcriptome assembly, raw data were scanned using Cassava software (version 1.8.1) and low-quality reads along with adapter-sequences were removed. The resulted high-quality reads were *de novo* assembled using program “Trinity” with the default settings (Grabherr et al. 2011). The clean reads were deposited in the NCBI Sequence Read Archive (SRA) database under the accession number SRP065381.

Functional annotation, classification and pathway analysis

Following *de novo* assembly, functional annotations of all assembled unigenes were searched against NCBI non-redundant sequence (Nr) database, NCBI non-redundant nucleotide database (Nt) and manually annotated protein sequence database Swiss-Prot (http://www.expasy.ch/sprot) under the threshold parameter of E-value cut-off$\leq 1e^{-5}$. Search against protein families database Pfam (version 27.0) (Finn et al. 2014) was performed using HMMER3 package software with an E-value ≤0.01.Unigenes were also aligned to EuKaryotic Orthologous Groups (KOG) protein databases (http://www.ncbi.nlm.nih.gov/COG/) (Tatusov et al. 2003) using BLASTx with an E-value $\leq 1e^{-3}$. Gene Ontology (GO) annotation was obtained by assigning molecular function, biological process and cellular component terms using Blast2GO program as described in (Conesa et al. 2005). Kyoto Encyclopaedia of Genes and Genomes (KEGG) annotation was performed to assign molecular interaction networks and metabolic pathways using the online KEGG Automatic Annotation Server (KAAS) (http://www.genome.jp/kegg/kaas/) (Moriya et al. 2007, Kanehisa et al. 2008).

Open reading frame (ORF) identification & development of SSR markers

The ORFs of the unigenes were searched using EMBOSS ‘getorf’ program with minimum nucleotide size of 100bp (Rice et al. 2000). The MISA tool (http://gramene.agrinome.org/db/searches/ssrtool) was used to identify dinucleotide to hexanucleotide SSRs using the default settings of the SSRIT tool.

Results

We first determined the pathogenic effects of eight different *C. gloeosporioides* isolates on *Stylosanthes* sp. Five isolates, including strains CH008, CH010, CH013, CH020 and CH247, showed high pathogenicity (DI > 45), while strains CH036, CH100 and CH450 showed low pathogenicity (DI < 25), as seen in Table 2.

Raw sequence reads were deposited in the NCBI sequence reads archive under accession no. SRP065381. The quality assessment statistics of the genome re-sequencing results of the seven *C. gloeosporioides* strains with different pathogenicity are listed in Table 3. Clean reads with a data size of 10.06 Gb were generated. The Q20 percentage (proportion of nucleotides with a quality value larger than 20 in the reads) was higher than 90% and the Q30 percentage was higher than 85%. The mean GC percentage of the clean reads was 51.74% (Table 3).

Table 2 Results of pathogenicity tests of *C. gloeosporioides* strains

NO.	Strain ID	Mean disease grade	DI
1	CH008	3.00	45.24
2	CH010	3.50	48.41
3	CH013	3.33	48.41
4	CH020	3.00	44.44
5	CH247	3.50	47.62
6	CH036	1.50	17.46
7	CH100	1.67	22.22
8	CH450	1.00	15.08

Table 3 Summary of Illumina genome re-sequencing results of seven *C. gloeosporioides* strains

BMK_ID	Raw Reads	Clean Reads	Clean Base	Q20 (%)	Q30 (%)	GC (%)
CH010	5,772,084	5,487,814	1,382,848,424	91.38	85.14	51.16
CH036	6,082,627	5,748,052	1,448,425,036	91.38	85.04	52.49
CH100	6,166,491	5,940,881	1,497,011,207	91.46	85.03	51.96
CH450	6,083,559	5,885,119	1,482,963,529	93.39	85.11	52.19
CH013	7,405,721	7,172,120	1,680,905,639	92.83	86.84	51.78
CH020	6,265,325	5,775,326	1,111,605,800	98.64	94.97	51.16
CH247	8,086,292	7,772,551	1,482,772,200	98.67	95.04	51.42

Whole genomes of the Stylosanthes anthracnose pathogens obtained in the present study were aligned with that of the highly pathogenic strain CH008 (data not published). The efficiency of the alignment is shown in Table 3. From 10.9–14.8 million clean reads generated from each genome, 61–90% were mapped to an unique position against the reference genome using BWA, with an average depth of 20× and average genome coverage of 79.71% (Table 4). Strains CH036, CH100 and CH450, which showed low pathogenicity (Table 2), had the lowest total mapped reads and genome coverage percentages (Table 4). These results indicate that the re-sequencing data were applicable for genetic difference analyses.

All uniquely mapped regions were subjected to specific single nucleotide polymorphisms (SNPs) calling. Comparisons between each of these *C. gloeosporioides* re-sequenced genomes with the *C. gloeosporioides* strain CH008 reference genome identified a total of 2,149,030 high quality SNPs with a transition/transversion ratio (Ti/Tv) between 0.51 and 2.19 (Table 5). Strains CH036, CH100 and CH450 have the highest SNP numbers and Ti/Tv ratios, consistent with their low pathogenicity compared to reference strain CH008. These results indicate abundant variations in SNPs between different strains of *C. gloeosporioides*.

Indels reflect variations between the samples and the reference genome. In addition, indels in the coding region lead to frame shift mutations and thus functional changes in genes. However, our results showed that there were fewer indels compared to SNPs. We called a total of 178,620 and 18,169 indels in the whole genomes and in the coding regions of the seven samples, respectively. Strains with the lowest pathogenicity (CH036, CH100 and CH450) showed the highest number of small indel variations, suggesting that the number of small indel variations might be negatively related with the pathogenicity of *C. gloeosporioides*. Therefore, small indel variations of *C. gloeosporioides* may play important roles in plant pathogenicity and their presence could potentially be used as markers for pathogenicity phenotypes. The small indel results are shown in Table 6.

Table 4 Illumina sequencing and Burrow-Wheeler Aligner mapping statistics

Samples	Clean reads number	Total mapped reads after filtering (%)	Genome coverage (%)	Genome coverage depth	Number of genes
CH010	10,975,628	67.50	80.64	19	21013
CH036	11,496,104	61.42	75.40	20	23838
CH100	11,881,762	61.45	75.38	20	23837
CH450	11,770,238	61.69	75.19	20	23789
CH013	14,344,240	90.60	89.63	28	24984
CH020	11,116,058	61.55	80.20	14	21078
CH247	14,827,722	68.85	80.33	21	21941

Table 5 Summary of SNPs identified from the genomes of seven *C. gloeosporioides* strains

Sample	SNP number	Transition	Transversion	Ti/Tv
CH010	712,605	486,665	225,940	2.15
CH036	1,072,656	736,729	335,927	2.19
CH100	1,076,066	738,982	337,084	2.19
CH450	1,064,121	730,868	333,253	2.19
CH013	14,969	5,115	9,854	0.51
CH020	707,918	483,613	224,305	2.15
CH247	773,246	526,893	246,353	2.13
Total	2,149,030	1,499,742	649,288	2.3

Table 6 Statistics of the structural variations identified in the genomes of seven *C. gloeosporioides* strains

Sample	SV	INS	DEL	INV	ITX	CTX	UN
CH010	246	69	129	3	0	44	1
CH036	183	23	138	7	1	13	1
CH100	230	76	134	6	0	13	1
CH450	178	40	119	5	0	12	2
CH013	1096	867	83	2	2	141	1
CH020	609	469	79	5	1	55	0
CH247	2443	2170	132	6	11	122	2

SV: Total number of structural variations; INS: Number of insertions; DEL: Number of deletions; INV: Number of inversions; ITX: number of intra-chromosomal translocations; CTX: Number of inter-chromosomal translocations; UN: Unknown structural variations.

Structural variations (SVs) indicate insertions, deletions, inversions and translocations of large fragments at genome level. In the current study, six kinds of SVs were identified using Break Dancer software (Chen et al. 2009), including number of insertions (INS), deletions (DEL), inversions (INV), intra-chromosomal translocations (ITX), inter-chromosomal translocations (CTX) and unknown SVs (UN). Among these variations, INS, DEL and CTX were the most abundant SVs in the seven re-sequenced *C. gloeosporioides* genomes. Strains with the lowest pathogenicity (CH036, CH100 and CH450) showed the lowest number of SVs, suggesting that the number of SVs might be positively related with the pathogenicity of *C. gloeosporioides*. The detailed SV results are shown in Table 6.

Average genome coverage was approximately 80%. The concentric circles shown in Fig. 1 are the different features drawn using the Circos program (Krzywinski et al. 2009), with 15 chromosomes portrayed along the perimeter of each circle. We found that strains with the lowest pathogenicity (CH036, CH100 and CH450) had similar SNP densities, indel densities and SV (INS, DEL and INV) distributions in each chromosome. All these variations in SNPs, SVs and in genes will result in virulence variation and then affect pathogenicity with functional mutants.

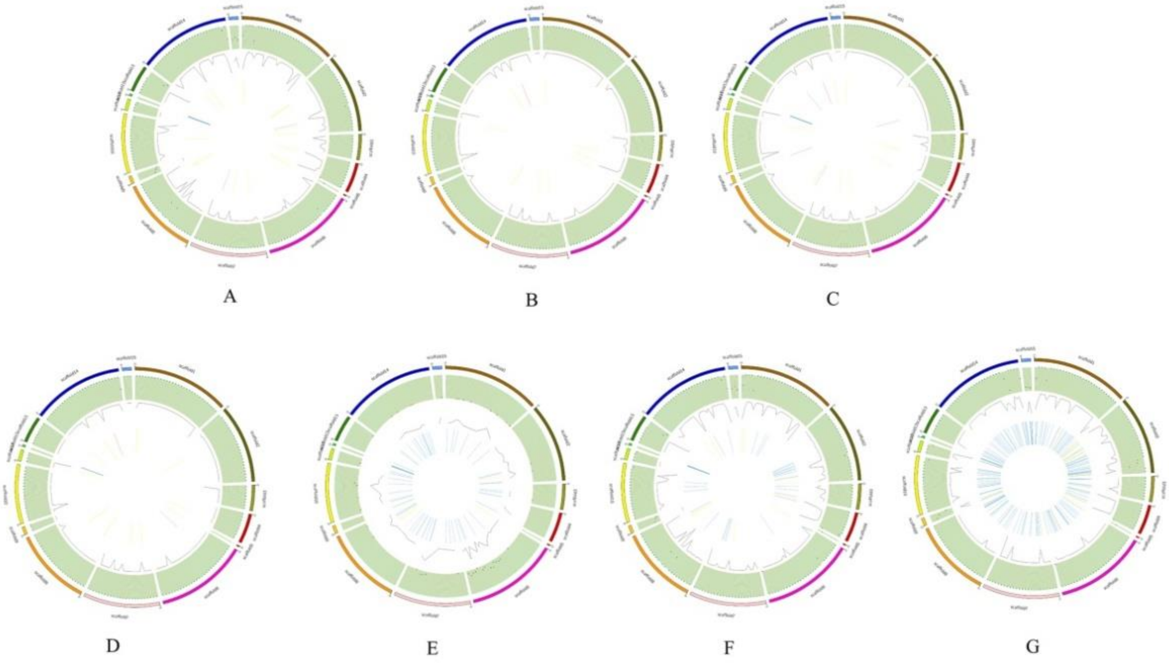

Fig. 1 – Summary of re-sequencing data of seven *C. gloeosporioides* strains. A: CH010; B: CH036; C: CH100; D: CH450; E: CH013; F: CH020; G: CH247; Patterns from outside to inside represent: chromosomes, SNP density, indel density and SV (INS, DEL and INV) distribution in each chromosome (Unit: Mega).

Phylogenies of the reference genome and the seven re-sequenced genomes were inferred using the maximum likelihood approach based on the SNPs. The results are shown in Fig 2. Highly pathogenic strains CH008, CH010, CH013, CH020 and CH247 were clustered in a single clade, while the lowest pathogenic strains CH036, CH100 and CH450 were clustered into another clade. Strain CH013 was the closest to the reference strain CH008.

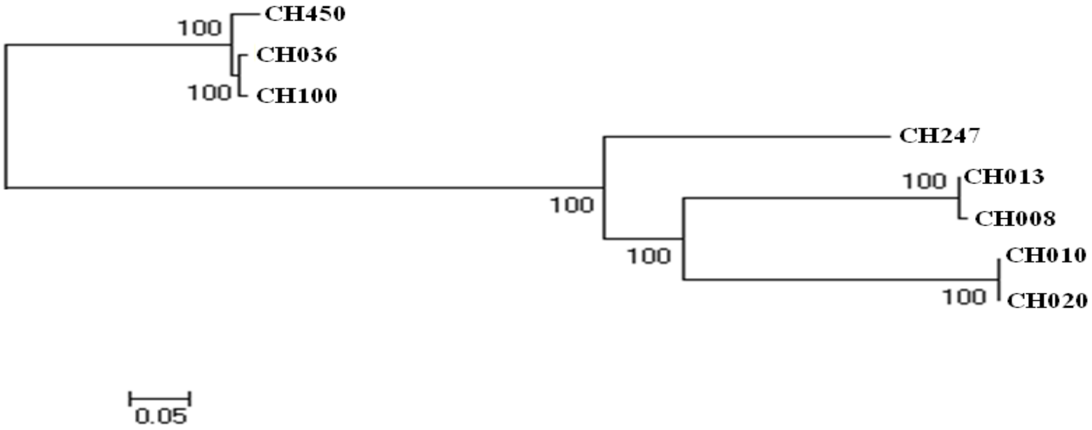

Figure 2 – Unrooted maximum likelihood phylogenomic tree of *C. gloeosporioides* strains used in the study. This phylogenomic tree was constructed on the basis of SNPs using the Dayhoff amino acid substitution model, shows the evolutionary relationships between the indicated eight *C. gloeosporioides* strains. Bootstrap values are shown near to the relevant tree branches.

To obtain a comprehensive overview of the CH008 transcriptome, a mixed cDNA sample from mycelia was prepared and sequenced using the HiSeq 2500 platforms. After stringent quality filtering, 30,455,527 clean reads (~6.1 G) with high-quality were obtained. The Q20 percentage (nucleotides with a quality larger than 20) was higher than 95 % and GC percentage of the clean reads was about 53.6%. All obtained short sequences were assembled into 33,529 transcripts with 1,459 bp mean length and N_{50} of 2,606bp. A total of 25,376 unigenes were obtained, among which 9,410 (37.08%) genes were greater than 1 kb (Table 7). The obtained unigenes exhibited variable lengths, ranging from 201 bp to 13,832 bp.

Table 7 Summary for the transcriptome of *C. gloeosporioides* strain CH008

Length(bp)	Number of transcripts	Number of unigenes
200-500bp	11,853	11,330
500-1kbp	5,716	4,636
1k-2kbp	6,887	4,534
>2kbp	9,073	4,876
Total	33,529	25,376

All identified unigenes were aligned using BLASTx and BLASTn tools against the Nr, Nt, KEGG, Swiss-Prot, Pfam, GO and KOG databases. Among 25,376 unigenes, total of 16,684 (65.74%) unigenes were successfully annotated in at least in one database. Among them, 15,498 (61.07%) were matched in the Nr protein database; 4,653 (18.33%) were matched in the Nt database; 12,398 (48.85%) had significant matches in the GO database; and 8,331 (32.83%) had similarities in the Swiss-Prot database.

Among all the annotated unigenes, 15,498 (61.07%) had significant matches in the Nr database, with the remaining 38.93% demonstrating no significant hits. The E-value distribution of the annotated unigenes in the Nr database showed 94.9% of the mapped sequences having strong homologies (E-value <10^{-30}) and 87.4% with significantly strong homologies (E-value <10^{-45}) to the available sequences (Fig. 3A). Further analysis showed that 60.8% of sequences were 95–100% similar (Fig. 3B). In terms of species distribution, 79.3% of unigenes were mapped to *C. gloeosporioides* and all together 91.2% were mapped to the genus *Colletotrichum*. Only 5.1% of the distinct sequences had matches to sequences from 'other' species (Fig. 3C).

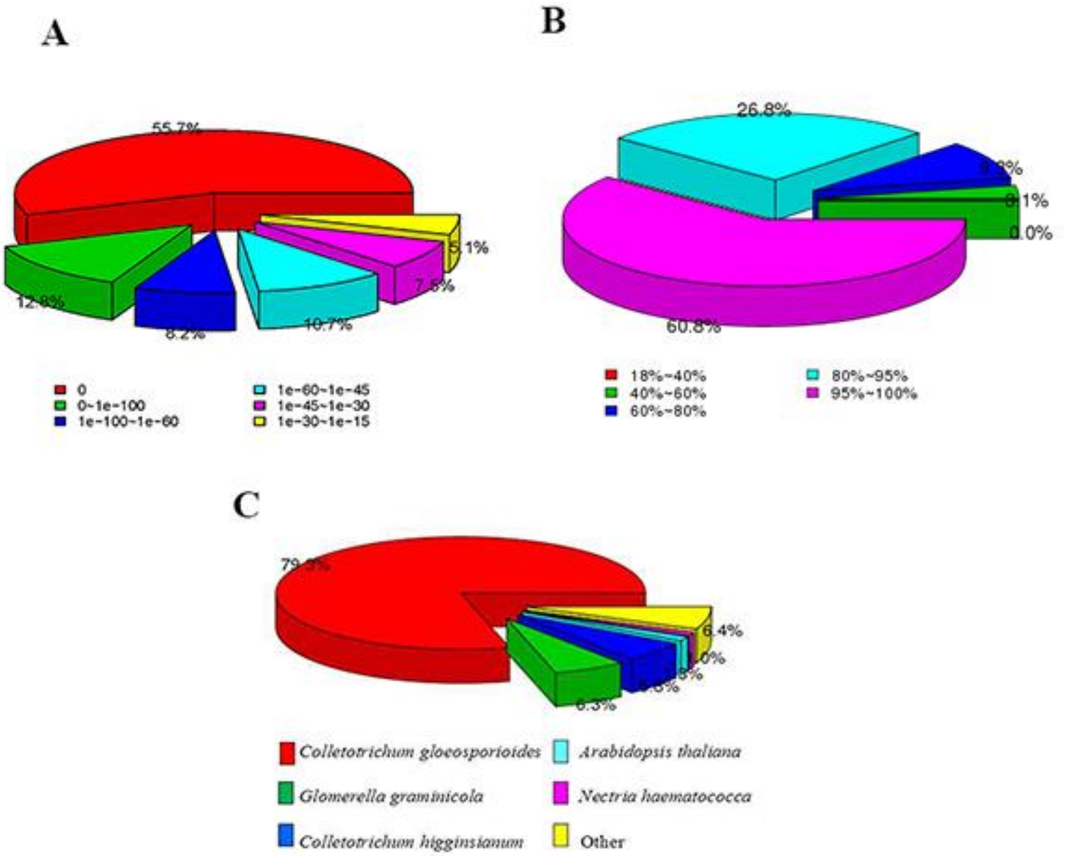

Figure 3 – Similarity analysis of unigenes for *Colletotrichum gloeosporioides* strain CH008. A. E-value distribution of BLAST hits for each unigene with a cutoff E-value of $1.0E^{-5}$, B. Similarity distribution of the top BLAST hits for each unigenes, C. Species distribution of the top BLAST hits for each unigenes in Nr database.

Gene ontology is an international, standardized functional classification system used for annotating and analyzing the functions of a large number of genes and their products. Among the 15,498 unigenes annotated in the Nr database, 12,398 unigenes were assigned with GO terms. They were divided into three main categories and 61 sub-categories (Fig. 4). These three main categories include biological process (25 sub-categories), cellular component (18 sub-categories) and molecular function (18 sub-categories). Among these three main categories, dominant sub-categories with highest number of genes include cellular process, metabolic process, cell, cell part, binding and catalytic activity.

The KOG database helps to classify orthologous gene products (Tatusov et al. 1997). Each KOG is a group of three or more proteins inferred to be orthologs, and the whole database is built on coding proteins with complete genomes as well as system evolutionary relationships of bacteria, algae and eukaryotes.

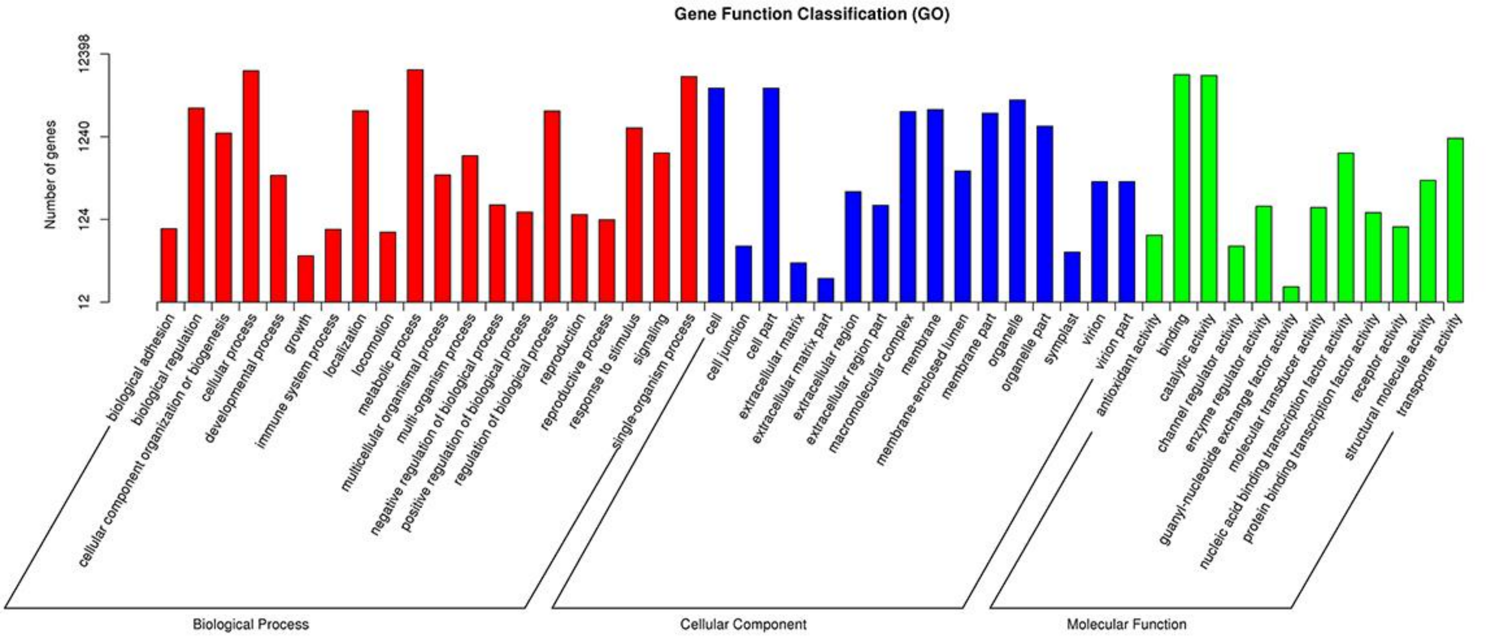

Figure 4 – Gene Ontology annotation and classification of assembled unigenes. A total of 12,398 unigenes with significant similarity in the Nr protein database were summarized into three main categories: biological processes, cellular components or molecular functions. Y-axis is number of genes in the category.

The current KOG database contains both prokaryotic and eukaryotic clusters (Rice et al. 2000). We aligned the unigenes to the KOG database to find homologous genes and classify possible functions (Fig. 5). A total of 5,258 unigenes (20.84%) had matches in the KOG database, with an E value $<1e^{-3}$. The possible functions of 5,258 unigenes were classified and subdivided into 25 KOG categories. The largest group was ‘General function prediction only’, followed by ‘Post-translational modification, protein turnover, chaperones’, ‘Secondary metabolites biosynthesis, transport and catabolism’, ‘Signal transduction mechanisms’, ‘Energy production and conversion’, ‘Amino acid transport and metabolism’ and ‘Translation, ribosomal structure and biogenesis’. The three smallest groups were ‘Cell motility’, ‘Extracellular structures’ and ‘Nuclear structure’.

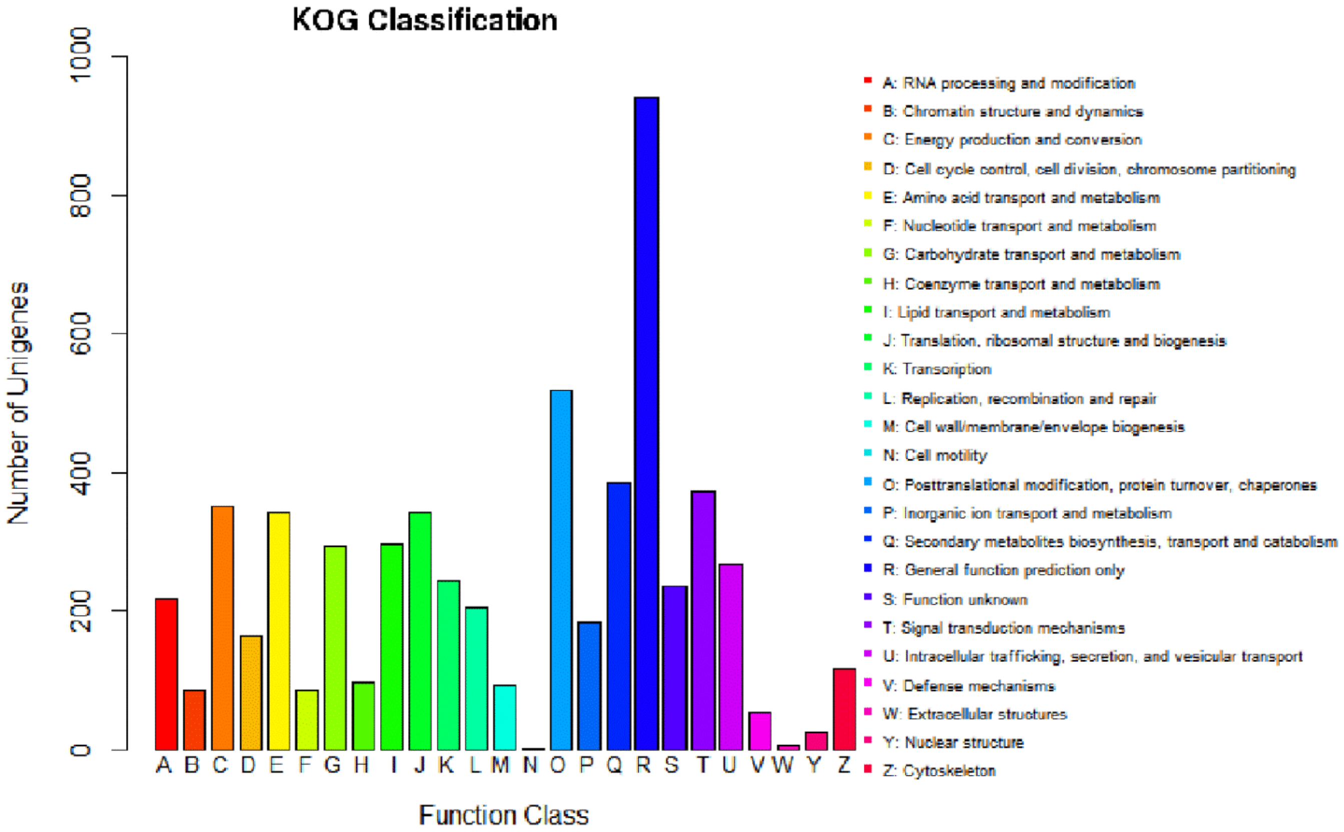

Figure 5 – Histogram showing the classification of Eukaryotic Orthologous Groups (KOG) of the annotated unigenes. All unigenes were aligned to the KOG database to predict and classify possible functions. Out of 15,498 Nr hits, 5,258 unigenes were grouped into 24 KOG classifications.

The KEGG pathway analysis is helpful to clarify unknown gene function and potential biological functions. Among the 25,376 of total unigenes, 3,717 (14.64%) were assigned to five main categories, including 264 predicted KEGG pathways. Among the five main categories, the largest group was genes responsible for different metabolisms (3,303, 53.89%), which included unigenes for carbohydrate metabolism (674), amino acid metabolism (601), overview (485), energy metabolism (342) and lipid metabolism (297), followed by organismal systems (899, 14.67%), genetic information processing (811, 13.23%), environmental information processing (572, 9.33%) and cellular processes (544, 8.88%, Fig. 6). The metabolic pathways predicted in this study would be useful for subsequent functional genomic research.

The EMBOSS software "getorf" function was used to identify the ORFs of the assembled sequences. Among 25,376 of assembled unigene sequences of *C. gloeosporioides*, 24,961 (98.36%) had an ORF longer than 100 bp, with an average length of 807 bp (min length = 102, max length = 13,100)

In order to develop new molecular markers for *C. gloeosporioides*, all 25,376 annotated unigenes were mined for potential microsatellites with a minimum of four repetitions. A total of 3,242 potential SSRs were identified by using MISA tool, 487 of which contained more than one SSR, and 149 SSRs were present in compound form (Table 8).

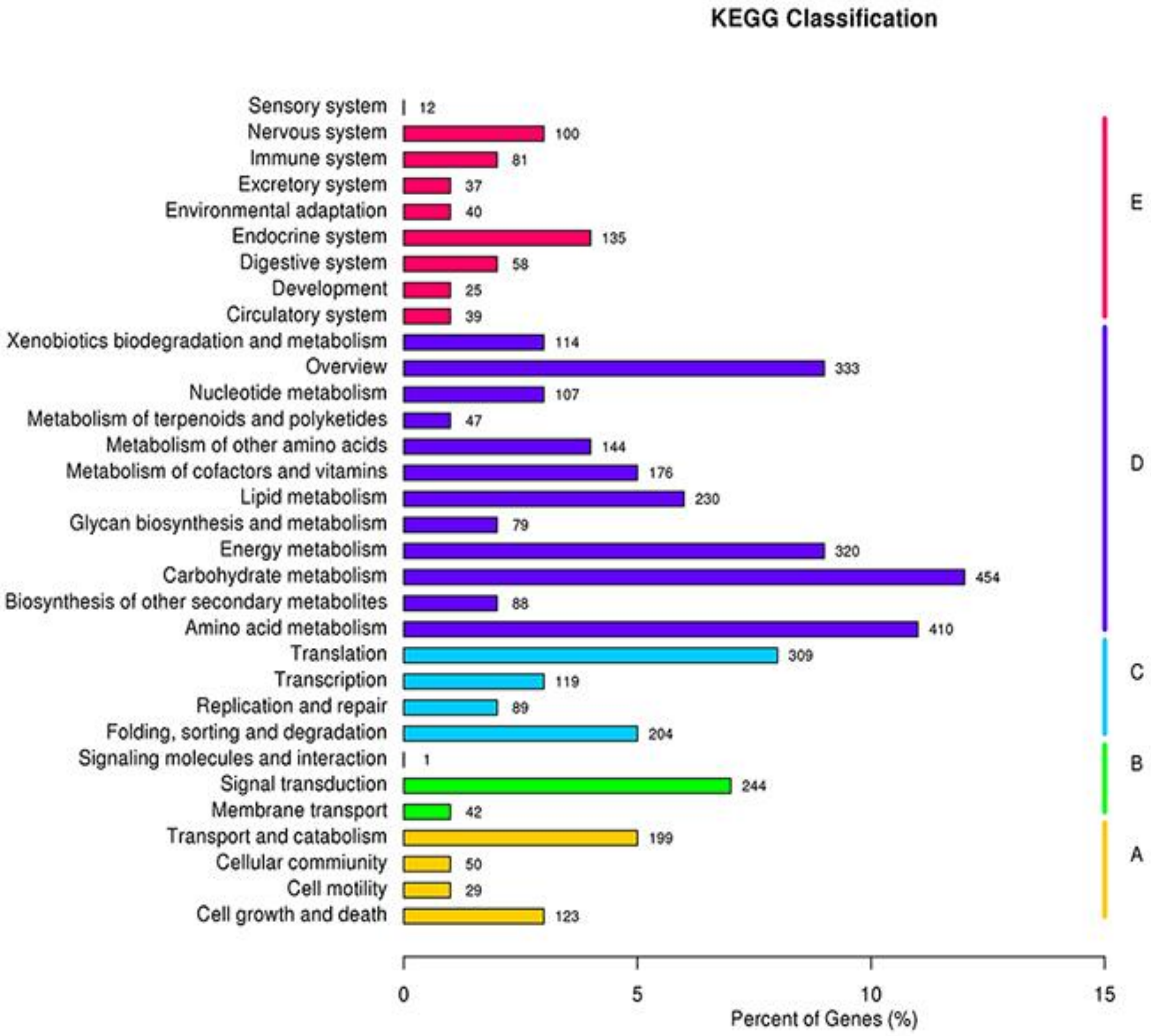

Figure 6 – Kyoto Encyclopedia of Genes and Genomes (KEGG) classification on the annotated unigenes. In total, 3,717 unigenes were grouped into 264 pathways. A. cellular processes, B. environmental information processing, C. genetic information processing, D. metabolism, E. organismal systems

Table 8 Summary of SSR markers developed for *C. gloeosporioides*

	Numbers
Total no. sequences examined	25,376
Total size of examined sequences (bp)	29,366,384
Total no. identified SSRs	3,242
No. SSR containing sequences	2,609
No. sequences containing more than one SSR	487
No. SSRs present in compound formation	149
Di-nucleotides	725
Tri-nucleotides	1,229
Tetra-nucleotides	78
Penta-nucleotides	23
Hexa-nucleotides	15

Considering the inaccuracy of SSRs with only one nucleotide motif, di to hexa-nucleotide SSRs were further used to perform type and distribution analyses. Among the remaining 2,070 SSRs, tri-nucleotide repeat motifs were the most abundant (1,229, 59.37%), followed by di- (725, 35.02%), tetra- (78, 3.77%), penta- (23, 1.11%) and hexa-nucleotide (15, 0.72%) repeat motifs (Table 8). We also found that the AC/GT di-nucleotide repeat was the most abundant motif detected in the newly developed SSRs (16.91%), followed by AG/CT (16.23%), AGC/CTG (15.89%), ACC/GGT (10.24%) and CCG/CGC (9.81%, Fig. 7).

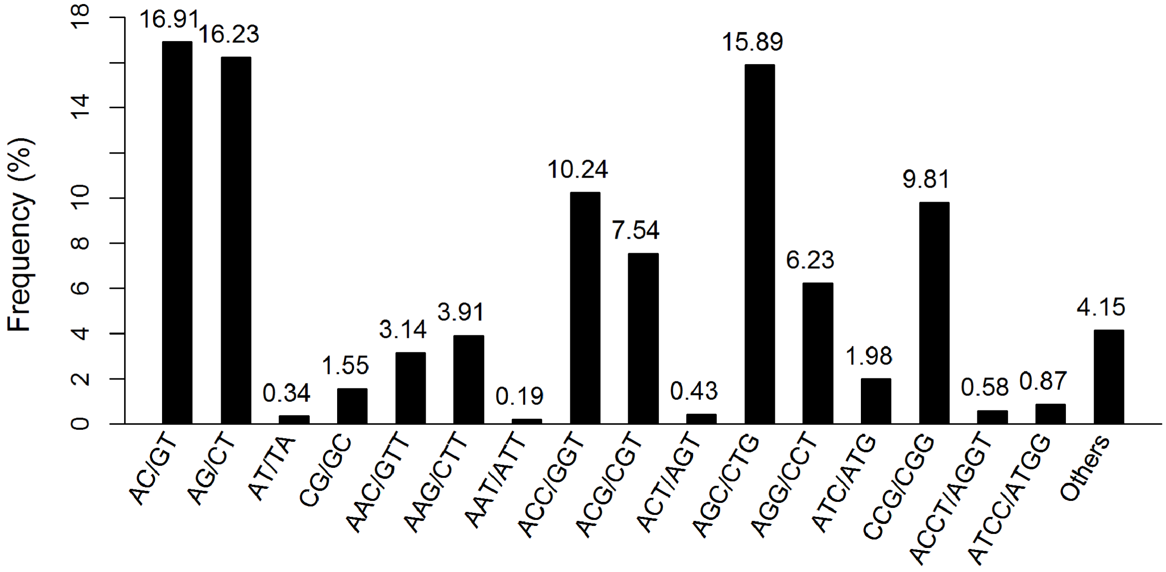

Figure 7 – Frequency distribution of SSRs based on motif sequence types of *C. gloeosporioides*

Based on the identified SSRs on the *C. gloeosporioides* genome, 1,505 SSR primers were designed. Randomly selected 20 SSR primers were used for validation of marker assay performance. Sixteen primer pairs resulted in successful PCR amplification. These results demonstrated that the potential identified SSRs would be a useful resource for the development of highly polymorphic SSR markers for *C. gloeosporioides*.

Discussion

Due to their low cost, high throughput, high accuracy and rapid methods, next generation sequencing technologies are widely applied for transcriptome sequencing and characterization (Bentley 2006, Maher et al. 2009). Among them, Illumina transcriptome sequencing and assembly have been successfully applied for a large variety of organisms including model (Hegedus et al. 2009, Li et al. 2010, Wang et al. 2010） and non-model organisms (Wu et al. 2010, Chen et al. 2013, Ge et al. 2014, Jiang et al. 2015, Verbruggen et al. 2015). However, due to the agronomical significance of these *Colletotrichum* species, they have been subjected to many studies involving fungal pathogenicity (Perfect et al. 1999, Münch et al. 2008). This highly diverse genus consists of different subgroups of species within a single species complex, a wide range of hosts and lifestyles varying from possible endophytes to destructive pathogens (Sharma et al. 2011). Therefore, studying these pathogens provide opportunities to analyse mechanisms underlying the diverse range of plant-pathogen interactions. Before the genomic and transcriptomic data of the highly destructive pathogen, *C. gloeosporioides* become available, studying the genomic and pathogenic diversity has been restricted to AFLP (Jiang et al. 2005), RFLP (Braithwaite et al. 1990, Chakraborty et al. 2002) and RAPD markers (Munaut et al.

2002, Weeds et al. 2003). However, with the availability of genomic and transcriptomic resources, many studies has been conducted to analyze genomic and transcriptomic data of *C. gloeosporioides* (Robinson et al. 1998, Stephenson et al. 2000, O'Connell et al. 2012, Gan et al. 2013) as well as other *Colletotrichum* species such as *C. higginsianum*, *C. orbiculare* (Narusaka et al. 2004, Kleemann et al. 2012, O'Connell et al. 2012, Yoshino et al. 2012, Gan et al. 2013).

In the current study, we carried out whole genome re-sequencing of stylo anthracnose pathogens using Illumina Hiseq 2500 high-throughput sequencing technology. We provide the first comprehensive re-sequencing data for high and low pathogenic *C. gloeosporioides* genomes. According to the results, genome assembly size (57.21 Gb) obtained by the current study for *C. gloeosporioides* strains was larger than the assembly size obtained in the previous studies for *C. gloeosporioides* (55.6 Mb) (Gan et al. 2013), *C. higginsianum* (53.4 Mb) (O'Connell et al. 2012), *C. orbiculare* (88.3 Mb) (Gan et al. 2013), *C. graminicola* (57.4 Mb) (O'Connell et al. 2012) and *C. acutatum* (52.1 Mb) (Han et al. 2016) as well as for other pathogenic fungi such as *Magnaporthe grisea* (38.8 Mb) (Dean et al. 2005). However, the average genome coverage of *C. gloeosporioides* strains (79.71%) in the current study is low compared to the genome coverage obtained from other studies for *C. gloeosporioides* (96.37%) (Gan et al. 2013), *C. orbiculare* (97.98%) (Gan et al. 2013) and *C. acutatum* (99.97%) (Han et al. 2016) as well as for *Magnaporthe grisea* (94%) (Dean et al. 2005). According to our transcriptomic analysis, N50 and the average length of 6.1 Gb paired-end reads produced by Illumina sequencing were 2,208 bp and 1,157 bp respectively. These relatively short reads were of high quality and suitable for novel gene discovery and SSR marker development. Among all the 25,376 unigenes annotated against Nt, Nr, Swiss-Prot, GO, KOG, and KEGG databases, 15,498 (61.07%) showed homology to sequences in the Nr database which is significantly higher than unigenes annotated in whitefly (Wang et al. 2010) and sweet potato (Wang et al. 2010). Unigenes annotated with significant matches in this study for *C. gloeosporioides* is similar to the number of genes predicted in the *C. gloeosporioides* (15,469) (Gan et al. 2013) *C. orbiculare* (13,479) (Gan et al. 2013), *C. graminicola* (12,006) (O'Connell et al. 2012), *C. acutatum* (13,559) (Han et al. 2016) and *C. higginsianum* (16,172) (O'Connell et al. 2012) genomes. Among the functional groups annotated for the unigenes of *C. gloeosporioides* CH008 transcriptome, 'cellular process' and 'metabolic process' are the most prominent sub categories belonging to the largest main category biological process. Similar results were found in *Houttuynia cordata* Thunb (Wei et al. 2014), whereas in the chickpea transcriptome, protein metabolism was selected as the dominant biological process (Garg et al. 2011). Similarly, most of the genes identified for *C. gloeosporioides* and *C. higginsianum* are associated with GO terms for primary metabolism and macromolecule biosynthesis (Gan et al. 2013).

By comparative genomic and phylogenetic analyses conducted between the seven *C. gloeosporioides* strains and the reference strain CH008, a large number of SNPs, small indels and SVs were discovered, indicating abundant polymorphism in stylo anthracnose pathogen. Compared with the CH008 strain, we identified two distinct features in the genomes of the highly pathogenic strains versus the low pathogenic strains. Specifically, genomes of the highly pathogenic strains had low single nucleotide polymorphism (SNPs) and smaller transition-to-transversion (Ti/Tv) ratio. We also found that highly pathogenic *C. gloeosporioides* has greater number of structural variations (SVs) in large fragments at the genome level.

In accordance to the pathogenicity assessment, the strains with high pathogenicity were found to be clustered in one group whereas the strains with low pathogenicity were found to be clustered in another group in the phylogenetic analyses. This indicates significant variations between strains with different pathogenicity. It can be further deduced that this variation might be caused by significant variations in the pathogenicity related genes between the two groups. When compare the annotated unigenes of these two groups, 13,357 gene differences were observed between the two groups. Using COG annotation system, 6810 different genes were annotated as indicated in the Fig. 8, whereas GO enrichment showed a high level of variability between different groups of genes among these two groups as shown in Fig. 9. According to GO classification, among the functional groups categorized under biological process, gene cluster responsible for mycelial development carry the significantly higher genetic differences between the strains of two groups. For the functional group summarized under cellular component, significantly highest genetic differences were shown in the gene cluster responsible for nucleus function. For the functional category summarized under molecular functions, gene clusters responsible for hydrolase activity, sequence-specific DNA binding RNA polymerase activity, zinc ion binding activity and protein serine/threonine kinase activity show the highest significant genetic differences among the two groups. These functional groups have also been reported in other studies for their effects in pathogenicity. For example serine protease gene family has shown to be implicating the pathogenicity of *Magnaporthe grisea*, *Neurospora crassa* (Idnurm & Howlett 2001, Dean et al. 2005), *Fusarium graminearum*, *F. oxysporum*, *C. higginsianum*, *C. graminicola*, *C. gloeosporioides* and *C. orbiculare* (Gan et al. 2013). Similar to the current study, it is observed that this serine protease family also expanded in existing *C. gloeosporioides* and *C. orbiculare* genomes (Gan et al. 2013). This observation explains the association of alkalinisation in host tissues of *C. gloeosporioides* with virulence during infection (Prusky et al. 2001). Apart from these, gene clusters with significant variations and responsible for hydrolase activity, cellulase activity and phospholipids binding activity have shown to expand in the transcriptomes of *C. higginsianum*, *C. graminicola* (O’Connell et al. 2012), *C. gloeosporioides* (Gan et al. 2013) and *C. orbiculare* (Gan et al. 2013). These enzymes help in degrading polysaccharides, hence important in establishing infection and also in accessing nutrients during different growth stages.

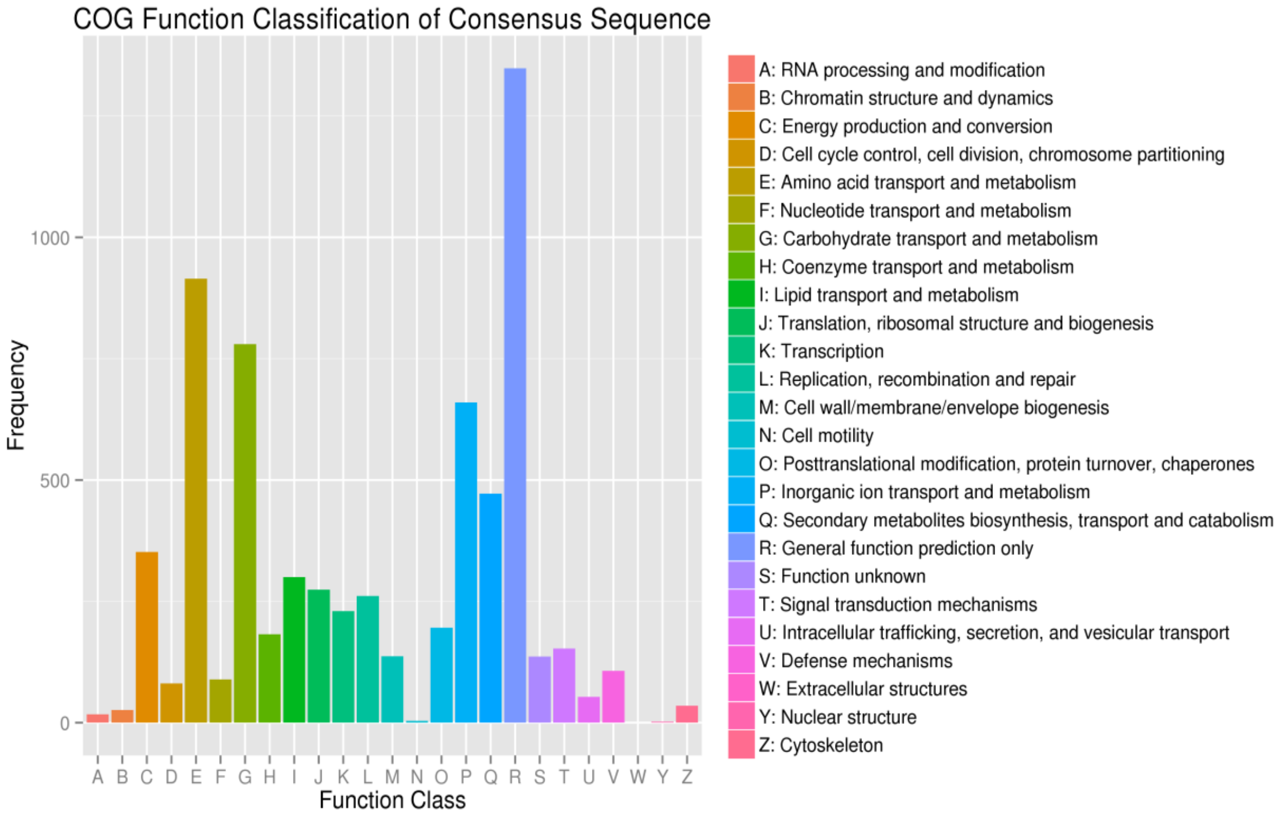

Figure 8 – Histogram showing the classification of gene differences between the highly pathogenic and weakly pathogenic groups of *Colletotrichum gloeosporioides* strains. This explains differences in the functional breakdown of COGs among the two groups.

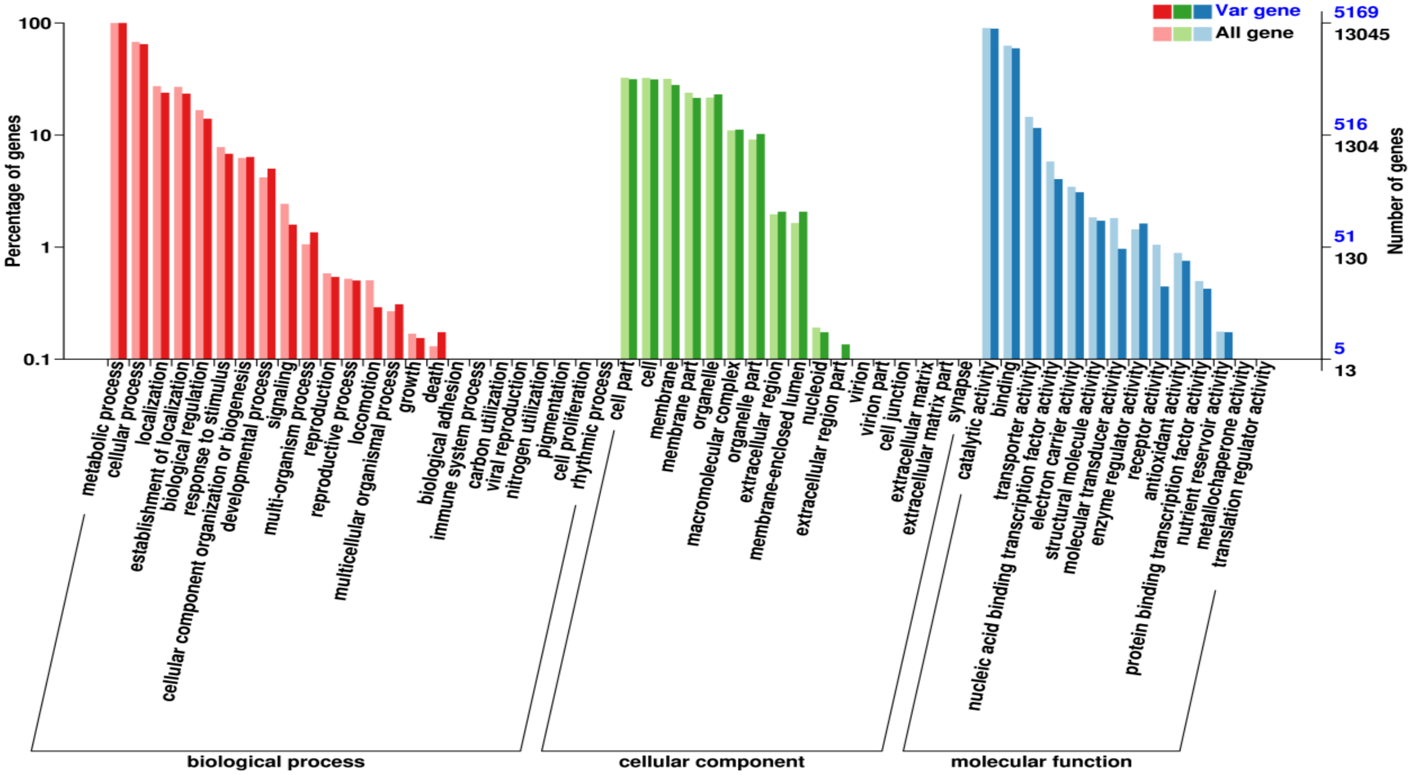

Figure 9 – Gene Ontology annotation and classification for the differences in the assembled unigenes among highly pathogenic and weakly pathogenic groups of *Colletotrichum gloeosporioides* strains. All the gene differences among these two groups were summarized into three main categories: biological processes, cellular components and molecular functions. Y-axis is number of genes in the category and the series with the light colors represent all the genes responsible for a particular function, whereas the dark colors represent the genes with the variability among the groups.

Due to the advantages of simple operation, good repetition, abundant polymorphism, large amount of genetic information, co-dominant inheritance and extensive genomic coverage (Powell et al. 1996), SSR markers have been widely applied in molecular, phenotypic, genetic and pathogenic studies in *Colletotrichum* (Kumar et al. 2011, Rampersad 2013, Saxena et al. 2014). In our study, 3,242 SSRs were developed for *C. gloeosporioides*. After the elimination of single nucleotide motifs, tri-nucleotide repeat motifs were the most abundant, followed by di-nucleotide repeats. This is consistent with the previous reports of the species. However, contrasting results were reported on plant species such as sweet potato (Wang et al. 2010), celery (Wang & Shen 2013) and *Oenanthe javanica* (Jiang et al. 2015), in which di-nucleotide motifs were the most abundant types. Further analysis showed that the AC/GT di-nucleotide repeat was the most abundant motif detected in our SSRs, followed by AG/CT. Finally, 1,505 SSR primers were successfully designed and 80% of randomly selected primers were validated by PCR. These results indicate that the SSR markers developed in our study will facilitate marker-assisted genomic and pathogenic diversity analysis in *C. gloeosporioides.*

The data generated from the genomes of seven diverse pathogenic *C. gloeosporioides* strains together with their transcriptomic data will provide a foundation for future in-depth studies on pathogenic mechanisms and pathogenicity related genes of *C. gloeosporioides*. In addition, the transcriptome sequences obtained can also be used as reference sequences for future gene expression studies and as a transcriptomic data resource for further molecular studies of *C. gloeosporioides*.

Acknowledgements

We thank the National Natural Science Foundation (31072076), the Special Fund for Agro-scientific Research in the Public Interest (201303057) and the Major Scientific and Technical Project of Hainan Province (ZDZX2013023-1) for supporting this research.

References

Abang MM, Abraham WR, Asiedu R, Hoffmann P, Wolf G, Winter S. 2009 – Secondary metabolite profile and phytotoxic activity of genetically distinct forms of *Colletotrichum gloeosporioides* from yam (*Dioscorea* spp.). Mycological Research 113(1), 130–140.

Amselem J, Cuomo CA, van Kan JA, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S, Fournier E, Gout L, Hahn M, Kohn L, Lapalu N, Plummer KM, Pradier JM, Quévillon E, Sharon A, Simon A, ten Have A, Tudzynski B, Tudzynski P, Wincker P, Andrew M, Anthouard V, Beever RE, Beffa R, Benoit I, Bouzid O, Brault B, Chen Z, Choquer M, Collémare J, Cotton P, Danchin EG, Da Silva C, Gautier A, Giraud C, Giraud T, Gonzalez C, Grossetete S, Güldener U, Henrissat B, Howlett BJ, Kodira C, Kretschmer M, Lappartient A, Leroch M, Levis C, Mauceli E, Neuvéglise C, Oeser B, Pearson M, Poulain J, Poussereau N, Quesneville H, Rascle C, Schumacher J, Ségurens B, Sexton A, Silva E, Sirven C, Soanes DM, Talbot NJ, Templeton M, Yandava C, Yarden O, Zeng Q, Rollins JA, Lebrun MH, Dickman M. 2011 – Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS Genetics 7(8), article e1002230.

Bentley DR. 2006 – Whole genome re-sequencing. Current Opinion in Genetics and Development 16(6), 545–552.

Braithwaite KS, Irwin JAG, Manners JM. 1990 – Restriction fragment length polymorphisms in *Colletotrichum gloeosporioides* infecting *Stylosanthes* spp. in Australia. Mycological Research 94(8), 1129–1137.

Conesa A, Götz S, García-Gómez JM, Terol J, Talón M, Robles M. 2005 – Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21(18), 3674–3676.

Chakraborty S, Fernandes CD, Charchar MJ, Thomas MR. 2002 – Pathogenic variation in *Colletotrichum gloeosporioides* infecting Stylosanthes spp. in a Center of diversity in Brazil. Phytopathology 92(5), 553–562.

Chen H, Hu C, Yi K, Huang G, Gao J, Zhang S, Zheng J, Liu Q, Xi J. 2014 – Cloning of insertion site flanking sequence and construction of transfer DNA insert mutant library in Stylosanthes *Colletotrichum*. PLoS One 9(10), article e111172.

Chen H, Zha J, Liang X, Bu J, Wang M, Wang Z. 2013 – Sequencing and de novo assembly of the Asian Clam (*Corbicula fluminea*) transcriptome using the Illumina GAIIx method. PLoS One 8(11), article e79516.

Chen K, Wallis JW, McLellan MD, Larson DE, Kalicki JM, Pohl CS, McGrath SD, Wendl MC, Zhang Q, Locke DP, Shi X, Fulton RS, Ley TJ, Wilson RK, Ding L, Mardis ER. 2009 – Break Dancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods 6, 677–681.

Chen W, Liu YX, Jiang GF. 2015 – De novo assembly and characterization of the testis transcriptome and development of EST-SSR Markers in the cockroach *Periplaneta Americana*. Science Reports 5, 11144.

Submitted 27 September 2016, Accepted 6 November 2016, Published 10 November 2016

Corresponding Author: Liang CH – e-mail – address: liangchunhao99@126.com

Cingolani P, Platts A, Wangle L, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012 – A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. Fly (Austin) 6, 80–92.

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, Read ND, Lee YH, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun MH, Bohnert H, Coughlan S, Butler J, Calvo S, Ma LJ, Nicol R, Purcell S, Nusbaum C, Galagan JE, Birren BW. 2005 – The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434, 980–986.

Feng S, Li F, He C, Lin J. 1994 – Bionomics and epidemiology of anthracnose on Stylosanthes spp. Chinese Journal of Tropical Crops 15, 87–94.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2014 – The Pfam protein families database: towards a more sustainable future. Nucleic Acids Research 44(D1), D222–D230.

Fu N, Wang Q, Shen HL. 2013 – De novo assembly, gene annotation and marker development using Illumina paired-end transcriptome sequences in Celery (*Apium graveolens* L.). PLoS One 8(2), article e57686.

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C, Birren B. 2003 – "The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422(6934), 859–868.

Gan P, Ikeda K, Irieda H, Narusaka M, O'Connell RJ, Narusaka Y, Takano Y, Kubo Y, Shirasu K. 2013 – Comparative genomic and transcriptomic analyses reveal the hemi biotrophic stage shift of *Colletotrichum* fungi. New Phytologist 197(4), 1236–1249.

Garg R, Patel RK, Tyagi AK, Jain M. 2011 – De novo assembly of Chickpea transcriptome using short reads for gene discovery and marker identification. DNA Research 18(1), 53–63.

Ge X, Chen H, Wang H, Shi A, Liu K. 2014 – De novo assembly and annotation of *Salvia splendens* transcriptome using the Illumina Platform. PLoS One 9(3), article e87693.

Goodwin SB, M'barek SB, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, Van der Lee TA, Grimwood J, Aerts A, Antoniw J, Bailey A, Bluhm B, Bowler J, Bristow J, van der Burgt A, Canto-Canché B, Churchill AC, Conde-Ferràez L, Cools HJ, Coutinho PM, Csukai M, Dehal P, De Wit P, Donzelli B, van de Geest HC, van Ham RC, Hammond-Kosack KE, Henrissat B, Kilian A, Kobayashi AK, Koopmann E, Kourmpetis Y, Kuzniar A, Lindquist E, Lombard V, Maliepaard C, Martins N, Mehrabi R, Nap JP, Ponomarenko A, Rudd JJ, Salamov A, Schmutz J, Schouten HJ, Shapiro H, Stergiopoulos I, Torriani SF, Tu H, de Vries RP, Waalwijk C, Ware SB, Wiebenga A, Zwiers LH, Oliver RP, Grigoriev IV, Kema GH. 2011 – Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genetics 7(6), article e1002070.

Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman N, Regev A. 2011 – Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nature Biotechnology 29(7), 644–652.

Han JH, Chon JK, Ahn JH, Choi IY, Lee YH, Kim KS. 2016 – Whole genome sequence and genome annotation of *Colletotrichum acutatum*, causal agent of anthracnose in pepper plants in South Korea. Genomics data 8, 45–46.

Hegedus Z, Zakrzewska A, Agoston VC, Ordas A, Rácz P, Mink M, Spaink HP, Meijer AH. 2009 – Deep sequencing of the zebrafish transcriptome response to mycobacterium infection. Molecular Immunology 46(15), 2918–2930.

Idnurm A, Howlett BJ. 2001 – Pathogenicity genes of phytopathogenic fungi. Molecular Plant Pathology 2, 241–255.

Jiang CS, Ma XR, Zhou DM, Zhang YZ. 2005 – AFLP analysis of genetic variability among Stylosanthes guianensis accessions resistant and susceptible to the stylo anthracnose. Plant Breeding 124(6), 595–598.

Jiang Q, Wang F, Tan HW, Li MY, Xu ZS, Tan GF, Xiong AS. 2015 – De novo transcriptome assembly, gene annotation, marker development, and miRNA potential target genes validation under abiotic stresses in Oenanthe javanica. Molecular Genetics and Genomics 290(2), 671–683.

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, Yamanishi Y. 2008 – KEGG for linking genomes to life and the environment. Nucleic Acids Research 36(Database issue), D480–D484.

Kelemu S, Badel JL, Moreno and CX, Miles JW. 1996 – Virulence spectrum of south American isolates of *Colletotrichum gloeosporioides* on selected Stylosanthes guianensis genotypes. Plant Disease 80(12), 1355–1358.

Kelemu S, Skinner DZ, Badel JL, Moreno CX, Rodriguez MX, Fernandes CD, Charchar MJ, Chakraborty S. 1999 – Genetic diversity in South American *Colletotrichum gloeosporioides* isolates from Stylosanthes guianensis, a tropical forage legume. European Journal of Plant Pathology 105(3), 261–272.

Kim YK, Liu ZM, Li D, Kolattukudy PE. 2000 – Two novel genes induced by hard-surface contact of *Colletotrichum gloeosporioides* conidia. Journal of Bacteriology 182(17), 4688–4695.

Kleemann J, Rincon-Rivera LJ, Takahara H, Neumann U, Ver Loren van Themaat E, van der Does HC, Hacquard S, Stüber K, Will I, Schmalenbach W, Schmelzer E, O'Connell RJ. 2012 – Sequential delivery of host-induced virulence effectors by appressoria and intracellular hyphae of the phytopathogen *Colletotrichum higginsianum*. PLoS Pathogens 8(4), article e1002643.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009 – Circos: An information aesthetic for comparative genomics. Genome Research 19(9), 1639–1645.

Kumar N, Jhang T, Sharma TR. 2011 – Molecular and pathological characterization of *Colletotrichum falcatum* infecting subtropical indian sugarcane. Journal of Phytopathology 159(4), 260–267.

Li H, Durbin R. 2010 – Fast and accurate long-read alignment with Burrows Wheeler transform. Bioinformatics 26(5), 589–595.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R; 1000 Genome Project Data Processing Subgroup. 2009 – The sequence alignment/map format and SAM tools. Bioinformatics 25(16), 2078–2079.

Li R, Li Y, Fang X, Yang H, Wang J, Kristiansen K, Wang J. 2009 – SNP detection for massively parallel whole-genome re-sequencing. Genome Research 19(6), 545–552.

Li R, Zhu H, Ruan J, Qian W, Fang X, Shi Z, Li Y, Li S, Shan G, Kristiansen K, Li S, Yang H, Wang J, Wang J. 2010 – De novo assembly of human genomes with massively parallel short read sequencing. Genome research 20(2), 265–272.

Maher CA, Palanisamy N, Brenner JC, Cao X, Kalyana-Sundaram S, Luo S, Khrebtukova I, Barrette TR, Grasso C, Yu J, Lonigro RJ, Schroth G, Kumar-Sinha C, Chinnaiyan AM.

2009 – Chimeric transcript discovery by paired-end transcriptome sequencing. Proceedings of the National Academy of Sciences of the United States of America 106(30), 12353–12358.

Manners JM, Masel AM, Braithwaite KS, et al. 1992 – Molecular analysis of *Colletotrichum gloeosporioides* pathogenic on the tropical pasture legumes Stylosanthes spp. Oxford, UK, CAB International, pp. 250–268.

Manners JM, Stephenson SA, He C, Maclean DJ. 2000 – Gene transfer and expression in *Colletotrichum gloeosporioides* causing anthracnose on Stylosanthes. St. Paul, USA, The American Phytopathological Society, pp. 180–194.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010 – The genome analysis toolkit: A map reduce framework for analyzing next-generation DNA sequencing data. Genome Research 20(9), 1297–1303.

Michalk DL, Nan-Ping F, Chin-Ming Z. 1993 – Improvement of dry tropical rangelands in Hainan island, China: 1. Evaluation of pasture legumes. Journal of Range Management 46(4), 331–339.

Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M. 2007 – KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Research 35(Web server issue), W182–W185.

Munaut F, Hamaide N, Maraite H. 2001 – Molecular and morphological characterization of *Colletotrichum gloeosporioides* from native Mexican Stylosanthes species. Plant Pathology 50(3), 383–396.

Munaut F, Hamaide N, Maraite H. 2002 – Genomic and pathogenic diversity in *Colletotrichum gloeosporioides* from wild native Mexican Stylosanthes spp. and taxonomic implications. Mycological Research 106(5), 579–593.

Münch S, Lingner U, Floss DS, Ludwig N, Sauer N, Deising HB. 2008 – The hemi biotrophic lifestyle of *Colletotrichum* species. Journal of Plant Physiology 165(1), 41–51.

Narduzzi-Wicht B, Jermini M, Gessler C, Broggini GAL. 2014 – Microsatellite markers for population studies of the ascomycete *Phyllosticta ampelicida*, the pathogen causing grape black rot. Phytopathologia Mediterranea 53(3), 470−479.

Narusaka Y, Narusaka M, Park P, Kubo Y, Hirayama T, Seki M, Shiraishi T, Ishida J, Nakashima M, Enju A, Sakurai T, Satou M, Kobayashi M, Shinozaki K. 2004 – RCH1, a locus in Arabidopsis that confers resistance to the hemi biotrophic fungal pathogen *Colletotrichum higginsianum*. Molecular Plant-Microbe Interactions 17(7), 749–762.

Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F, Hesse CN, Kosti I, LaButti K, Lindquist EA, Lucas S, Salamov AA, Bradshaw RE, Ciuffetti L, Hamelin RC, Kema GH, Lawrence C, Scott JA, Spatafora JW, Turgeon BG, de Wit PJ, Zhong S, Goodwin SB, Grigoriev IV. 2012 – Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathogens 8(12), article e1003037.

O'Connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, Damm U, Buiate EA, Epstein L, Alkan N, Altmüller J, Alvarado-Balderrama L, Bauser CA, Becker C, Birren BW, Chen Z, Choi J, Crouch JA, Duvick JP, Farman MA, Gan P, Heiman D, Henrissat B, Howard RJ, Kabbage M, Koch C, Kracher B, Kubo Y, Law AD, Lebrun MH, Lee YH, Miyara I, Moore N, Neumann U, Nordström K, Panaccione DG, Panstruga R, Place M, Proctor RH, Prusky D, Rech G, Reinhardt R, Rollins JA, Rounsley S, Schardl CL, Schwartz DC, Shenoy N, Shirasu K, Sikhakolli UR, Stüber K, Sukno SA, Sweigard JA, Takano Y, Takahara H, Trail F, van der Does HC, Voll LM, Will I, Young S, Zeng Q, Zhang J, Zhou S, Dickman MB, Schulze-Lefert P, Ver Loren van Themaat E, Ma LJ, Vaillancourt LJ. 2012 – Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. Nature Genetics 44(9), 1060–1065.

Perfect SE, Hughes HB, O'Connell RJ, Green JR. 1999 – *Colletotrichum*: A model genus for studies on pathology and fungal-plant interactions. Fungal Genetics and Biology 27(2–3), 186–198.

Powell W, Morgante M, Andre C, Hanafey M, Vogel J, Tingey S, Rafalski A. 1996 – The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Molecular Breeding 2(3), 225–238.

Powell NT. 1971 – Disease complexes in tobacco involving *Meloidogyne incognita* and certain soil-borne fungi. Phytopathology 61(11), 1332–1337.

Prusky D, McEvoy JL, Leverentz B, Conway WS. 2001 – Local modulation of host pH by *Colletotrichum* species as a mechanism to increase virulence. Molecular Plant-Microbe Interactions 14(9), 1105–1113.

Rampersad SN. 2013 – Genetic structure of *Colletotrichum gloeosporioides* sensu lato isolates infecting papaya inferred by multi locus ISSR markers. Phytopathology 103(2), 182–189.

Rice P, Longden I, Bleasby A. 2000 – EMBOSS: The European molecular biology open software suite. Trends in Genetics 16(6), 276–277.

Robinson M, Riov J, Sharon A. 1998 – Indole-3-acetic acid biosynthesis in *Colletotrichum gloeosporioides* f. sp. *aeschynomene*. Applied and Environmental Microbiology 64(12), 5030–5032.

Saxena A, Raghuwanshi R, Singh HB. 2014 – Molecular, phenotypic and pathogenic variability in *Colletotrichum* isolates of subtropical region in north-eastern India, causing fruit rot of chillies. Journal of Applied Microbiology 117(5), 1422–1434.

Sharma KR, Bhagya N, Sheik S, et al. 2011– "Isolation of endophytic *Colletotrichum gloeosporioides* Penz. From *Salacia chinensis* and its antifungal sensitivity. Journal of Phytology 3, 20–22.

Stamatakis A. 2014 – RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9), 1312–1313.

Stephenson SA, Hatfield J, Rusu AG, Maclean DJ, Manners JM. 2000 – CgDN3: An essential pathogenicity gene of *Colletotrichum gloeosporioides* necessary to avert a hypersensitive-like response in the host *Stylosanthes guianensis*. Molecular Plant-Microbe Interactions 13(9), 929–941.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA. 2003 – The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4, 41.

Tatusov RL, Koonin EV, Lipman DJ. 1997 – A genomic perspective on protein families. Science 278(5338), 631–637.

Wang B, Guo G, Wang C, Lin Y, Wang X, Zhao M, Guo Y, He M, Zhang Y, Pan L. 2010 – Survey of the transcriptome of *Aspergillus oryzae* via massively parallel mRNA sequencing. Nucleic Acids Research 38(15), 5075–5087.

Wang XW, Luan JB, Li JM, Bao YY, Zhang CX, Liu SS. 2010 – De novo characterization of a whitefly transcriptome and analysis of its gene expression during development. BMC Genomics 11, 400.

Wang Z, Fang B, Chen J, Zhang X, Luo Z, Huang L, Chen X, Li Y. 2010 – De novo assembly and characterization of root transcriptome using Illumina paired-end sequencing and development of cSSR markers in sweetpotato (*Ipomoea batatas*). BMC Genomics 11, 726.

Wei L, Li S, Liu S, He A, Wang D, Wang J, Tang Y, Wu X. 2014 – Transcriptome analysis of *Houttuynia cordata* Thunb. by Illumina paired-end RNA sequencing and SSR marker discovery. PLoS One 9(1), article e84105.

Weeds PL, Chakraborty S, Fernandes CD, d'A Charchar MJ, Ramesh CR, Kexian Y, Kelemu S. 2003 – Genetic diversity in *Colletotrichum gloeosporioides* from Stylosanthes spp. at centers of origin and utilization. Phytopathology 93(2), 176–185.

Wu T, Qin Z, Zhou X, Feng Z, Du Y. 2010 – Transcriptome profile analysis of floral sex determination in cucumber. Journal of Plant Physiology 167(11), 905–913.

Verbruggen B, Bickley LK, Santos EM, Tyler CR, Stentiford GD, Bateman KS, van Aerle R. 2015 – De novo assembly of the *Carcinus maenas* transcriptome and characterization of innate immune system pathways. BMC Genomics 16(1), 1–17.

Xing M, Lv H, Ma J, Xu D, Li H, Yang L, Kang J, Wang X, Fang Z. 2016 – Transcriptome profiling of resistance to *Fusarium oxysporum* f. sp. *conglutinans* in Cabbage (*Brassica oleracea*) Roots. PLoS One 11(2), article e0148048.

Xue M, Yang J, Li Z, Hu S, Yao N, Dean RA, Zhao W, Shen M, Zhang H, Li C, Liu L, Cao L, Xu X, Xing Y, Hsiang T, Zhang Z, Xu JR, Peng YL. 2012 – Comparative analysis of the genomes of two field isolates of the rice blast fungus *Magnaporthe oryzae*. PLoS Genetics 8(8), article e1002869.

Yap HY, Chooi YH, Fung SY, Ng ST, Tan CS, Tan NH. 2015 – Transcriptome analysis revealed highly expressed genes encoding secondary metabolite pathways and small cysteine-rich proteins in the sclerotium of *Lignosus rhinocerotis*. PLoS One 10(11), article e0143549.

Yi KX. 2001 – Stylo anthracnose and current research progresses on Anthracnose resistance breeding. Chinese Journal of Grassland 23(4), 59–65.

Yin Z, Liu H, Li Z, Ke X, Dou D, Gao X, Song N, Dai Q, Wu Y, Xu JR, Kang Z, Huang L. 2015 – Genome sequence of Valsa canker pathogens uncovers a potential adaptation of colonization of woody bark. New Phytologist 208(4), 1202–1216.

Yoshino K, Irieda H, Sugimoto F, Yoshioka H, Okuno T, Takano Y. 2012 – Cell death of *Nicotiana benthamiana* is induced by secreted protein NIS1 of *Colletotrichum orbiculare* and is suppressed by a homologue of CgDN3. Molecular Plant-Microbe Interactions 25(5), 625–636.

Zhang Y, Zhang K, Fang A, Han Y, Yang J, Xue M, Bao J, Hu D, Zhou B, Sun X, Li S, Wen M, Yao N, Ma LJ, Liu Y, Zhang M, Huang F, Luo C, Zhou L, Li J, Chen Z, Miao J, Wang S, Lai J, Xu JR, Hsiang T, Peng YL, Sun W. 2014 – Specific adaptation of *Ustilaginoidea virens* in occupying host florets revealed by comparative and functional genomics. Nature Communications 5, 3849.

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

Zhou SM, Chen LM, Liu SQ, Wang XF, Sun XD. 2015 – De Novo assembly and annotation of the Chinese Chive (*Allium tuberosum* Rottler ex Spr.) transcriptome using the Illumina Platform. PLoS One 10(5), article e0133312.