Genomics Data 13 (2017) 54–63

Contents lists available at ScienceDirect

Genomics Data

journal homepage: www.elsevier.com/locate/gdata

A comparative genome analysis of *Cercospora sojina* with other members of the pathogen genus *Mycosphaerella* on different plant hosts

Fanchang Zeng[a,b,1], Xin Lian[a,1], Guirong Zhang[b], Xiaoman Yu[a], Carl A. Bradley[c,d], Ray Ming[b,e,*]

[a] State Key Laboratory of Crop Biology, College of Agronomy, Shandong Agricultural University, Tai'an, Shandong 271018, China
[b] Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA
[c] Department of Plant Pathology, University of Kentucky, Princeton, KY 42445, USA
[d] Department of Crop Sciences, University of Illinois, Urbana, IL 61801, USA
[e] FAFU and UIUC-SIB Joint Center for Genomics and Biotechnology, Fujian Agriculture and Forestry University, Fuzhou, Fujian 350002, China

ARTICLE INFO

Keywords:
Phytopathogenic fungi
Mycosphaerella pathogens
Genome sequence
Comparative genomics

ABSTRACT

Fungi are the causal agents of many of the world's most serious plant diseases causing disastrous consequences for large-scale agricultural production. Pathogenicity genomic basis is complex in fungi as multicellular eukaryotic pathogens. Here, we report the genome sequence of *C. sojina*, and comparative genome analysis with plant pathogen members of the genus *Mycosphaerella* (*Zymoseptoria. tritici* (synonyms *M. graminicola*), *M. pini*, *M. populorum* and *M. fijiensis* - pathogens of wheat, pine, poplar and banana, respectively). Synteny or collinearity was limited between genomes of major *Mycosphaerella* pathogens. Comparative analysis with these related pathogen genomes indicated distinct genome-wide repeat organization features. It suggests repetitive elements might be responsible for considerable evolutionary genomic changes. These results reveal the background of genomic differences and similarities between *Dothideomycete* species. Wide diversity as well as conservation on genome features forms the potential genomic basis of the pathogen specialization, such as pathogenicity to woody vs. herbaceous hosts. Through comparative genome analysis among five *Dothideomycete* species, our results have shed light on the genome features of these related fungi species. It provides insight for understanding the genomic basis of fungal pathogenicity and disease resistance in the crop hosts.

1. Introduction

A number of genome sequences of plant pathogenic fungi in the genus *Mycosphaerella* that cause economically important disease of major crop hosts have been released [1–4]. In addition, the fungus *Cercosopora sojina* is a plant pathogen that threatens global soybean supplies. The teleomorphs of *Cercospora* species with identified sexual stages are in the genus *Mycosphaerella* [5]. Recently, we sequenced and released the genome sequence of *C. sojina*, which would greatly expand the range for comparative analysis of the closely related members in the genus *Mycosphaerella*, and may provide new insight into the genomic basis of phytopathogenicity biology. It is essential for designing strategies to manage destructive disease in different major crop hosts effectively.

The genus *Mycosphaerella* and its associated anamorphs comprise one of the largest groups of plant-pathogenic fungi. Many *Mycosphaerella* species are important pathogens causing leaf spotting diseases in a wide variety of economically important crops including cereals, banana, woody plants, citrus, eucalypts, soft fruits and horticultural crops. Two of the most important pathogens of wheat and banana are *Z. tritici* (formerly known as synonyms *M. graminicola*) and *M. fijensis*, which cause Septoria leaf blotch and black Sigatoka leaf spot, respectively [6,7]. These diseases occur in most wheat- and banana-producing areas throughout the world every year. *Mycosphaerella pini* and *Mycosphaerella populorum* are foliar pathogens of many pine and poplars species respectively, causing serious economic losses on forests and ecological deterioration world-wide. Pines account for the majority of commercial forest products and important members of native forests in many countries. And poplars, as the model organism for forest tree research, are valued as a future source for biofuel. Because of the undisputed economic and ecological importance, understanding these foliar pathogens at the genome level is the basis for developing new methods to manage the disease. These pathogens together represent the *Mycosphaerella* branch of the fungal evolutionary tree. Phylogenetically, species of *Mycosphaerella* are close relatives of the soybean frogeye leaf spot pathogen, *C. sojina* [8]. No sexual

* Corresponding author at: Department of Plant Biology, University of Illinois, Urbana, IL 61801, USA.
E-mail addresses: fczeng@sdau.edu.cn (F. Zeng), grzhang@illinois.edu (G. Zhang), carl.bradley@uky.edu (C.A. Bradley), rming@life.uiuc.edu, rming@life.illinois.edu (R. Ming).
[1] F. Z. and X. L. contributed equally to this work.

http://dx.doi.org/10.1016/j.gdata.2017.07.007
Received 17 February 2017; Received in revised form 17 May 2017; Accepted 7 July 2017
Available online 08 July 2017

(teleomorphic) stage of *C. sojina* has been identified, but the teleomorphs of *Cercospora* species with identified sexual stages are in the genus *Mycosphaerella*. However, the host diversity of these pathogens suggests pathogen specialization besides the common pathogenicity mechanisms in fungi. Therefore, the availability of genome data from *C. sojina* as well as the closely related species provides the opportunity for comparative genome analysis. It could be valuable in identifying the genome features that can be exploited for better control of disease epidemics.

Through high-throughput DNA sequencing and large-scale comparative genomics of phytopathogenic fungi in this report, we present a genome profile of plant pathogenic fungi in the genus *Mycosphaerella*, and we investigate the genome sequence background of related pathogens to reveal the differences on genome features involved in pathogenicity to different hosts such as woody vs. herbaceous plants, small crop vs. large trees, tropical and temperate zones crops.

2. Material and methods

2.1. C. sojina whole-genome sequencing and assembly

Genome of *C. sojina* strain S9, from a soybean field in Georgia, was sequenced using the Illumina GA IIx next generation technology by paired-end sequencing method to a depth of 239 × at Keck Center at University of Illinois Urbana-Champaign. The produced sequences had read length of 124 base pairs (bp). A total of 29,619,123 reads from each end were produced for a total of 59,238,246 reads from one lane. *C. sojina* genomes were assembled using Velvet algorithm to obtain optimized results with high quality assembly.

2.2. C. sojina genome annotation

The *C. sojina* genes were predicted with ab initio gene finders (FGENESH, FGENESH +, and GENEWISE). We referred the gene models from *Z. tritici* (*M. graminicola*) as the closest species with *C. sojina* to train the gene finding programs. BlastX against publicly available nonredundant protein and BlastN against ESTs databases are used to validate and curate predicted complete coding regions of the gene models. The entire DNA sequence was also compared against the nonredundant protein databases in all six reading frames, using BlastX with threshold E < 1e-5 to identify any possible coding sequences previously missed by using ARTEMIS to collate data and facilitate annotation. Finally, a non-redundant set of gene models is produced, in which a single best gene model per locus is selected, preferring the candidate annotation with supporting evidence of homolog protein/EST sequence in public database and complete coding sequence region.

2.3. Genomics synteny mapping, genome wide orthologous genes annotation and evolutionary relationships across multiple species

Comparison of the genome of *C. sojina* with Joint Genome Institute (JGI) released other four related fungal genomes (*M. graminicola* v2.0, *M. pini* v1.0, *M. populorum* v1.0, *M. fijiensis* v2.0) was performed using Synteny Mapping and Analysis Program package (SyMAP v3.4) for detecting and displaying syntenic relationships between sequenced genome [9,10] in compliance with instruction from the package. Genome wide comparison and annotation of orthologous genes across multiple species was performed using OrthoVenn [11] with the default settings. Basic cladogram for evolutionary relationships analysis was carried out with the software Mauve with the default settings from whole-genome ortholgous gene sequence data [12].

2.4. Genomics repeat structure profile and mapping

For the genome repeat sequences features of *C. sojina* and other four related fungal, genome repeat sequences structure were detected and repeat organization map were generated by Pygram pipeline [13], as an efficient genome repeat analysis tool, which provide an representation of the organization of repeated structures including frequency visualization in multi-genomes for discovering new structure features and specific repeat properties.

2.5. Genomics functional annotation

All predicted genes are annotated for function and physiology pathway using Blast2Go function annotation system [14,15], according to Gene Ontology (GO), eukaryotic orthologous groups (KOGs), and KEGG metabolic pathways. For annotated *C. sojina* genes, where possible, assigned predicted protein functions using a combination of sequence comparison with BlastP and domains/motif identification with interProScan [14] and PFAM [16].

2.6. Inter-species genome-wide genes annotation comparison

Large-scale genome-wide genes GO functional comparison in all these related fungi were explored and plotted by program WEGO [17]. Comparison histogram were displayed with all items at different GO level separately, including the default second level and the third level, as well as level limited to only items with significant relationship for the genome dataset compared base on Pearson Chi-Square test (Significance level is below the 0.05, expected item counts are greater than 5).

3. Results

3.1. Overall genome comparison

The *C. sojina* genome was compared with the four fungal genomes of close relatives in the same genus. Limited similarity was present among these phylogenetically close fungal genomes with the exception of the *C. sojina* and *Z. tritici* (*M. graminicola*) genomes, as shown in dot plot alignment mapping and 3D schematic view (Fig. 1). The homologous genome blocks between *C. sojina* and each of the other four species displayed 81.0% to 85.9% nuclear acid identity with an average of only 82.9%, for they are from close relatives (Additional file 1 Table S1). However, if counting the non-homologous regions at whole genome level, the sequences similarity decreases dramatically.

Whole-genome orthologous genes were comprehensively investigated in five close species. The overall comparison analysis result was displayed as Venn diagram in Fig. 2A. Total 5020 orthologous genes were shared among all five species, which highlights these five species are close relatives with considerable common coding gene. Whole genome phylogenetic analysis in our study reveals the evolutionary relationships of these five additional important close species which were not included in previous study of Goodwin SB, 2001 except specie of *C. sojina* [8]. Among the five relatives, *M. populorum* was found as the closest specie to *C. sojina* (Fig. 2B).

3.2. Genome synteny and genomic changes

The availability of these related fungal genome sequences has allowed comparisons of synteny among different species. These five fungi genomes provide an opportunity to study eukaryotic genome differences and similarities between the genomes. The genomes of *C. sojina* shared 92% overall synteny with that of *Z. tritici* (*M. graminicola*) (Additional file 1 Table S1). In contrast, only 46%, 63%, 66% of *C. sojina* genome assembly could be mapped to conserved syntenic blocks of the other three genomes of *M. pini*, *M. fijiensis*, and *M. populorum*, respectively (Additional file 1 Table S1). While the overall synteny between *C. sojina* and *Z. tritici* (*M. graminicola*) genomes was conserved, considerable rearrangements were detected among genomes of *C. sojina* and other three fungi (Fig. 3 and Additional file 2 Fig. S1). Previous