Genomics Data 5 (2015) 284–291

Contents lists available at ScienceDirect

Genomics Data

journal homepage: http://www.journals.elsevier.com/genomics-data/

Genome analysis of rice-blast fungus *Magnaporthe oryzae* field isolates from southern India

Malali Gowda [a,*], Meghana D. Shirke [a], H.B. Mahesh [a,c], Pinal Chandarana [b], Anantharamanan Rajamani [a], Bharat B. Chattoo [b,*]

[a] Genomics Laboratory, Centre for Cellular and Molecular Platforms, Bangalore 560065, India
[b] Centre for Genome Research, Department of Microbiology and Biotechnology Centre, Faculty of Science, Maharaja Sayajirao University of Baroda, Vadodara 390002, India
[c] Marker Assisted Selection Laboratory, Department of Genetics and Plant Breeding, University of Agricultural Sciences, Bangalore, India

ARTICLE INFO

Article history:
Received 7 May 2015
Accepted 3 June 2015
Available online 20 June 2015

Keywords:
Genome comparison
Next generation sequencing
Magnaporthe
Single nucleotide polymorphism
Isolate specific genes

ABSTRACT

The Indian subcontinent is the center of origin and diversity for rice (*Oryza sativa* L.). The *O. sativa* ssp. *indica* is a major food crop grown in India, which occupies the first and second position in area and production, respectively. Blast disease caused by *Magnaporthe oryzae* is a major constraint to rice production. Here, we report the analysis of genome architecture and sequence variation of two field isolates, B157 and MG01, of the blast fungus from southern India. The 40 Mb genome of B157 and 43 Mb genome of MG01 contained 11,344 and 11,733 predicted genes, respectively. Genomic comparisons unveiled a large set of SNPs and several isolate specific genes in the Indian blast isolates. Avr genes were analyzed in several sequenced *Magnaporthe* strains; this analysis revealed the presence of Avr-Pizt and Avr-Ace1 genes in all the sequenced isolates. Availability of whole genomes of field isolates from India will contribute to global efforts to understand genetic diversity of *M. oryzae* population and to track the emergence of virulent pathotypes.

1. Introduction

Rice (*Oryza sativa*) is the staple food for more than half of the world population. Rice blast disease caused by the Ascomycetes fungus *Magnaporthe oryzae* is a predominant biotic stress affecting rice production worldwide. The outbreak of wheat (*Triticum aestivum*) blast in Brazil in early 1990s is an important example of host shift and coevolution of this fungus in recent times [1]. The blast fungus is also known to infect other food crops such as finger millet, small millets, barley, and most of the growth stages in rice including leaf, stem, nodes, panicle and root [2,3]. Thus, blast disease is a serious constraint in cereal crop production in India and at the global level. High genetic variability in *M. oryzae* isolates poses a major challenge to rice breeders and pathologists to control blast disease.

The genome of *M. oryzae* strain 70-15 was the first to be sequenced among plant pathogenic fungi using Sanger sequencing method [4]. The 70-15 isolate was derived from a cross between isolates of rice and weeping lovegrass and further backcross to rice isolate [5]. Subsequently, several field isolates of blast have been sequenced using next generation sequencing (NGS). Field isolates from Japan (Ina168 and P131) and China (Y34) [6,7] were sequenced using 454 sequencing. More recently two field isolates, FJ81278 and HN19311 from China have been sequenced using Illumina technology [8]. Interestingly, whole genome sequencing of multiple isolates has revealed over a mega-base pairs of novel genomic regions and hundreds of novel genes. This could be due to race evolution over a period of time by geographical separation, chromosomal variation and variability in repetitive elements [4,7,8]. Multiple clonal lineages of *Magnaporthe* are known to exist around various cropping zones in the world.

The Indian subcontinent is a center of origin and diversity for *Magnaporthe* species complex, but we lack the information about genomic variability among *M. oryzae* isolates. To understand the genomic variation within field isolates of *M. oryzae*, we carried out whole genome sequencing of two Indian strains, B157 and MG01 using Illumina sequencing technology. B157 is commonly used virulent reference strain by several research groups in India for many years [9,10] and MG01 is recently isolated virulent strain from HR12 cultivar. In addition, we used RNAseq to understand strand specific gene expression, which is not explored in pathogenic fungi. This is the first of its kind to compare blast fungal isolates in India at the genome level. This work will certainly be useful to pathologists and breeders to understand *Magnaporthe* virulence and improve blast resistance in rice.

* Corresponding authors. Tel.: +91 80 67185113.
E-mail addresses: malalig@ccamp.res.in (M. Gowda), bharat.chattoo@bcmsu.ac.in (B.B. Chattoo).

http://dx.doi.org/10.1016/j.gdata.2015.06.018

Table 1
Genome assembly and annotation of *M. oryzae* strains, B157 and MG01.

	De novo based assembly		*Reference* based assembly	
Isolate name	B157	MG01	B157	MG01
Illumina reads (millions)	35.83	38.81	35.83	38.81
Coverage (X)	89	97	89	97
No. of contigs	6815	7722	5364	6046
No. of scaffolds	3534	7303	2508	3060
Largest scaffold size (Kb)	649	324	489	292
N50 (Kb)	91.16	55.03	92.4	54.6
Assembly size (Mb)	38	40	41	43
% repeats	3.01	3.23	10.4	10.39
No. of predicted genes with Augustus	11,340	11,744	12,535	13,135
No. of predicted genes (>200 bp)	–	–	11,334	11,733

2. Results

2.1. Genome assembly

The *M. oryzae* strain B157 was isolated during 1990s from Maruteru near Hyderabad, India and is a widely used standard strain in our laboratory experiments [9,10]. MG01 strain was isolated in 2012 from Karnataka in India. The genomes of these field isolates were initially assembled using both *de novo* and reference approaches using Velvet algorithm. However, reference based assembly was preferred for further analysis since it yielded better assembly quality and resolution of repeat elements. The genome assembly and annotation statistics are summarized in Table 1. Iterative mapping, contig ordering and scaffolding further improved the quality of reference-based assembly (Supplementary Tables S1 and S2). The reference-guided assembly analysis of B157 genome yielded 2508 scaffolds, N50 of 92 Kb and the largest scaffold of 489 Kb (Table 1). The reference-based assembly for MG01 comprised of 3060 scaffolds with N50 of 55 Kb and largest scaffold of 292 Kb. The genome size of 41 Mb and 43 Mb for B157 and MG01, respectively, was obtained from reference based assembly (Table 1).

2.2. Analysis of repetitive DNA sequences

Using *de novo* assembly, we obtained 3.01 Mb and 3.23 Mb of repeat elements for B157 and MG01, respectively (Table 1). Similar repeat content has been reported for other *Magnaporthe* isolates using *de novo* approaches [8]. However, the overall detection of repeat sequences *in silico* was dramatically increased in the reference-guided assembly, 4 Mb (10.4%) for B157 and 4.5 Mb (10.39%) for MG01 (Table 1). This repeat percentage is comparable to the first draft of *Magnaporthe* genome [4]. The predicted repeats in B157 and MG01 are largely composed of retrotransposons such as Pot2, MGR583, MAGGY and Pyret (Fig. 1). Pot2 copy number was higher in B157 (366 copies) and MG01 (396 copies) as compared to the laboratory strain 70-15 (272 copies). In contrast, Mg-SINE copy number is higher in 70-15 (172 copies) as compared to Indian isolates, B157 (65 copies) and MG01 (97 copies).

2.3. Gene prediction and annotation

The *de novo* gene prediction from Augustus tool resulted in 11,340 and 11,744 genes in B157 and MG01, respectively. After discarding short coding sequences (<200 bp), we obtained 11,334 and 11,733 genes in the genomes of B157 and MG01, respectively (Table 1; Supplementary Tables S3 and S4). The average gene length was 1400 bp and protein length was 480 amino acids. To further validate *de novo* gene prediction, a total of 87,947 ESTs from NCBI were mapped to the coding sequences of 70-15, B157 and MG01 genes. A total of 64.8%, 63.1% and 63.1% ESTs from NCBI were mapped to 52.6%, 56.3% and 54.3% of coding sequence in 70-15, B157 and MG01 strains of *M. oryzae*, respectively, provided validation of *de novo* gene prediction. Out of the mapped ESTs, 20%, 18.8% and 19.1% coding sequences were mapped with one EST, while 80%, 81.2% and 80.9% of coding sequences were mapped by two or more ESTs in 70-15, B157 and MG01, respectively.

2.4. Strain specific genes in Magnaporthe

Annotated genes of B157 and MG01 were aligned against 12,827 genes of reference strain 70-15 genome using BLAST program. No BLAST alignment was found for 489 and 596 genes of B157 and MG01, respectively. These genes from B157 and MG01 were further aligned against the reference genome of 70-15 using FASTA tool [11]. This analysis resulted in 54, 73 and 134 isolate specific genes in B157, MG01 and 70-15, respectively. B157 and MG01 specific genes were compared with recently sequenced *Magnaporthe* field isolates, Y34 and P131 [6]. This comparison has resulted in 17 isolates specific genes in B157 and 24 genes specific to MG01 (Fig. 2; Supplementary Tables S5 and S6). About 44% of novel genes from MG01 were found to have expression evidence and functional annotation.

2.5. Genome-wide variation in M. oryzae

There were 8650 and 10,797 ICVs in B157 and MG01, respectively (Fig. 3). We also observed that a large number of these variations occurred on chromosome 1, 2, 6 and 7. All possible inter-chromosomal translocations are shown in the innermost circle of Circos [12] (Fig. 3). All translocated regions were screened for *M. oryzae* repeat elements (Fig. 3). About 51.6% and 54.59% of rearrangements present across the seven chromosomes of B157 and MG01 genomes were due to repeats. These repeat elements were further analysed and found to be mainly composed of retroelements (40% in B157 and 39.53% in MG01) and DNA transposable elements (10.85% in B157 and 14.56% in MG01).

We used *M. oryzae* reference genome 70-15 to identify SNPs and short InDels in the genomes of Indian isolates, B157 and MG01. We identified 11,736 SNPs in B157 and 14,117 SNPs in MG01 genomes. In addition, we have also obtained 2.32 and 2.43 ratio of transitions (A ↔ G or C ↔ T) to transversions (A ↔ C or G ↔ T) Ts/Tv for B157 and MG01, respectively (Supplementary Fig. S4).

In comparison to the reference *Magnaporthe* strain (70-15), we have obtained SNPs for genic regions, UTRs, coding regions, introns, and intergenic regions (Supplementary Fig. S2; Supplementary Tables S7 and S8). We also screened synonymous and non-synonymous amino acid substitutions for the coding regions of annotated genes. This analysis revealed non-synonymous SNPs at exonic regions of 2423 and 2948 genes for B157 and MG01, respectively. In addition, SNPs in 151 genes from B157 and 155 genes from MG01 were found to have changes at the start and stop codons (gained and lost). Around 6260 and 5695 InDels were identified in B157 and MG01, respectively, by comparing with the genome of 70-15 (Supplementary Fig. S3, Tables S9 and S10).

2.6. Analysis of host specificity factors and avirulence genes in Magnaporthe

We compared host specificity factors and Avr genes in all sequenced *M. oryzae* strains including 70-15, B157 and MG01 (Table 2). The host specificity factor PWL1 is absent in all sequenced isolates expect MG01 whereas PWL2 is absent in B157, P131 and 4091-59-8. PWL3 and PWL4 are present in all isolates of *M. oryzae* except Indian isolates, B157 and MG01 (Table 2). The Avirulence gene, Avr-CO39 is absent in all strains except 4091-59-8 (Table 2). The avirulence factor (Avr-ACE1) belongs to PKS/NRPS family, which is present in all sequenced strains of *Magnaporthe*. We validated *in silico* analyzed host specificity factors and Avr genes by PCR amplification using gene-specific primers (Supplementary Table S14). This validation confirmed

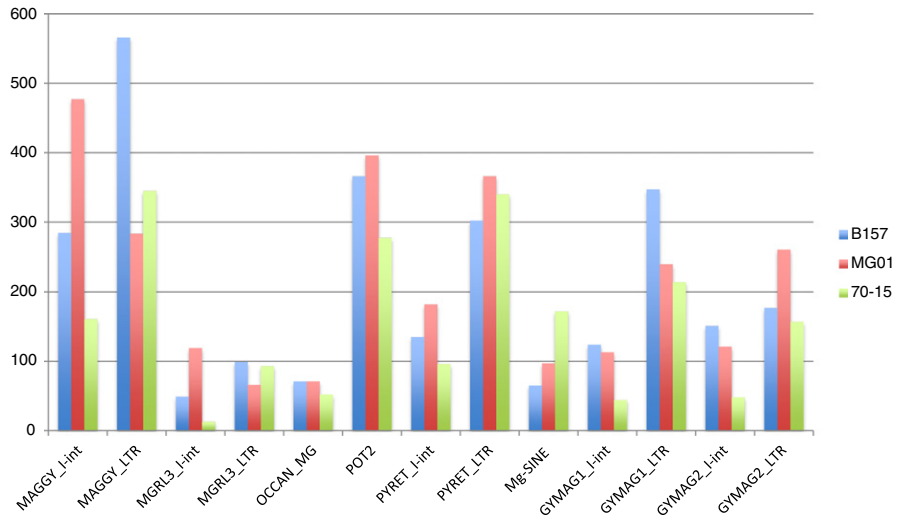

Fig. 1. Major repeat elements in the genomes of *M. oryzae* strains B157, MG01 and 70-15.

the accuracy of prediction of the presence and absence of the aforementioned Avr genes in B157 and MG01 (Supplementary Fig. S5).

2.7. Strand specific RNA-seq analysis

The strand specific RNA-seq analysis revealed 8338 (71%) genes expressed in MG01 (Fig. 4A, Supplementary Table S11). Expressed transcripts were further classified into sense [7405 genes (89%); Fig. 4B], antisense [338 genes (4.05%); Fig. 4C] and sense/antisense [595 genes (7.15%); Fig. 4D]. Each of the main classes was sub-categorized as low (14.04%), medium (72.27%) and highly (13.68%) expressed genes based on the FPKM values. We identified 835 genes that potentially generate overlapping transcripts of which 575 genes have overlapping sense and antisense transcripts (Table 3). To validate our ssRNA-seq results, we have used MPSS and RL-SAGE data from previous study [13]. This validation analysis showed, 17% of genes (58 out of 338 genes) have evidence for antisense transcription from MPSS and RL-SAGE data and 12% of genes (75 out of 595 genes) have MPSS evidence for producing sense and antisense transcripts (Table 3).

2.8. Comparative analyses in Ascomycota fungi

We compared the core set of *Magnaporthe* proteins (10,778 genes) with all sequenced non-pathogenic Ascomycetes fungi including *Neurospora crassa*, *Aspergillus niger*, *Aspergillus clavatus*, *Aspergillus oryzae*, *Aspergillus flavus*, *Aspergillus nidulans* and *Aspergillus terreus*. *Magnaporthe* protein sequences that have no homology in non-pathogenic fungi were scanned for Pfam domain. From this analysis, we identified a few pathogenicity genes; Enterotoxin A (PF01375), GSP synthase (PF03738) and YgcI-YcgG (PF08892), which are only present in *Magnaporthe* and absent in non-pathogenic Ascomycetous fungi such as *N. crassa*, *A. niger*, *A. clavatus*, *A. oryzae*, *A. flavus*, *A. nidulans* and *A. terreus*. Enterotoxin A gene cluster consists of seven genes in 70-15 and five genes in other sequenced strains including B157, MG01, P131, Y34, KJ201, 4091-5-8, FJ81278 and HN19311 (Table 4). Among these gene clusters, enterotoxin A cluster (MGG_05465, MGG_00390, MGG_16989) was expressed in mycelial stage in MG01 isolate (Table 4). However, other plant pathogenic fungi were found to have a single copy of enterotoxin A gene including *Phaeosphaeria nodorum*, *Glomerella graminicola*, *Colletotrichum higginsianum* and *Verticillium dahliae* (Table 4). We also looked for enterotoxin A domain containing genes in non-pathogenic Ascomycetous fungi. We were unable to find any homologs for enterotoxin A gene in non-pathogenic fungi including *N. crassa*, *A. niger*, *A. clavatus*, *A. oryzae*, *A. flavus*, *A. nidulans* and *A. terreus.*

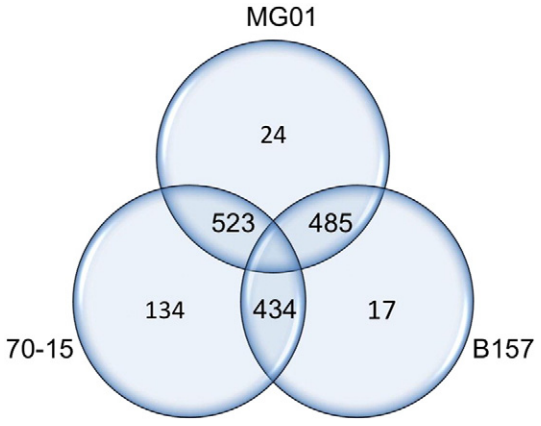

Fig. 2. Isolate specific genes from B157 and MG01 based on the reference (70-15) genome assembly.

3. Discussion

The comparative analysis of pathogenic field isolates of *Magnaporthe* from different locations of the world will help in understanding the fungal virulence spectrum. With advent of next generation sequencing technologies, now it is possible to sequence multiple genomes of *Magnaporthe* species. In this study we sequenced two field isolates (B157 and MG01) of *Magnaporthe* from different regions of southern India. The genome size for B157 and MG01 was slightly larger than 70-15. Over 50% coding sequences in B157 and MG01 strains have EST evidences. To understand genomic variations and isolate specific gene content we compared the predicted gene sets of *Magnaporthe* isolates. Comparisons of the genes of two field isolates with 70-15 and two recently published isolates revealed 54 and 73 isolates specific genes in B157 and MG01, respectively, as compared to 70-15. There were 17 and 24 isolate specific genes in B157 and MG01 when compared with Y34 and P131 [6], respectively. Majority of the isolate specific genes did not have annotations, however 47% of isolate specific genes from MG01 showed the RNA seq expression evidence.

In general, Avr-ACE1 and Avr-Pizt genes were having orthologs in all the sequenced isolates. We surveyed for SNPs and InDels in the coding regions of Avr genes in the sequenced genomes of the rice blast isolates.

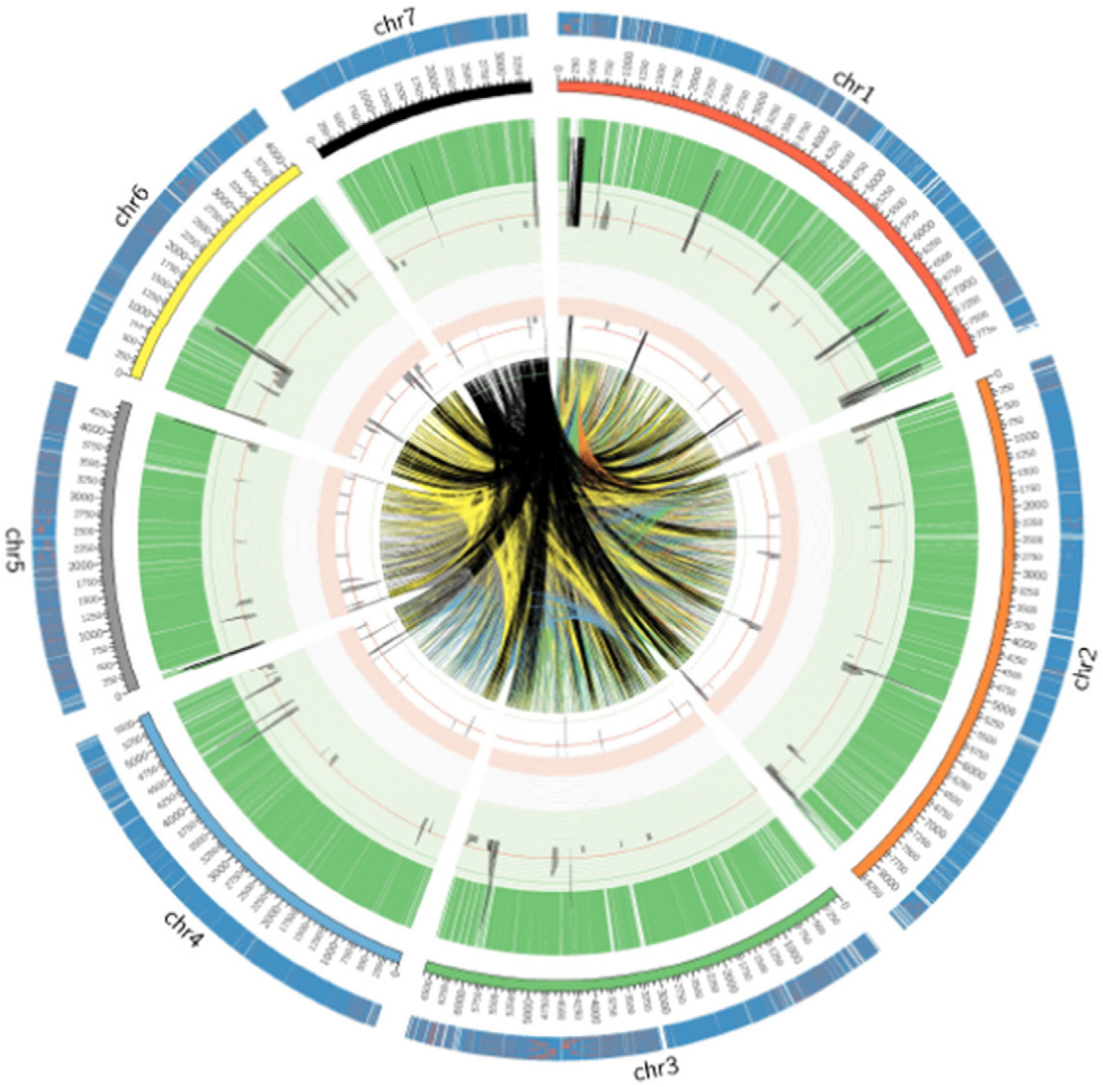

Fig. 3. Distribution of genes, SNPs and inter-chromosomal variations in Indian isolates (B157 and MG01) in comparison with 70-15 reference genome. The outermost circle shows the distribution of genes in MG01, and the second circle shows seven chromosomes of 70-15 with their sizes in 1 Kb intervals. The third circle shows the distribution of genes in B157, and the fourth and fifth circles show SNP distribution for MG01 and B157, respectively at 1 Kb intervals. The high SNP density regions are marked as peaks. The innermost circle shows possible translocation patterns in both B157 and MG01.

However, we could not find any non-synonymous mutation in any of the cloned host specificity and Avr genes. This indicates that these genes are likely to be important for survival of the blast fungus in nature. Interestingly MG01 genome contains more host specificity factors and Avr genes (PWL1, PWL2, Avr-Pita, Avr ACE1, Avr-Pizt and Avr-Pia) as compared to B157 (Avr-ACE1, AvrPiz-t and Avr-Pik) (Table 2).

Repetitive elements have played an important role in *Magnaporthe* genome evolution. Although individual repeat classes varied in copy

Table 2
Host specificity factors and avirulence (Avr) genes in sequenced strains of *M. oryzae*.

M. oryzae genes	*M. oryzae* strains								
	B157	MG01	70-15	*HN19311*	*FJ81278*	Y34	P131	KJ201	4091-59-8
Host specificity factors									
PWL1	−	+	−	−	−	−	−	−	−
PWL2	−	+	+	+	+	+	−	+	−
PWL3	−	−	+	+	+	+	+	+	+
PWL4	−	−	+	+	+	+	+	+	+
Avirulence (Avr) genes									
Avr-CO39	−	−	−	−	−	−	−	−	+
Avr-Piz-t	+	+	+	+	+	+	+	+	+
Avr-Pia	−	+	−	−	−	+	+	−	−
Avr-Pii	−	−	−	−	−	−	−	−	−
Avr-Pik	+	−	+	+	+	+	−	+	−
Avr-Pita	−	+	+	+	+	+	+	+	+
Avr ACE1	+	+	+	+	+	+	+	+	+

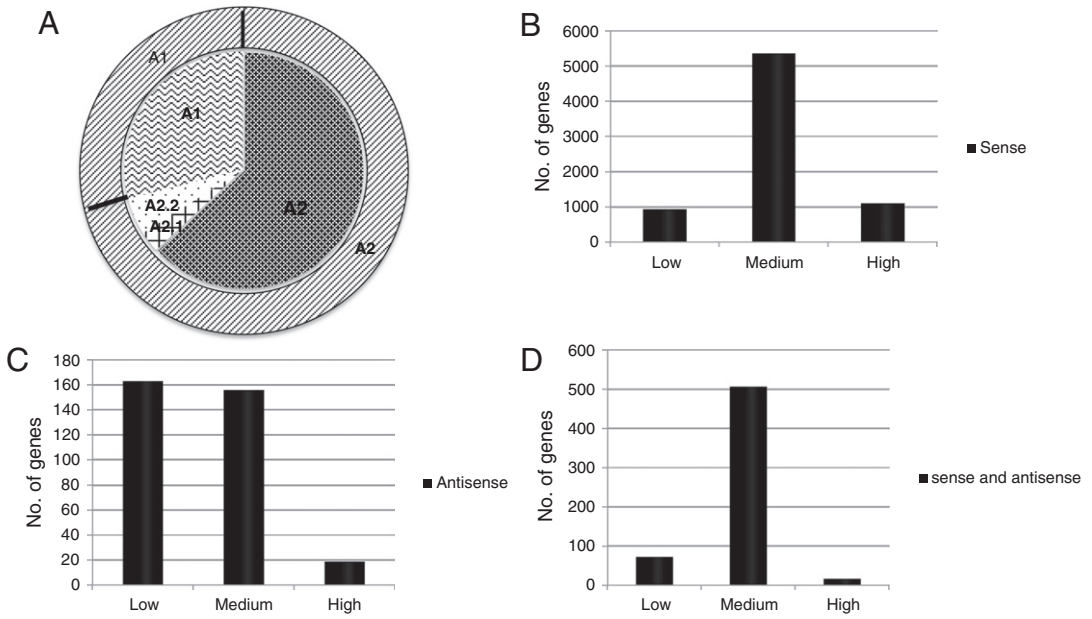

Fig. 4. Sense and antisense transcript expression support for annotated genes of MG01. A, the overall distribution of genes validated by strand-specific RNA-sequence data. A1, number of genes with no RNA-seq evidence; A2, number of genes with RNA-seq evidence. A2 is subdivided into sense (A2.1), antisense (A2.2) and both sense and antisense (A2.3). All expressed genes were classified based on their strand being expressed into three main classes as sense (B), antisense (C) or both (D) (sense and antisense). Each main class was subcategorized into three subclasses based on the FPKM value (Low, medium and high). The FPKM value was categorized as low (if FPKM ≤10), medium (if FPKM >10 to ≤200) or high (>200).

numbers between these two isolates and overall distribution of repeat content remain similar (Fig. 1; Table 1). The percentage of majority of repeats elements like MAGGY, PYRET and POT2 was slightly higher in the genomes of these isolates. Copy number of Mg-SINE was found to be higher in the 70-15 genome as compared to field isolates (Fig. 1). A large number of genome-wide translocations were found to be associated with retrotransposons (40% in B157 and 39.53% in MG01) and transposons (11% in B157 and 15% in MG01). Absence of Avr-CO39 in rice isolates indicated that this gene must have been lost due to accumulation of mutations and transpositions over a period of time [14]. Insertion of repeats in Avr genes has been shown to be associated with gain of virulence. It has been shown that insertion of Pot3 element in the promoter of Avr-Pita has led to gain in virulence towards rice varieties carrying R gene, Pi-ta [15]. Thus analysis of repeat elements and their insertion sites in the genome will help to clarify the emergence of virulent *Magnaporthe* strains.

We hypothesize that the genomic differences among fungal isolates may be due to variation in environmental factors, host factors, mating types, and other micro variations such as SNPs and InDels, chromosomal abnormalities and repetitive DNA elements [4,7,8]. We could observe about one SNP per 3448 nts and 2866 nts in the genomes of B157 and MG01, respectively. We identified 11,736 SNPs in B157 and 14,117 SNPs in MG01 genomes. Non-synonymous mutations were much higher as compared to synonymous substitutions in MG01 and B157 isolates (Supplementary Tables S6 and S7). Among these, 151 genes from B157 and 155 genes from MG01 have gained start or stop codons due to non-synonymous SNPs, which will have a significant effect on gene function. The larger genome size, higher percentage of repeat elements and non-synonymous mutations indicate the adaptive evolution of these isolates under field conditions.

Natural antisense transcripts have been identified in several fungi including *Saccharomyces cerevisiae* [16], *Candida albicans* [17], *A. flavus* [18], *Cryptococcus neoformans* [19] and *Ustilago maydis* [20,21]. However, natural antisense transcripts and their role in pathogenicity have not been studied in fungal pathogens. For this purpose, we adopted strand orientation based RNA sequencing using Illumina HiSeq chemistry. In this study we identified sense and antisense transcripts in *Magnaporthe* isolate, MG01 for several genes including MGG_09706 (aerobactin siderophore biosynthesis protein; Supplementary Fig. S6A) and MGG_07705 (acyl-CoA ligase; Supplementary Fig. S6B), and few pathogenicity genes like beta-glucanase (MGG_06493), serine/threonine protein kinase (MGG_03207) and endoglucanase-4 (MGG_08020). This is the first report of cataloguing antisense transcription across *Magnaporthe* genome, which will shed light on role of antisense gene regulation in blast disease biology.

We compared *Magnaporthe* genomes with non-pathogenic Ascomycetes fungi (*N. crassa*, *A. niger*, *A. clavatus*, *A. oryzae*, *A. flavus*, *A. nidulans* and *A. terreus*). This comparison revealed the exclusive

Table 3
Strand specific RNA (ssRNA) seq analysis for MG01 strain.

Gene feature	B157	MG01
Total no. of genes predicted (>200 bp)	11,334	11,733
No. of genes having RNA-seq evidence	8672	8338
a) No. of genes having sense expression	–	7404
b) No. of genes having antisense expression	–	338 (58[a])
c) No. of genes having sense and antisense expression	–	595 (75[b])
No. of genes with no RNA-seq evidence	2662	3395

[a] Genes are confirmed by MPSS and RL-SAGE data [13].
[b] Genes confirmed by MPSS data [13].

Table 4
Enterotoxin-A domain encoding genes in pathogenic fungi including the sequenced *M. oryzae* strains.

Pathogenic fungi	Host	No. of genes
P. nodorum	Plant	1
G. graminicola	Plant	1
C. higginsianum	Plant	1
V. dahliae	Plant	1
M. oryzae strains		
70-15	Plant	7
B157, *MG01*, *P131*, *Y34*, *FJ81278*, *HN19311*, *4091-5-8*, *KJ201*		5
M. acridum	Animal	7
M. anisopliae	Animal	16
C. militaris	Animal	4

presence of pathogenicity genes like Enterotoxin A (PF01375), GSP synthase (PF03738) and YgcI-YcgG (PF08892) in *Magnaporthe* isolates. Other genes including transcriptional regulators like MGG_12865 (HOX7), MGG_11346 (CDTF1) and MGG_07218 (TF) were found only in *Magnaporthe* and absent in all non-pathogenic Ascomycetes fungi such as *N. crassa*, *A. niger*, *A. clavatus*, *A. oryzae*, *A. flavus*, *A. nidulans* and *A. terreus* (Supplementary Table S12). HOX family is the member of homeobox transcription factors, which are stage specific regulators in *Magnaporthe* developmental process and HOX7 is specifically involved in appressoria formation [22]. CDTF1 is a transcriptional regulator involved in appressoria development [23] and MGG_07218 is hypothetical protein involved in host pathogenicity [24]. Interestingly, the homologs of these genes were identified in other pathogenic fungi including *Fusarium*, *Colletotrichum*, *Verticillium* and *Gaeumannomyces graminis*, *V. dahliae*, *Glomerella cingulata*, *Colletotrichum graminicola* and *C. higginsianum*. The PTH11 (MGG_0587), a trans-membrane protein is absent in majority of non-pathogenic Ascomycetes except *A. clavatus* (Supplementary Table S13). PTH1 is known to involve in appressoria development in *M. oryzae* [25]. The Con7p (MGG_05287) is a transcription factor necessary for appressoria development and *in planta* establishment of *Magnaporthe* [26] and this gene is absent in *A. clavatus*, *A. oryzae* and *A. flavus* (Supplementary Table S13). Tps1 affects virulence-associated gene expression by modulation of a set of NADP-dependent GATA factor/Nmr transcriptional regulators (NMR1: MGG_10017, NMR2: MGG_02860 and NMR3: MGG_09705), which are implicated in both fungal development and pathogenicity [27]. NMR genes are absent in non-pathogens including *N. crassa*, *A. niger*, *A. clavatus* and *A. terreus* (Supplementary Table S13). The exclusive presence of pathogenic gene clusters in *Magnaporthe* indicates the selective retention of these genes in pathogenic strains and loss in non-pathogenic Ascomycetes counterparts as previously reported by [28] range of pathogenic Ascomycetes fungi.

In summary, we have sequenced and analyzed genomes of two field isolates of *Magnaporthe* from Indian subcontinent where rice is grown in large areas. The incidence of rice blast disease is also very high in India. The variability of host specificity genes, avirulence genes and other pathogenicity genes from this study will be valuable resource for functional genomic studies in *M. oryzae*. The availability of genomes of host plant (HR12; NCBI accession number AZTA00000000) and its corresponding virulent *M. oryzae* isolates (B157 and MG01) from India will accelerate study of host-pathogen interaction and development of strategies for resistance breeding in rice.

4. Materials and methods

4.1. M. oryzae field isolates

Monoconidial isolation approach was followed to obtain pure cultures of B157 and MG01. These strains were isolated from the rice cultivar HR12 belonging to *indica* subspecies. HR12 is widely used blast susceptible check cultivar in India, since it is extremely sensitive to blast pathogen. Oatmeal agar medium was used for growth and maintenance of *M. oryzae* isolates. These cultures were stored on filter paper disks at −20 °C for long-term storage.

4.2. Nucleic acid isolation

M. oryzae strains were grown in liquid YEG medium (Glucose 1 g, yeast extract 0.2 g in 100 ml of distilled water) for 3-days in dark at 200 rpm at 28 °C. Mycelia were filtered and genomic DNA was extracted using GenElute plant genomic Miniprep kit (Sigma Cat. No. G2N70-1KT). Total RNA was isolated from 3-days old mycelia on YEG liquid medium from MG01 using TRIzol method [13].

4.3. DNA library construction and Illumina sequencing

Libraries were prepared with 1 μg of DNA using TrueSeq DNA sample preparation kit (Illumina Cat. No. FC-121-2001). The genomic DNA (1 μg) was fragmented using ultrasonicator (S220: Covaris, USA) to obtain an average of 350 bp fragments. This was followed by end repair, A-tailing, ligation with Illumina adapters, size selection and PCR amplification. The prepared libraries were quantified using Bioanalyzer and quantitative PCR (qPCR). The clusters were generated using cBOT and paired-end sequencing was carried out with Illumina HiSeq 1000 instrument at Center for Cellular and Molecular Platforms (CCAMP), Bangalore, India.

4.4. Genome assembly

Illumina paired-end reads were quality filtered using FastX tool kit (version 0.0.13.2). Adapter sequences were clipped using Cutadapt version 1.2.1 [29]. Then paired reads having at least 80% of bases with quality score greater than Q30 (Q score is quality score specified by Illumina, which indicates probability of errors in base calling. Q30 means a probability of incorrect base call is in 1 in 1000) were chosen for further analysis. We attempted both *de novo* and *reference* based assembly of genomes using Velvet 1.2.09, however reference based assembly was used for further analysis since it yielded better assembly [30]. *M. oryzae* 70-15 was used as a reference strain for reference based assembly. The whole genome assembly is available at NCBI/DDBJ/EMBL with the accession AXDJ01000000 for B157 and AYPX01000000 for MG01.

Contig ordering, gap filling and re-scaffolding was performed using various integrated tools in order to improve assembly quality. We used the ABACAS tool for contig ordering with reference [31]. The Iterative Mapping and Assembly for Gap Elimination (IMAGE) [32] method was used to fill the gaps in the assembly. The pre-assembled contigs were merged back to scaffolds after successful completion of iterative assembly using SSPACE (SSAKE-based Scaffolding of Pre-Assembled Contigs after Extension) [33].

4.5. Gene prediction and annotation

Genes were predicted using Augustus [34] from reference-based assembly of B157 and MG01 strains of *M. oryzae*. The predicted genes were aligned to 70-15 protein sequence using BLAST based homology with e-value cutoff of $\leq 10^{-5}$. To validate the predicted genes by Augustus, ESTs of *M. oryzae* were mapped onto 70-15, B157 and MG01. EST data sets were downloaded from NCBI and mapped onto coding sequences of predicted genes of B157 and MG01.

De novo gene prediction was performed using Augustus [34] from Indian isolates, B157 and MG01 whereas 70-15 genes were predicted by FGENESH [35]. Thus, in order to identify isolate specific genes, FASTA v36 program [11] was used for comparison of genes to genome to extract isolate specific genes. Annot8r tool [36] was used for annotation of genes predicted by Augustus.

4.6. Repetitive DNA prediction

Repeat detection and masking was carried out using RepeatMasker 4.0.2 tool [37] *Magnaporthe* species repeat library was used for repeat prediction. The predicted repeats were further classified into major classes/families of elements.

4.7. Variant analysis

In order to detect possible inter chromosomal variations (ICVs), sequencing reads were mapped against the 70-15 reference genome. Anomalous read pairs (ARPs) (mapping distance between the read pairs is beyond the actual sequencing library insert size) were extracted

from mapping file. To avoid any false positive hit, we removed duplicate reads prior to structural variation detection. ARPs mapping to different chromosomes were extracted and checked for ICVs. Alignment obtained from reference mapping tool was used for short variants detection using SAM tool version 0.1.19 [38]. Both single nucleotide variations (SNVs) and short InDels were detected for B157 and MG01 in comparison with reference sequence of 70-15. Predicted variants were filtered based on mapping quality greater than 25 read, read depths greater than 10 and strand level evidence (at least one read from both the direction). These variants were further annotated using SnpEff tool [39] based on their chromosomal location and biological effects such as synonymous/non-synonymous SNPs, upstream/downstream, UTRs, intergenic etc. Also the transition and transversion (Ts/Tv) ratio was calculated for single nucleotide variations.

4.8. Analysis of host specificity and Avr genes in M. oryzae

Nucleotide sequences for host specificity genes [PWL1 (U36923.1), PWL2 (U26313.1), PWL3 (1045533), PWL4 (1045535)] and avirulence (Avr) genes [AvrPita (12642087), Avr-CO39 (27450408), AvrACE1 (47109413), AvrPiz-t (194293523), Avr-Pia (237858322), Avr-Pii (237858324) and Avr-Pik (237858326)] were downloaded from NCBI database. BLASTn was performed to identify Avr genes in the genomes of *M. oryzae* strains, *N. crassa* and *A. niger*. Genomic sequence of other *M. oryzae* strains including 70-15, Y34 (AHZS00000000), P131 (AHZT00000000), FJ81278 (ATNU00000000) and HN19311 (ATNT00000000) were downloaded from NCBI database.

4.9. PCR validation of host specificity and Avr genes

PCR primers were designed for 5′ and 3′ untranslated region (UTR) of AVR genes (Supplementary Table S14). Ten milliliters PCR reaction was set up with, 20 ng of genomic DNA, 1 μl of 10× buffer, 0.4 μl of 20 mM of dNTP mix, 0.5 μl of 10 mM of each forward and reverse primer, 0.75 Units of Dream Taq DNA polymerase (Fermentas, Cat. No EP0712, PA, USA). PCR program with initial denaturation at 95 °C for 5 min, 30 cycles of 95 °C for 15 s, Ta (variable as per each primer) for 30 s, 72 °C for 1 min and final extension at 72 °C for 5 min was set up for 30 PCR cycles. PCR product was separated on 2% agarose gels.

4.10. Strand specific RNA (ssRNA) library preparation and sequencing

About 1μg of total RNA from MG01 strain was used for mRNA purification using Illumina TruSeq RNA sample preparation kit v2 (Cat. No. RS-122-2001). mRNA was fragmented (100–150 bases) by chemical method. The first strand cDNA was synthesized using dNTPs and followed by second strand cDNA synthesis using dUTPs along with dNTPs, random hexamers and reverse transcriptase. Ampure beads were used to select double stranded cDNA template, end repaired, adenylated by adding single 'A' base to end of the cDNA and adapters were ligated. Fifteen cycles of PCR was performed to enrich the library. Bioanalyzer was used to validate RNA library and paired end sequencing (2 × 100 cycles) was performed using Illumina HiSeq 1000 at C-CAMP, Bangalore, India.

4.11. Strand specific RNA seq (ssRNAseq) data analysis

Genes predicted from Augustus pipeline was used for expression validation. The gene sequence was indexed using Bowtie2 [40] tool. We used about 35 million paired ssRNAseq reads (2 × 100 nts) from mycelial tissue of MG01 strain. The spliced read mapper TopHat2 [41] was used to map ssRNA-seq reads to the indexed gene set. The SAM file (TopHat2 output) was processed further to remove reads with secondary alignments. Uniquely aligned reads were assembled into transcripts and estimated relative abundance of transcripts for genes using Cufflinks [42]. The transcript abundance per gene was expressed in fragments per kilobase of exon model per million mapped fragments (FPKM). The output of Cufflinks was processed manually to count the genes with sense and antisense expression. We used in-house Perl script to identify the genes for overlapping sense and antisense transcripts. Genes with overlapping transcripts from both sense and antisense strands are annotated using Annot8r programme.

4.12. Pathogenicity genes analyses in M. oryzae

Pathogenic (*M. oryzae*) and non-pathogenic Ascomycetous fungi were compared for pathogenicity genes using whole genome annotation data. Protein sequences of the reference stain 70-15 (assembly version MG8) and Indian strains, B157 and MG01 were compared to identify the core set of genes in *M. oryzae*. Protein sequences of *N. crassa* OR74A, *A. clavatus*, *A. oryzae*, *A. flavus*, *A. nidulans* and *A. terreus* were retrieved from Broad Institute fungal database (http://www.broadinstitute.org) and *A. niger* CBS 513.88 sequences retrieved from KEGG database (http://www.genome.jp/kegg). BLASTp analysis was performed for *M. oryzae* core proteins against the non-pathogenic Ascomycetous proteins. The significant homology (e-10^{-5} and greater than 55% query sequence length coverage) was used for homology alignment. Additionally, we also used a gradient of percent identity (I > 30%, I > 50% and I > 70%) to check the variation in gene numbers. These sequences were further classified into known, predicted and hypothetical proteins based on the information available at fungal database at Broad Institute (http://www.broadinstitute.org). Further, hypothetical proteins were scanned through Pfam domain database to identify conserved domains from *M. oryzae*, which are present only in pathogenic fungi and absent in non-pathogenic fungi.

M. oryzae genes that are involved in pathogenesis were obtained from PHIbase V-3.4 database [43]. Reciprocal BLASTp was performed to identify homologous genes in sequenced strains of *M. oryzae*, *N. crassa*, *A. niger*, *A. clavatus*, *A. oryzae*, *A. flavus*, *A. nidulans* and *A. terreus*. Genes were filtered based on percent identity and query coverage (PHIbase analysis parameter: identity ≥30% and query coverage ≥50%).

Supplementary data to this article can be found online at http://dx.doi.org/10.1016/j.gdata.2015.06.018.

Acknowledgments

This work is supported by grants from the Department of Biotechnology, Govt. of India (BT/HRD/35/02/2006) to Malali Gowda (Ramalingaswami Fellowship) and Bharat B Chattoo. We thank Naweed Naqvi, Temasek Life Sciences Laboratory, Singapore for critical reading of this manuscript. We also acknowledge Council of Scientific & Industrial Research (CSIR) fellowship awarded to Meghana Deepak Shirke by Government of India. We thank Russiachand, Ramya LR, Chandana S and Shilpa S from Genomics Facility at CCAMP for their help in sequencing of *Magnaporthe* isolates. We acknowledge Genomics facility (BT/PR3481/INF/22/140/2011) at Centre for Cellular and Molecular Platforms, Bengaluru for sequencing of *Magnaporthe* isolates.

References

[1] M. Kohli, Y. Mehta, E. Guzman, L. Viedma, L. Cubilla, Pyricularia blast—a threat to wheat cultivation. Czech J. Genet. Plant Breed. 47 (2011) S130–S134.
[2] S. Marcel, R. Sawers, E. Oakeley, H. Angliker, U. Paszkowski, Tissue-adapted invasion strategies of the rice blast fungus *Magnaporthe oryzae*. Plant Cell Online 22 (2010) 3177–3187.
[3] A. Sesma, A.E. Osbourn, The rice leaf blast pathogen undergoes developmental processes typical of root-infecting fungi. Nature 431 (2004) 582–586.
[4] R.A. Dean, N.J. Talbot, D.J. Ebbole, M.L. Farman, T.K. Mitchell, M.J. Orbach, M. Thon, R. Kulkarni, J.-R. Xu, H. Pan, The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434 (2005) 980–986.
[5] H. Leung, E.S. Borromeo, M.A. Bernardo, J.L. Notteghem, Genetic analysis of virulence in the rice blast fungus *Magnaporthe grisea*. Phytopathology 78 (1988) 1227–1233.
[6] M. Xue, J. Yang, Z. Li, S. Hu, N. Yao, R.A. Dean, W. Zhao, M. Shen, H. Zhang, C. Li, Comparative analysis of the genomes of two field isolates of the rice blast fungus *Magnaporthe oryzae*. PLoS Genet. 8 (2012) e1002869.

[7] K. Yoshida, H. Saitoh, S. Fujisawa, H. Kanzaki, H. Matsumura, K. Yoshida, Y. Tosa, I. Chuma, Y. Takano, J. Win, Association genetics reveals three novel avirulence genes from the rice blast fungal pathogen *Magnaporthe oryzae*. Plant Cell Online 21 (2009) 1573–1591.
[8] C. Chen, B. Lian, J. Hu, H. Zhai, X. Wang, R. Venu, E. Liu, Z. Wang, M. Chen, B. Wang, Genome comparison of two *Magnaporthe oryzae* field isolates reveals genome variations and potential virulence effectors. BMC Genomics 14 (2013) 887.
[9] P. Kachroo, S. Leong, B. Chattoo, Pot2, an inverted repeat transposon from the rice blast fungus *Magnaporthe grisea*. Mol. Gen. Genet. MGG 245 (1994) 339–348.
[10] P. Kachroo, S.A. Leong, B.B. Chattoo, Mg-SINE: a short interspersed nuclear element from the rice blast fungus. Proc. Natl. Acad. Sci. 92 (1995) 11125–11129.
[11] W.R. Pearson, D.J. Lipman, Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. 85 (1988) 2444–2448.
[12] M. Krzywinski, J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S.J. Jones, M.A. Marra, Circos: an information aesthetic for comparative genomics. Genome Res. 19 (2009) 1639–1645.
[13] M. Gowda, R. Venu, M.B. Raghupathy, K. Nobuta, H. Li, R. Wing, E. Stahlberg, S. Couglan, C.D. Haudenschild, R. Dean, Deep and comparative analysis of the mycelium and appressorium transcriptomes of *Magnaporthe grisea* using MPSS, RL-SAGE, and oligoarray methods. BMC Genomics 7 (2006) 310.
[14] Y. Zheng, W. Zheng, F. Lin, Y. Zhang, Y. Yi, B. Wang, G. Lu, Z. Wang, W. Wu, AVR1-CO39 is a predominant locus governing the broad avirulence of *Magnaporthe oryzae* 2539 on cultivated rice (*Oryza sativa* L.). Mol. Plant-Microbe Interact. 24 (2011) 13–17.
[15] S. Kang, M.H. Lebrun, L. Farrall, B. Valent, Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. Mol. Plant-Microbe Interact. 14 (2001) 671–674.
[16] M. Yassour, J. Pfiffner, J.Z. Levin, X. Adiconis, A. Gnirke, C. Nusbaum, D.-A. Thompson, N. Friedman, A. Regev, Strand-specific RNA sequencing reveals extensive regulated long antisense transcripts that are conserved across yeast species. Genome Biol. 11 (2010) R87.
[17] B.B. Tuch, Q.M. Mitrovich, O.R. Homann, A.D. Hernday, C.K. Monighetti, M. Francisco, A.D. Johnson, The transcriptomes of two heritable cell types illuminate the circuit governing their differentiation. PLoS Genet. 6 (2010) e1001070.
[18] C.A. Smith, D. Robertson, B. Yates, D.M. Nielsen, D. Brown, R.A. Dean, G.A. Payne, The effect of temperature on Natural Antisense Transcript (NAT) expression in *Aspergillus flavus*. Curr. Genet. 54 (2008) 241–269.
[19] B.J. Loftus, E. Fung, P. Roncaglia, D. Rowley, P. Amedeo, D. Bruno, J. Vamathevan, M. Miranda, I.J. Anderson, J.A. Fraser, The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science 307 (2005) 1321–1324.
[20] E.C. Ho, M.J. Cahill, B.J. Saville, Gene discovery and transcript analyses in the corn smut pathogen *Ustilago maydis*: expressed sequence tag and genome sequence comparison. BMC Genomics 8 (2007) 334.
[21] E.N. Morrison, M.E. Donaldson, B.J. Saville, Identification and analysis of genes expressed in the *Ustilago maydis* dikaryon: uncovering a novel class of pathogenesis genes. Can. J. Plant Pathol. 34 (2012) 417–435.
[22] S. Kim, S.-Y. Park, K.S. Kim, H.-S. Rho, M.-H. Chi, J. Choi, J. Park, S. Kong, J. Park, J. Goh, Homeobox transcription factors are required for conidiation and appressorium development in the rice blast fungus *Magnaporthe oryzae*. PLoS Genet. 5 (2009) e1000757.
[23] X. Yan, Y. Li, X. Yue, C. Wang, Y. Que, D. Kong, Z. Ma, N.J. Talbot, Z. Wang, Two novel transcriptional regulators are essential for infection-related morphogenesis and pathogenicity of the rice blast fungus *Magnaporthe oryzae*. PLoS Pathog. 7 (2011) e1002385.
[24] H. Zhang, K. Liu, X. Zhang, W. Tang, J. Wang, M. Guo, Q. Zhao, X. Zheng, P. Wang, Z. Zhang, Two phosphodiesterase genes, PDEL and PDEH, regulate development and pathogenicity by modulating intracellular cyclic AMP levels in *Magnaporthe oryzae*. PLoS One 6 (2011) e17241.
[25] T.M. DeZwaan, A.M. Carroll, B. Valent, J.A. Sweigard, *Magnaporthe grisea* pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues. Plant Cell Online 11 (1999) 2013–2030.
[26] D. Odenbach, B. Breth, E. Thines, R.W. Weber, H. Anke, A.J. Foster, The transcription factor Con7p is a central regulator of infection-related morphogenesis in the rice blast fungus *Magnaporthe grisea*. Mol. Microbiol. 64 (2007) 293–307.
[27] R.A. Wilson, R.P. Gibson, C.F. Quispe, J.A. Littlechild, N.J. Talbot, An NADPH-dependent genetic switch regulates plant infection by the rice blast fungus. Proc. Natl. Acad. Sci. 107 (2010) 21902–21907.
[28] A. Sánchez-Rodríguez, C. Martens, K. Engelen, Y. Van de Peer, K. Marchal, The potential for pathogenicity was present in the ancestor of the Ascomycete subphylum Pezizomycotina. BMC Evol. Biol. 10 (2010) 318.
[29] M. Martin, Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 17 (2011) 10–12.
[30] D.R. Zerbino, E. Birney, Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18 (2008) 821–829.
[31] S. Assefa, T.M. Keane, T.D. Otto, C. Newbold, M. Berriman, ABACAS: algorithm-based automatic contiguation of assembled sequences. Bioinformatics 25 (2009) 1968–1969.
[32] I.J. Tsai, T.D. Otto, M. Berriman, Method improving draft assemblies by iterative mapping and assembly of short reads to eliminate gaps. Genome Biol. 11 (2010) R41.
[33] M. Boetzer, C.V. Henkel, H.J. Jansen, D. Butler, W. Pirovano, Scaffolding preassembled contigs using SSPACE. Bioinformatics 27 (2011) 578–579.
[34] M. Stanke, B. Morgenstern, AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 33 (2005) W465–W467.
[35] A. Salamov, V. Solovyev, Fgenesh multiple gene prediction program. http://genomic.sanger.ac.uk/1998.
[36] R. Schmid, M.L. Blaxter, annot8r: GO, EC and KEGG annotation of EST datasets. BMC Bioinformatics 9 (2008) 180.
[37] A. Smit, R. Hubley, P. Green, RepeatMasker 4.0. Institute for Systems Biology, Seattle, WA, 2013.
[38] H. Li, B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin, The sequence alignment/map format and SAMtools. Bioinformatics 25 (2009) 2078–2079.
[39] P. Cingolani, A. Platts, L.L. Wang, M. Coon, T. Nguyen, L. Wang, S.J. Land, X. Lu, D.M. Ruden, 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 6 (2012) 80–92.
[40] B. Langmead, S.L. Salzberg, Fast gapped-read alignment with Bowtie 2. Nat. Methods 9 (2012) 357–359.
[41] D. Kim, G. Pertea, C. Trapnell, H. Pimentel, R. Kelley, S.L. Salzberg, TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 14 (2013) R36.
[42] C. Trapnell, A. Roberts, L. Goff, G. Pertea, D. Kim, D.R. Kelley, H. Pimentel, S.L. Salzberg, J.L. Rinn, L. Pachter, Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. 7 (2012) 562–578.
[43] R. Winnenburg, M. Urban, A. Beacham, T.K. Baldwin, S. Holland, M. Lindeberg, H. Hansen, C. Rawlings, K.E. Hammond-Kosack, J. Köhler, PHI-base update: additions to the pathogen–host interaction database. Nucleic Acids Res. 36 (2008) D572–D576.