Chapter 5

Results:

In silico analyses of secondary metabolite biosynthetic gene clusters in different host-specific lineages of the blast fungus

5.1 Identification of biosynthetic gene clusters (BGCs) in host-specific strains of *M. oryzae*

In this investigation, we employed a dataset comprising of 68 *M. oryzae* genomes, which were sourced from six distinct host plants, namely rice (*Oryza sativa*), finger millet *(Eleusine coracana*), foxtail millet *(Setaria sp.*), wheat (*Triticum aestivum*), perennial ryegrass *(Lolium sp.*) and weeping lovegrass (*Eragrostis curvula*) **(Fig. 5.1A, Table 5.1)**. This dataset encompasses an earlier sequenced 15 field isolates, collected from different parts of India **(Fig. 5.1A, Table 5.1)**. The remaining 53 genomes were obtained from publicly available genome sequences. Furthermore, we incorporated sequence data from three *M. grisea* strains, previously isolated from crabgrass (*Digitaria sp.),* to be used as control **(Table 5.1)**. Sixteen out of the publicly available assemblies were derived from long-read sequencing technology, resulting in highly contiguous datasets. Subsequently, gene predictions were performed on all 71 assemblies and analysis of BUSCO genes indicated robust gene prediction quality and assembly completeness, with a BUSCO score exceeding 90% **(Table 5.1)**.

Utilizing a set of 2655 BUSCO genes found within the 68 *M. oryzae* genomes, we constructed a phylogenomic tree. The resulting topology of the species tree confirms the existence of multiple genetic lineages within *M. oryzae,* each specialized on different host plants, including *Oryza, Setaria, Eleusine, Eragrostis, Triticum* and *Lolium* **(Fig. 5.1B)**. This phylogenetic arrangement aligns with the previously reported population structure (Gladieux, Condon, et al., 2018). A similar phylogenomic assessment was conducted using an additional three genomes from *M. grisea* (*Digitaria* isolates), reaffirming the divergence of *M. grisea* as a distinct species from *M. oryzae*, as in accordance with prior findings (**Fig. 5.2**; Gladieux et al., 2018).

Our methodology involved the utilization of a comprehensive pipeline to discern the biosynthetic diversity within various lineages of *M. oryzae* **(Fig. 5.3)**. This approach encompassed the initial identification of biosynthetic gene clusters (BGCs) responsible for secondary metabolite (SM) production within genomic regions, followed by a subsequent analysis employing similarity networks to explore the genetic variations among the predicted BGCs.

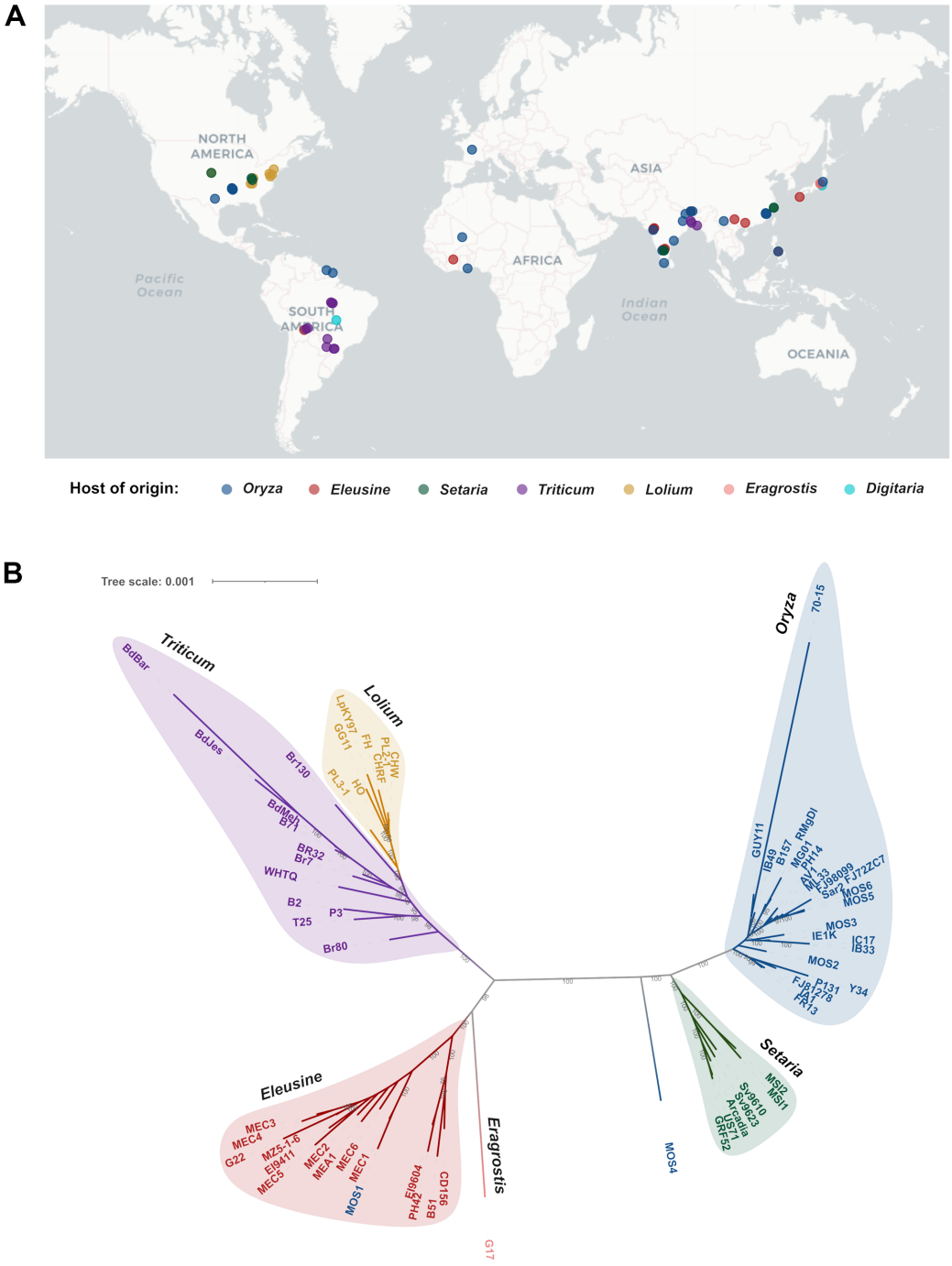

Figure 5.1: *M. oryzae* worldwide populations are organized in host-specific lineages. A) Geographic distribution of sequenced *M. oryzae* strains. The color-coded dots denote the host plant of origin of the strains. B) Phylogenetic tree constructed based on concatenation of 2655 BUSCO genes present in all 68 *M. oryzae* genomes used in this study. Colored shapes in the background depict different host-specific genetic lineages.

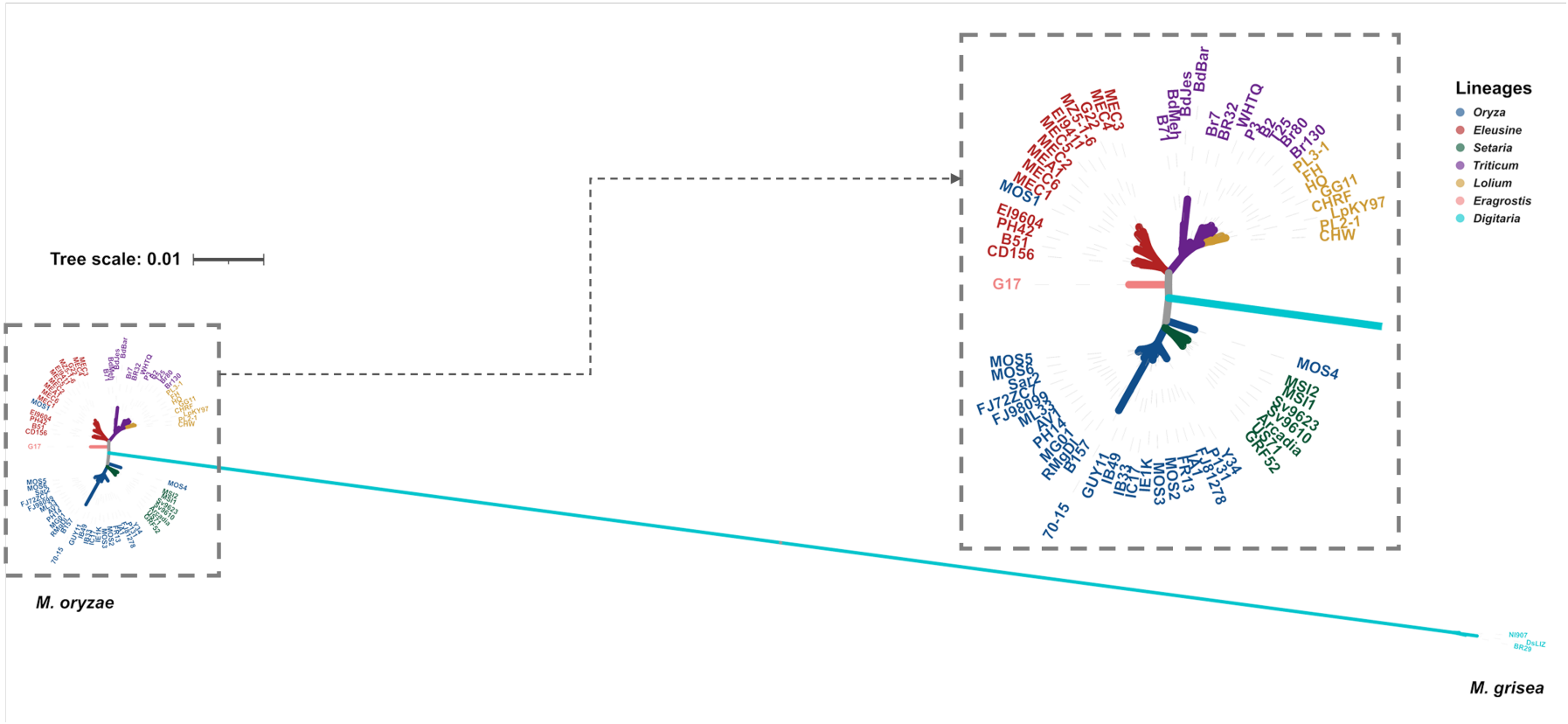

Figure 5.2: *M. oryzae* and *M. grisea* are evolutionarily distinct species adapted to different hosts. Maximum likelihood tree constructed based on concatenation of a total 2557 BUSCOs present in all 71 genomes of *M. oryzae* and *M. grisea* strains used in the study. Colored branches depict different host-specific genetic lineages of *M. oryzae*.

Table 5.1: *M. oryzae* and *M. grisea* genomes used in this study.

Strains	Synonyms	Host of Isolation	Country/Region	Year of Collection	Phylogenetic lineage	Assembly Accession IDs	References	BUSCO (%)
70-15		*Oryza sativa*	n/a	-	*Oryza*	GCF_000002495.2	Dean et al. 2005	98.6
Arcadia		*Setaria viridis*	Lexington, Kentucky, USA	1998	*Setaria*	GCA_002925445.1	Rahnama et al. 2021	95.7
AV1-1-1		*Oryza sativa*	Ghana: Aveyime	2015	*Oryza*	GCA_011799965.1 *	Zhong et al. 2020	97.1
B157		*Oryza sativa*	India: Maruteru, Andhra Pradesh	1989	*Oryza*	GCA_000832285.1	Gowda et al. 2015	95.8
B2		*Triticum aestivum*	Bolivia: Okinawa Uno	2011	*Triticum*	GCA_002218465.1	Rahnama et al. 2021	96.5
B51		*Eleusine indica*	Bolivia: Quirusillas	2012	*Eleusine*	GCA_002925415.1	Farman et al. 2017	96.7
B71		*Triticum aestivum*	Bolivia: Okinawa Uno	2012	*Triticum*	GCA_004785725.2 *	Peng et al. 2019	97.1
BdBar	BdBar16-1	*Triticum aestivum*	Bangladesh: Barisal	2016	*Triticum*	GCA_001675615.1	Rahnama et al. 2021	91.5
BdJes	BdJes16-1	*Triticum aestivum*	Bangladesh: Jessore district	2016	*Triticum*	GCA_001675595.1	Rahnama et al. 2021	93.7
BdMeh	BdMeh16-1	*Triticum aestivum*	Bangladesh: Mehepur district	2016	*Triticum*	GCA_001675605.1	Rahnama et al. 2021	96.9
BR0032	BR32	*Triticum aestivum*	Brazil	1991	*Triticum*	GCA_900474545.3 *	Langner et al. 2021	97
Br130		*Triticum aestivum*	Brazil: Mato Grosso do Sul	1990	*Triticum*	GCA_002925325.1	Rahnama et al. 2021	94.6
BR29		*Digitaria*	Brazil	-	*Digitaria*	BR29	Chiapello et al. 2015	97.2
Br7		*Triticum aestivum*	Brazil: Parana	1990	*Triticum*	GCA_002925335.1	Rahnama et al. 2021	97
Br80		*Triticum aestivum*	Brazil	1991	*Triticum*	GCA_002925345.1	Rahnama et al. 2021	96.9
CD156	CD0156	*Eleusine indica*	Ivory Coast, Ferkessedougou	1989	*Eleusine*	GCA_900474475.3 *	Langner et al. 2021	97.2
CHRF		*Lolium perenne*	Siler Springs, MD, USA	1996	*Lolium*	GCA_002925295.1	Rahnama et al. 2021	97.1

CHW		*Lolium perenne*	USA: Annapolis, MD	1996	*Lolium*	GCA_002925285.1	Rahnama et al. 2021	96.8
DsLIZ		*Digitaria*	Lexington, Kentucky, USA	2000	*Digitaria*	GCA_002925245.1 *	Rahnama et al. 2021	96
EI9411		*Eleusine indica*	China,Fujian	1994	*Eleusine*	GCA_001548775.1	Zhong et al. 2016	96.8
EI9604		*Eleusine indica*	China,Zhejiang	1996	*Eleusine*	GCA_001548785.1	Zhong et al. 2016	96.9
FH		*Lolium perenne*	USA: Hagerstown, MD	1997	*Lolium*	GCA_002925225.1	Rahnama et al. 2021	95.7
FJ72ZC7-77		*Oryza sativa*	China: Fujian	1992	*Oryza*	GCA_011799905.1 *	Zhong et al. 2020	96.9
FJ81278		*Oryza sativa*	China: Fujian	1981	*Oryza*	GCA_002368475.1 *	Bao et al. 2017	97
FJ98099		*Oryza sativa*	China: Fujian	1998	*Oryza*	GCA_011799925.1 *	Zhong et al. 2020	97.1
FR13	FR0013	*Oryza sativa*	France	1990	*Oryza*	GCA_900474655.3 *	Langner et al. 2021	97.1
G17	K76-79	*Eragrostis curvula*	Japan	1976	*Eragrostis*	GCA_002925205.1	Rahnama et al. 2021	96.9
G22	WGG-FA40	*Eleusine coracana*	Japan	1976	*Eleusine*	GCA_002925165.1	Gladieux et al. 2018	96.9
GG11		*Lolium perenne*	USA: Lexington, KY	1997	*Lolium*	GCA_002925155.1	Rahnama et al. 2021	96.9
GRF52		*Setaria viridis*	Lexington, Kentucky, USA	2001	*Setaria*	GCA_002925145.1	Rahnama et al. 2021	96.9
GY11	GY0011, Guy11	*Oryza sativa*	French - Guyana	1988	*Oryza*	GCA_002368485.1 *	Bao et al. 2017	97
HO		*Lolium perenne*	USA: Richmond, PA	1996	*Lolium*	GCA_002925105.1	Rahnama et al. 2021	97
IA1	ARB114	*Oryza sativa*	Arkansas, USA	2009	*Oryza*	GCA_002925085.1	Rahnama et al. 2021	97
IB33		*Oryza sativa*	Texas, USA	-	*Oryza*	GCA_002925065.1	Rahnama et al. 2021	97.1
IB49	ZN61	*Oryza sativa*	AR, USA	1992	*Oryza*	GCA_002925045.1	Rahnama et al. 2021	97.1
IC17	ZN57	*Oryza sativa*	AR, USA	1992	*Oryza*	GCA_002925025.1	Rahnama et al. 2021	97.1
IE1K	TM2	*Oryza sativa*	AR, USA	2003	*Oryza*	GCA_002924985.1	Rahnama et al. 2021	97.1
LpKY97	LpKY97-1	*Lolium perenne*	USA	1997	*Lolium*	GCA_012272995.1 *	Rahnama et al. 2021	96.6
MEA1		*Eleusine africana*	GKVK, Bengaluru, KA, India	2015	*Eleusine*		This study	96.9
MEC1		*Eleusine coracana*	Waghai, Dangs, GJ, India	2015	*Eleusine*		This study	96.9

MEC2		*Eleusine coracana*	Waghai, Dangs, GJ, India	2015	*Eleusine*		This study	96.9
MEC3		*Eleusine coracana*	GKVK, Bengaluru, KA, India	2015	*Eleusine*		This study	96.8
MEC4		*Eleusine coracana*	GKVK, Bengaluru, KA, India	2015	*Eleusine*		This study	96.8
MEC5		*Eleusine coracana*	Dentam, west sikkim, SK, India	2015	*Eleusine*		This study	96.4
MEC6		*Eleusine coracana*	South Sikkim, SK, India	2015	*Eleusine*		This study	96.8
MG01		*Oryza sativa*	India: Mandya	2011	*Oryza*	GCA_000969745.1	Gowda et al. 2015	95.4
ML33		*Oryza sativa*	Mali	1995	*Oryza*	GCA_002924965.1	Rahnama et al. 2021	97
MOS1		*Oryza sativa*	Waghai, Dangs, GJ, India	2013	*Eleusine*		This study	97
MOS2		*Oryza sativa*	Waghai, Dangs, GJ, India	2013	*Oryza*		This study	97
MOS3		*Oryza sativa*	Hazaribag, JH, India	2010	*Oryza*		This study	96.9
MOS4		*Oryza sativa*	Kanyakumari, TN, India	2015	*Oryza*		This study	97.1
MOS5		*Oryza sativa*	Upper Lingthem, North Sikkim, SK, India	2015	*Oryza*		This study	97
MOS6		*Oryza sativa*	Geysing, West Sikkim, SK, India	2015	*Oryza*		This study	96.9
MSI1		*Setaria Italica*	GKVK, Bengaluru, KA, India	2015	*Setaria*		This study	96.8
MSI2		*Setaria Italica*	GKVK, Bengaluru, KA, India	2015	*Setaria*		This study	96.7
MZ5-1-6		*Eleusine coracana*	Japan,Miyazaki	1976	*Eleusine*	GCA_004346965.1 *	Luciano et al 2019	97.1
NI907		*Digitaria*	Japan:Tochigi	1974	*Digitaria*	GCF_004355905.1 *	Gomez Luciano et al. 2019	97.3
P131		*Oryza sativa*	Japan	1976	*Oryza*	GCA_000292605.1	Xue et al. 2012	96.7
P3		*Triticum durum*	Paraguay: Canindeyu	2012	*Triticum*	GCA_002924885.1	Rahnama et al. 2021	97
PH0014-rn	PH0014, PH14	*Oryza sativa*	Philippines	-	*Oryza*	PH14-rn	Chiapello et al. 2015	96.6

PH42		*Eleusine coracana*	Philippines	1983	*Eleusine*	GCA_002924865.1	Pieck et al. 2017	95.5
PL2-1		*Lolium multiflorum*	Pulaski Co, KY, USA	2002	*Lolium*	GCA_002924835.1	Rahnama et al. 2021	97
PL3-1		*Lolium multiflorum*	Pulaski Co, KY, USA	2002	*Lolium*	GCA_002924825.1	Rahnama et al. 2021	96.7
RMg-Dl		*Oryza sativa*	India: Madhubani, Bihar	2012	*Oryza*	GCA_001853415.2 *	Reddy et al. 2021	93.7
Sar-2-20-1		*Oryza sativa*	Suriname: Saramacca	2013	*Oryza*	GCA_011799915.1 *	Zhong et al. 2020	97
Sv9610		*Setaria viridis*	China,Zhejiang	1996	*Setaria*	GCA_001548845.1	Zhong et al. 2016	97
Sv9623		*Setaria viridis*	China,Zhejiang	1996	*Setaria*	GCA_001548855.1	Zhong et al. 2016	96.8
T25		*Triticum aestivum*	Brazil: Parana	1988	*Triticum*	GCA_002924745.1	Rahnama et al. 2021	96.8
US0071	US71	*Setaria spp.*	USA	1998	*Setaria*	GCA_900474175.3 *	Langner et al. 2021	97.1
WHTQ		*Triticum aestivum*	Brazil	-	*Triticum*	GCA_002924665.1	Rahnama et al. 2021	93.8
Y34		*Oryza sativa*	China: Yunnan	1982	*Oryza*	GCA_000292585.1	Xue et al. 2012	96.7

* Denotes assemblies obtained using long-read next generation sequencing technologies

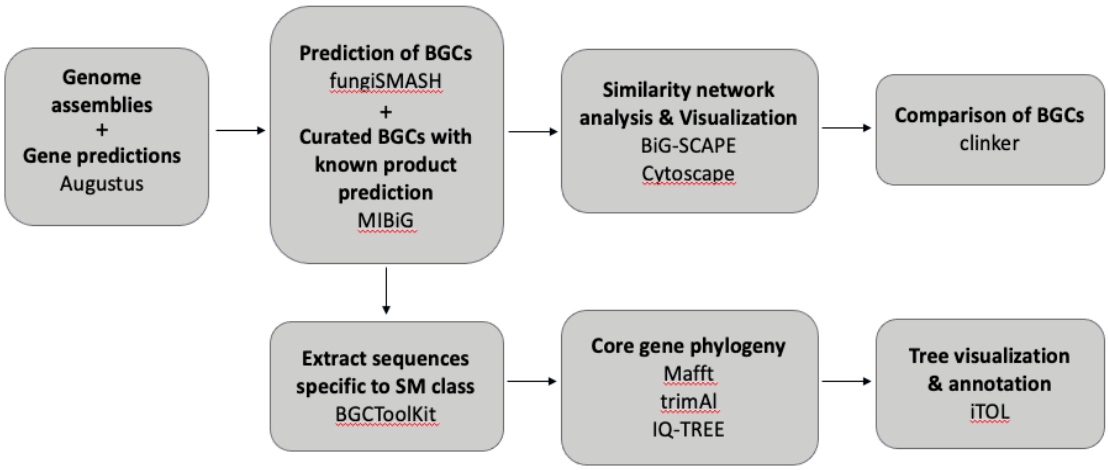

Figure 5.3: Workflow for exploring fungal biosynthetic diversity.

The process of pinpointing genomic regions harboring SM BGCs was executed across all 71 genomes using fungiSMASH (Blin et al., 2019). This investigation yielded a total of 4224 BGCs predictions, with an average of approximately 59 BGCs per strain. These projected BGCs were categorized based on their core biosynthetic genes, such as genes encoding polyketide synthases (PKSs), non-ribosomal peptide synthetases (NRPSs) or terpene cyclases (TCs). It is noteworthy that all lineages specific to particular host plants exhibited a similar number of BGCs across various BGC classes, with type I PKSs being the most prevalent **(Fig. 5.4)**. Overall, these analyses underscore the substantial potential of *M. oryzae* to synthesize SMs, some of which might play pivotal roles in virulence and/or host specialization.

5.2 Similarity network analyses to identify biosynthetic diversity in host-specific lineages

To ascertain whether any potential biosynthetic gene cluster (BGC) is linked to the ability to infect specific host plants, we performed a network similarity analysis using BiG-SCAPE (Navarro-Muñoz et al., 2019). This analysis encompassed a dataset of the 4224 predicted

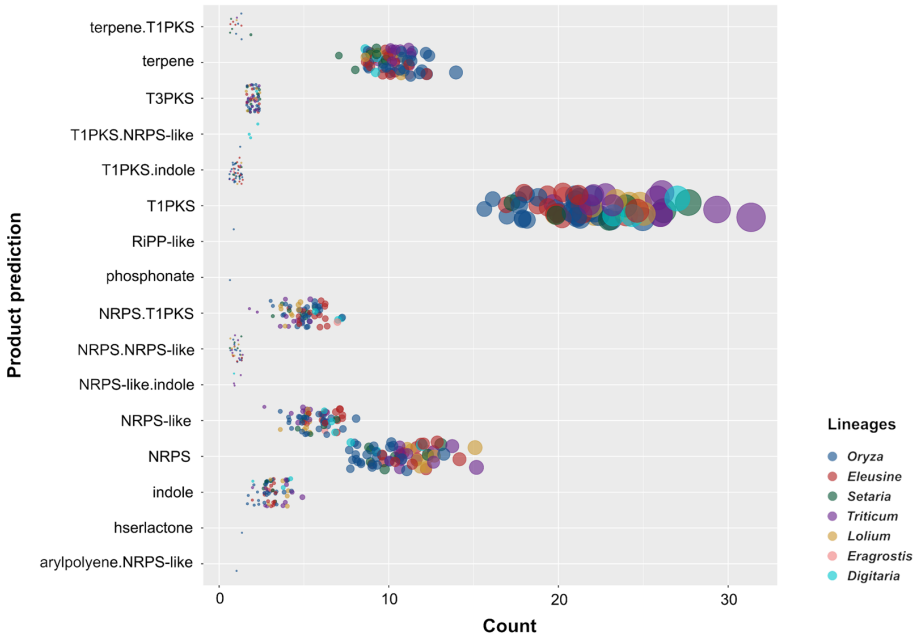

Figure 5.4: Occurrence of BGCs associated with specific classes of SM in individual genomes of *M. oryzae* and *M. grisea*. Area of a given circle is directly proportional to the total number of BGC associated with a class of specific product in a particular genome. Color of a given circle denotes the host-specific lineage it belongs to.

BGCs, complemented by 277 characterized BGCs sourced from the MIBiG database (Kautsar et al., 2019) for reference purposes. This resulted in a total of 4501 BGCs, which were subsequently clustered into 283 gene cluster families (GCFs) or subnetworks, of which 180 represents singletons. Among these, 160 belonged to the reference characterized BGCs **(Fig. 5.5)**. Our BiG-SCAPE analysis unveiled that, while the majority of the BGCs, regardless of their SM classification, are distributed across various lineages specific to different host plants (multi-colored closed circles grouped together; **Fig. 5.5**), only a limited number of GCFs or subnetworks exhibited similarities with reference BGCs derived from the MIBiG database. This observation implies that a significant portion of these BGCs remains uncharacterized.

The likely products associated with a specific BGC can be inferred through the examination of homology with reference BGCs known to encode pathways for SMs, particularly those identified in different fungi. As a result of our analysis, we have identified a subset of BGCs

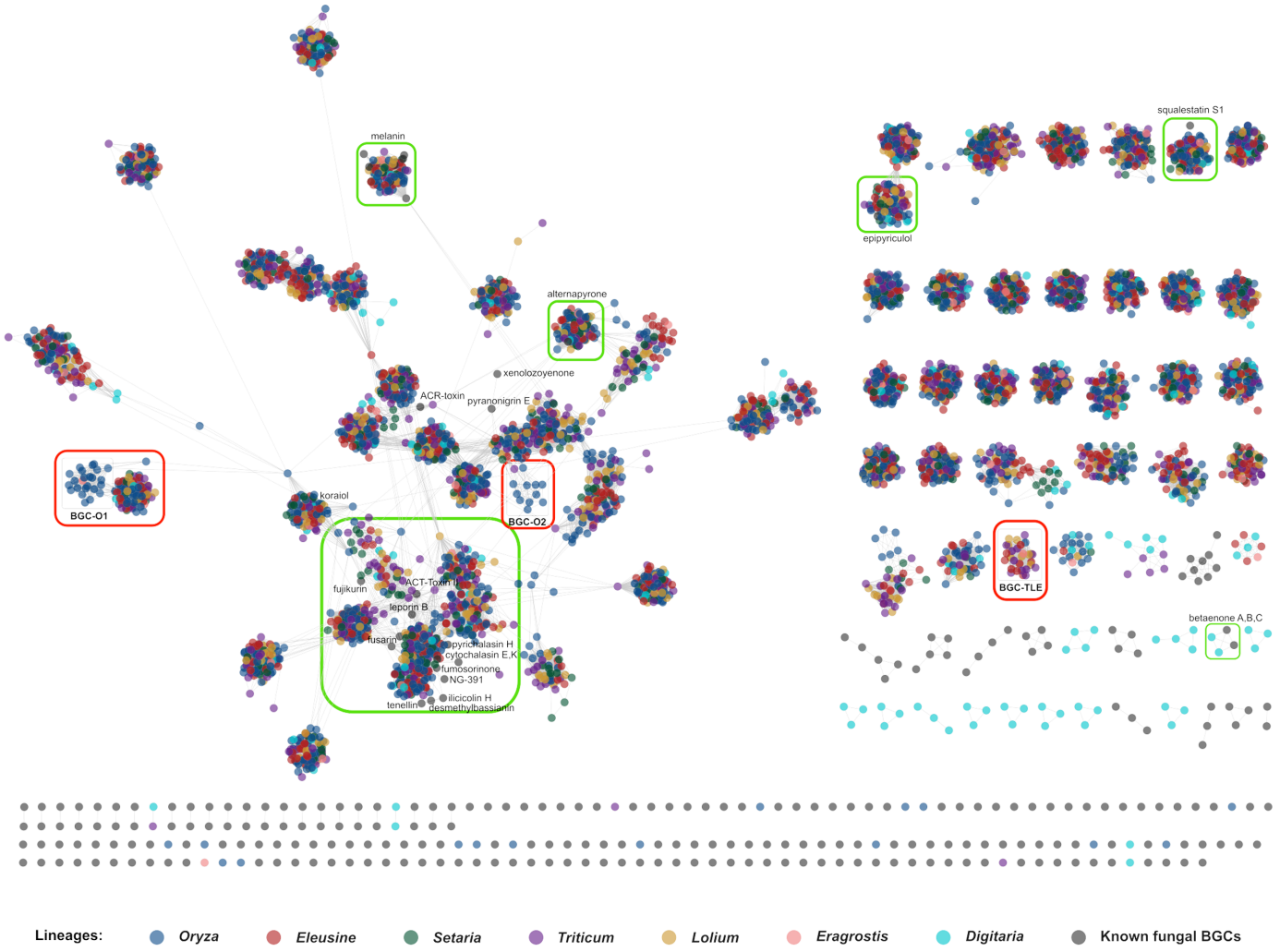
melanin
alternapyrone
xenolozoyenone
ACR-toxin
pyranonigrin E
koraiol
BGC-O1
BGC-O2
fujikurin
ACT-Toxin I
leporin B
fusarin
pyrichalasin H
cytochalasin E,K
fumosorinone
NG-391
tenellin
iiicicolin H
desmethylbassianin
squalestatin S1
epipyriculol
BGC-TLE
betaenone A,B,C
Lineages:
Oryza
Eleusine
Setaria
Triticum
Lolium
Eragrostis
Digitaria
Known fungal BGCs

Figure 5.5: Similarity network analysis of biosynthetic gene clusters (BGCs) from *M. oryzae* and *M. grisea*. A BiG-SCAPE analysis with a cutoff c0.5 depicts similarity of 4224 BGCs from *M. oryzae* or *M. grisea* with 277 reference BGCs from MIBiG database. Each dot represents a BGC and is color-coded according to the lineage. Gene cluster families (GCF; subnetworks) marked with green boxes share significant homology with reference MIBiG BGCs (grey-colored circle). GCFs marked with red boxes are found to be unique to host-specific lineages. The length of the gray lines is proportional to the genetic distance between BGCs. Singletons are shown as individual dots at the bottom.

that are likely associated with the synthesis of SMs such as melanin, cytochalasans, epipyriculol, squalestatin, Fusarin, fujikurin, alternapyrone, cercosporin, ACT-ToxinII, and pyranonigrin, in *M. grisea* and/or different lineages of *M. oryzae* (green boxes; **Fig. 5.5**). Within the 91 GCFs present in *Magnaporthe* strains, twelve GCFs were found to comprise characterized BGCs, primarily associated with the production of DHN melanin, epipyriculol, alternapyrone, squalestatin and cytochalasans in *M. oryzae*, and betaneone in *M. grisea* **(Fig. 5.5-5.10)**. Nonetheless, the majority of the remaining BGCs studied here, did not display any homology with the known/reference BGCs in other fungi. Forty-four GCFs are found in several *M. oryzae* lineages as well as in *M. grisea*, most of them remain uncharacterized. Interestingly, 13 GCFs were specific to *M. grisea*, while 14 GCFs appeared to be exclusive to *M. oryzae*, hinting at their potential roles in pathogenesis on *Digitaria* and other relevant host plants **(Fig. 5.5)**. Taken together, our comprehensive analyses strongly indicate that certain SM BGCs exhibit significant diversity among various *M. oryzae* lineages adapted to specific host plants and/or geographical locations.

5.3 Gene Cluster Families unique to host-specific lineages

Within the set of GCFs specific to *M. oryzae*, three GCFs could potentially be associated with host specialization. Specifically, BGC-O1 and BGC-O2 exhibited a predominant presence within the *Oryza* lineage, while BGC-TLE was unique to the *Tritici*, *Lolium, Eleusine* and *Eragrostis* lineages **(Fig. 5.5)**. These lineages shared a common ancestry and diverged from the *Oryza* and *Setaria* lineages **(Fig. 5.1)**. BGC-O1 encompasses genes encoding a Type 1 reducing polyketide synthase (rPKS) and tailoring enzymes **(Fig. 5.14)**, whereas BGC-O2 is restricted to a solitary T1 PKS gene **(Fig. 5.11)**.

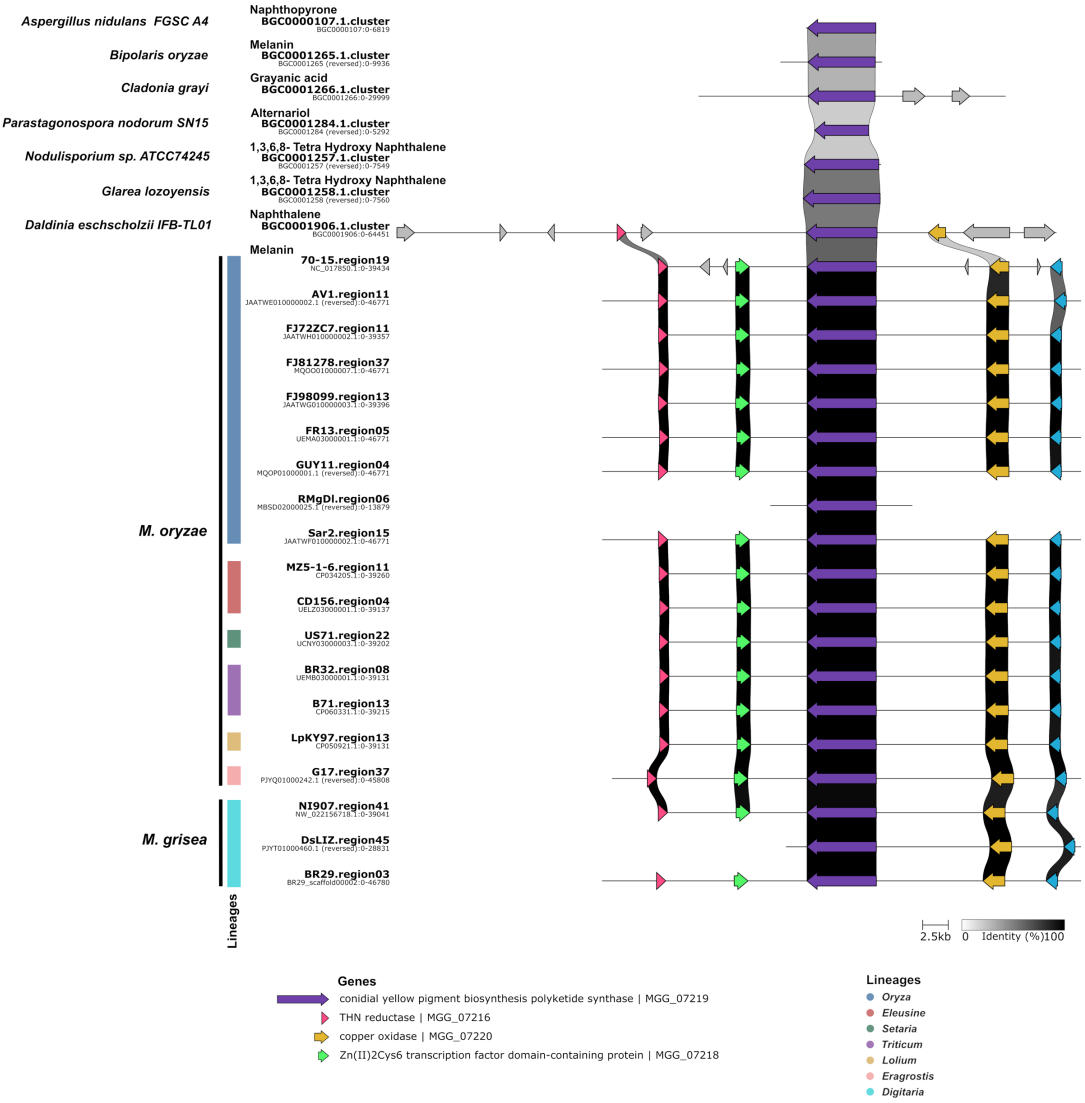

Figure 5.6: Homology of melanin-associated *M. oryzae* and *M. grisea* gene cluster family with reference BGCs in MIBiG database. The map depicts comparison of BGC loci with reference BGCs associated with melanin or melanin-like SM product in MIBiG. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

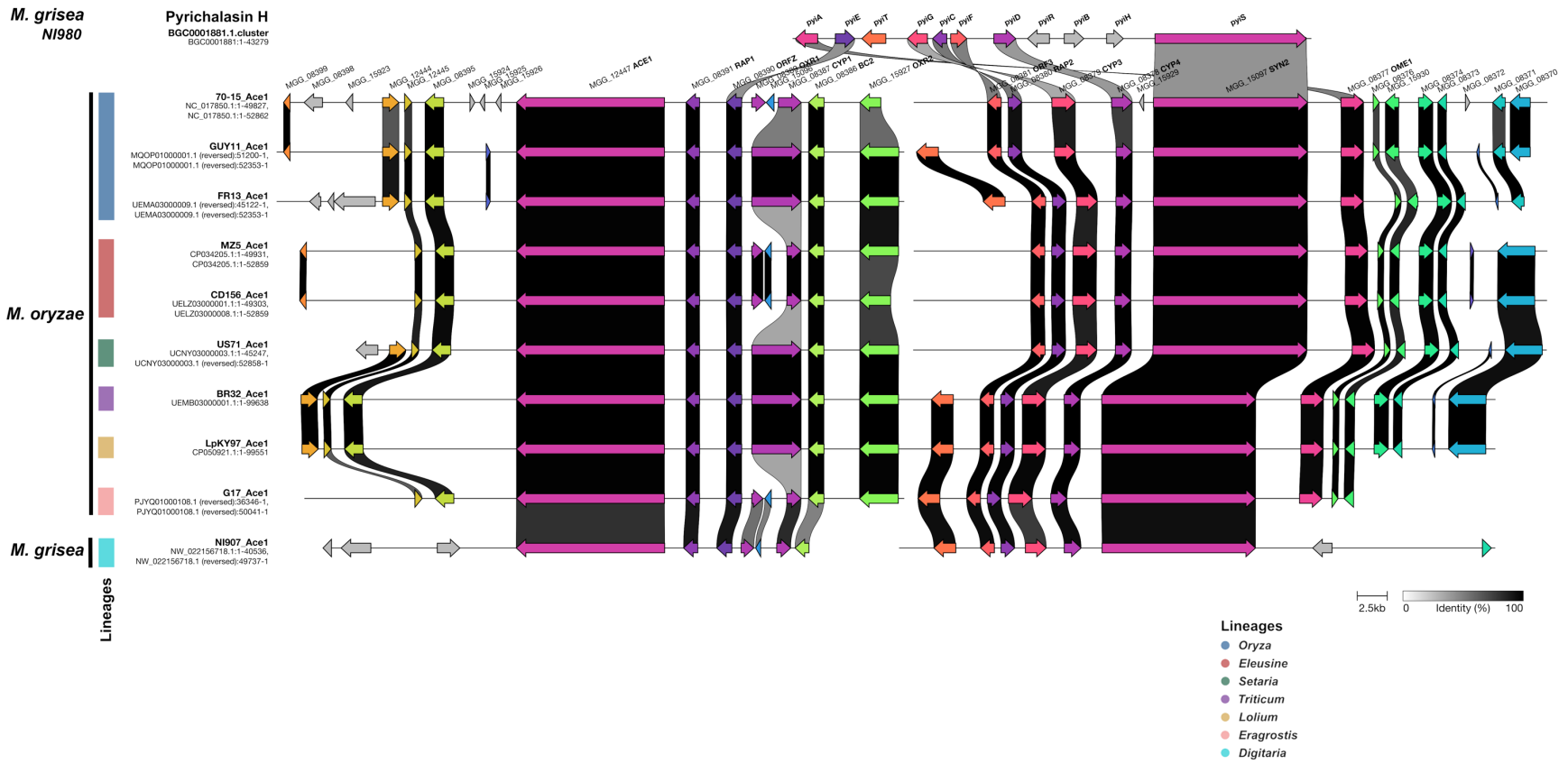

Figure 5.7: Analysis of homology between *ACE1* gene cluster family in *M. oryzae* and *M. grisea* and reference BGCs in MIBiG database. The map depicts comparison of BGC loci with reference BGCs associated with pyrichalasin H in MIBiG. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

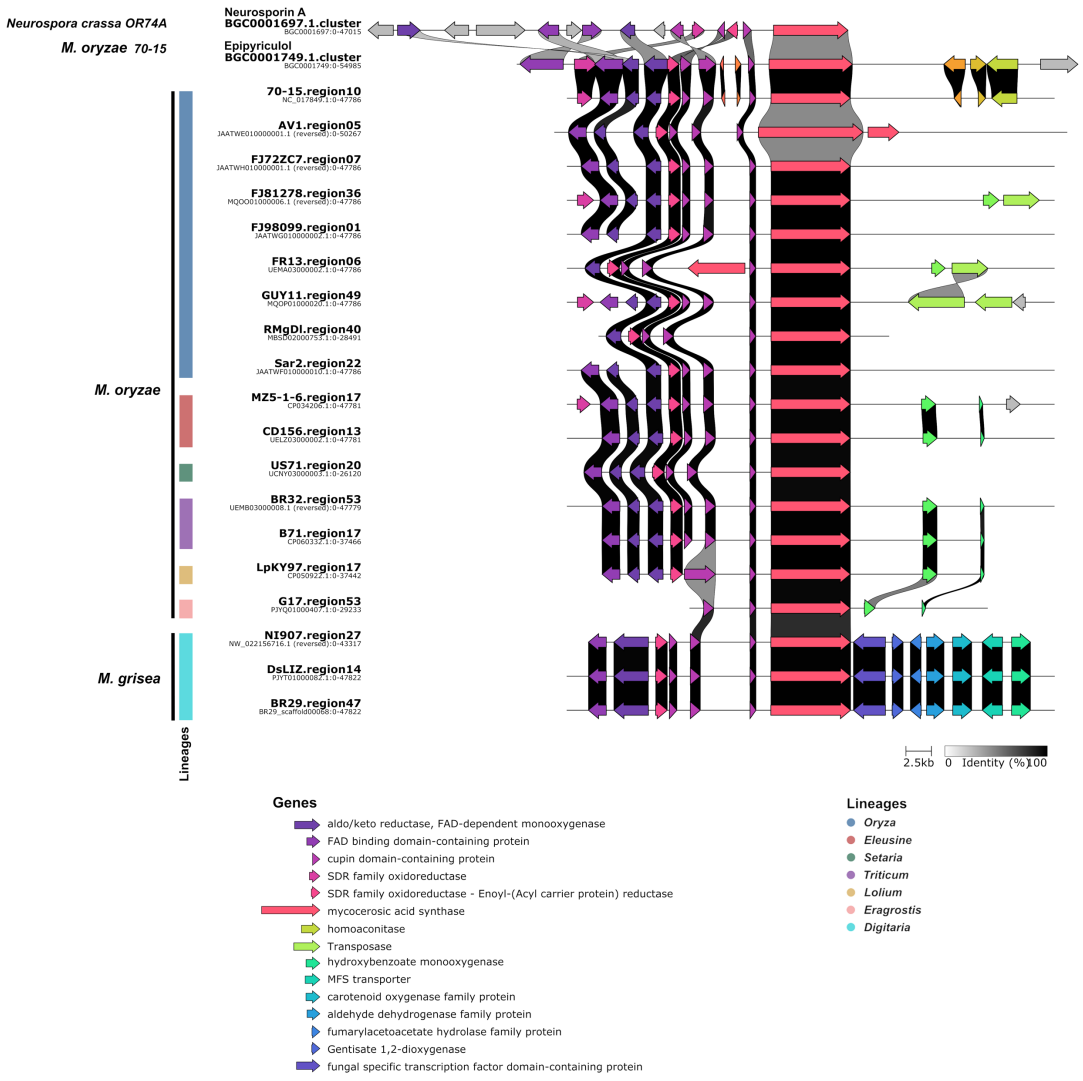

Figure 5.8: Analysis of homology between putative epipyriculol-associated gene cluster family in *M. oryzae* and *M. grisea* and reference BGCs in MIBiG database. The map depicts comparison of BGC loci with reference BGCs associated with epipyriculol in MIBiG. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

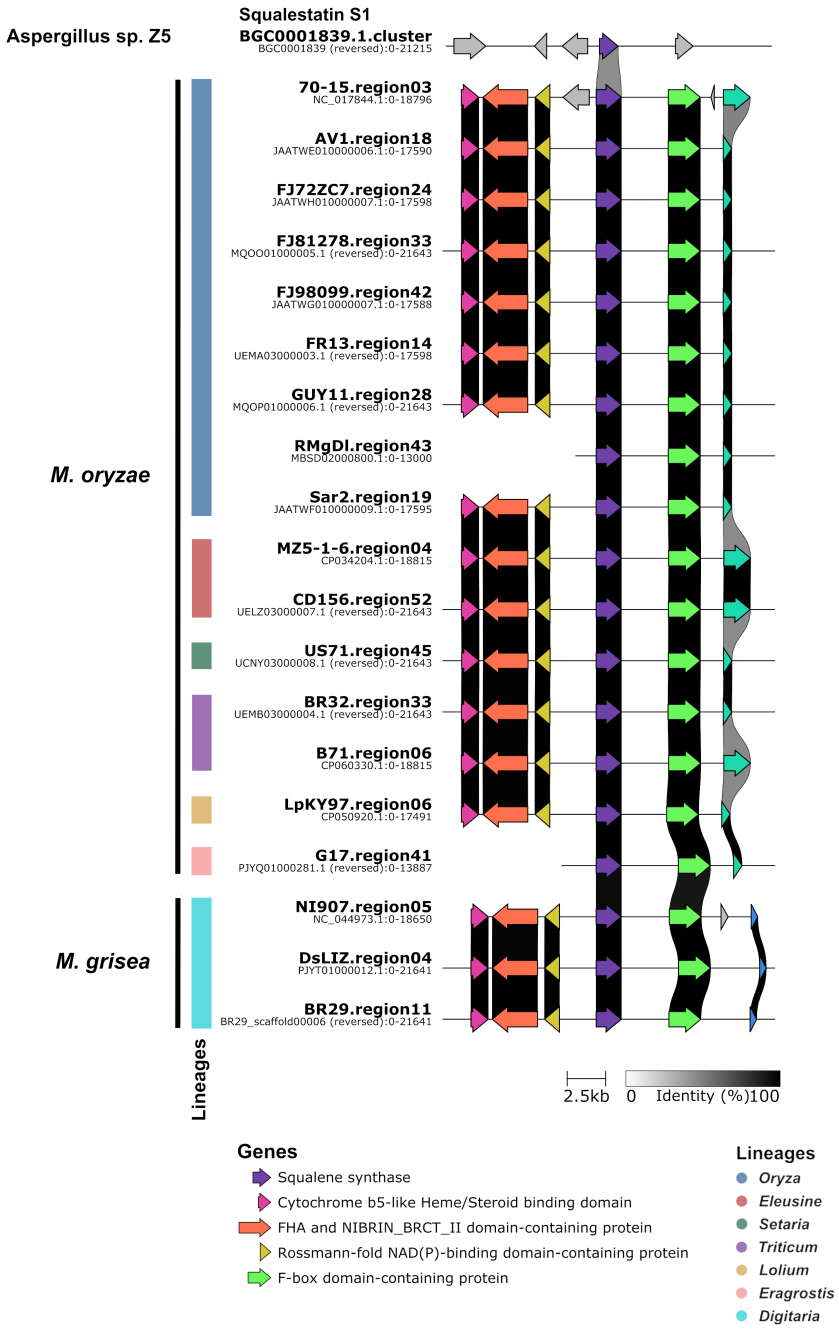

Figure 5.9: Homology of putative Squalestatin S1-associated *M. oryzae* and *M. grisea* gene cluster family with reference BGCs in MIBiG database. The map depicts comparison of BGC loci with reference BGCs associated with Squalestatin S1 in MIBiG. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

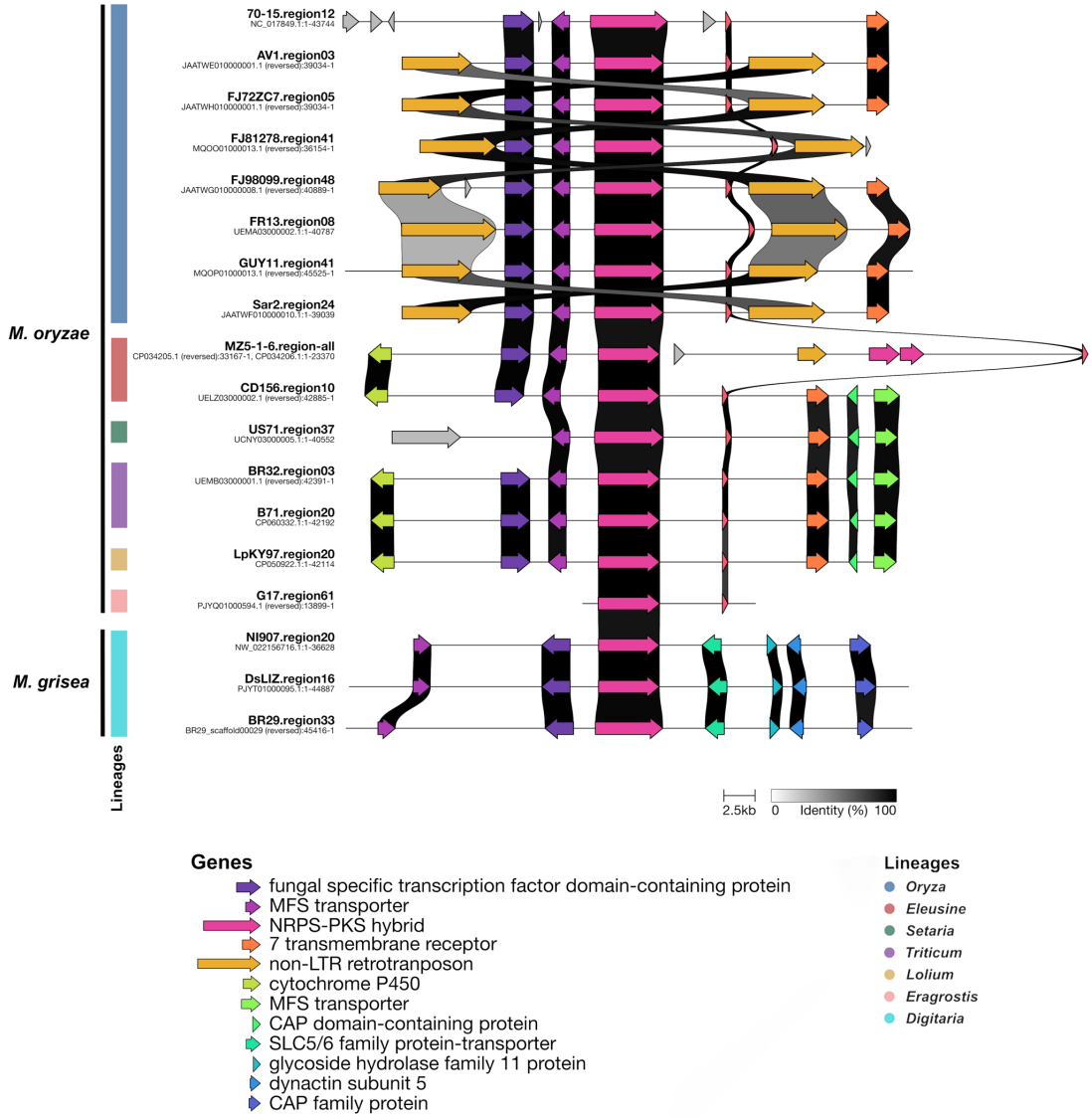

Figure 5.10: Analysis of homology between Tenuazonic acid-associated gene cluster families in host-adapted *M. oryzae* and *M. grisea* isolates. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

In the case of BGC-TLE, although the standalone core biosynthetic NRPS gene therein is present in all the strains, there were notable differences in the gene structure **(Fig.5.12)**. Through manual curation of the NRPS gene, it was discerned that the *Triticum* and *Lolium* lineages had a deletion of 557 bps, corresponding to second exon observed in the *Eleusine* lineage (CD156 strain). Furthermore, the presence of a stop codon within the Amp-binding domain of the NRPS gene in the *Triticum* and *Lolium* lineages suggested pseudogenization of the NRPS gene and, consequently, the possible non-functionality of the corresponding BGC in those strains **(Fig. 5.13)**. It is conceivable that BGC-TLE may either play a pivotal role in virulence exclusively on *Eleusine* host plants, or its product could serve an avirulence effector-like function in *Triticum* and *Lolium* hosts. In response to an evolutionary arms-race, pathogen from the *Triticum* and *Lolium* lineages have likely undergone adaptations leading to the loss of a functional NRPS gene, allowing them to continue infecting these hosts. Therefore, among the three candidate BGCs, our further exploration has focused on BGC-O1, as it holds the potential for involvement in specialization on the rice host.

5.4 Identification of a novel reducing polyketide synthase BGC unique to *Oryza* lineage

BGC-O1 was detected in 23 out of the 24 strains belonging to *Oryza* lineage that were employed in this study, as well as in a single strain from the *Eragrostis* lineage **(Fig. 5.14)**. This GCF seems to encompass two distinct networks, and a comparative analysis of the genomic loci indeed clearly distinguishes between two potential BGCs - one conserved across all the lineages, while the other, BGC-O1 is specifically found in the *Oryza* and *Eragrostis* lineages **(Fig. 5.14)**. BGC-O1 comprises a novel core biosynthetic gene, a reducing type I polyketide synthase (rPKS) gene, encoded by MGG_08236, along with adjacent tailoring genes. These tailoring genes include one methyl transferase (MGG_15107), one Co-A transferase (MGG_15108) and two cytochrome P450 monooxygenases (MGG_12496 and MGG_12497).

The predominance of BGC-O1 in the *Oryza*-specific lineage strongly suggests that the product generated by the core PKS might play a role in specialization or adaptation to the rice host. Crucially, it is noteworthy that BGC-O1 did not exhibit any similarity with any of the reference BGCs present in the MIBiG database.

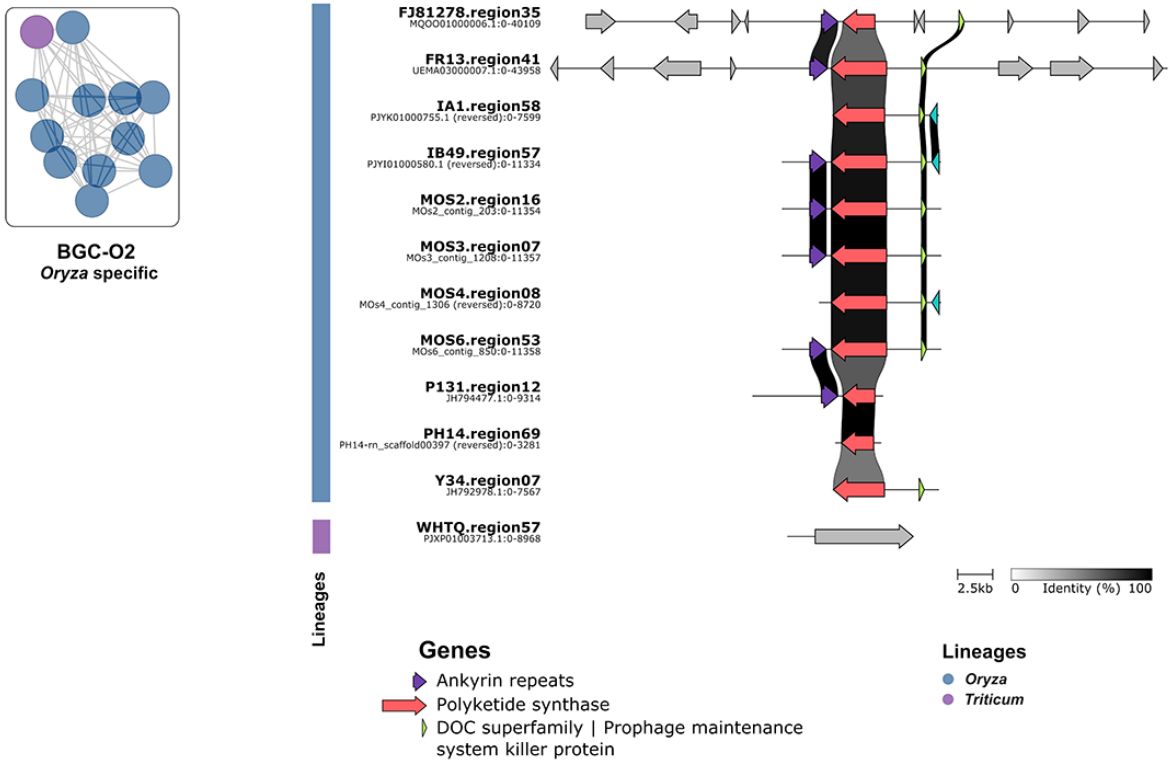

Figure 5.11: BGC-O2 is predominantly present in *Oryza*-specific lineage of *M. oryzae*. BiG-SCAPE analysis showing BGC-O2 GCF with all the BGCs specifically present in 11 genomes from *Oryz*a lineage and one genome from *Triticum* lineage. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

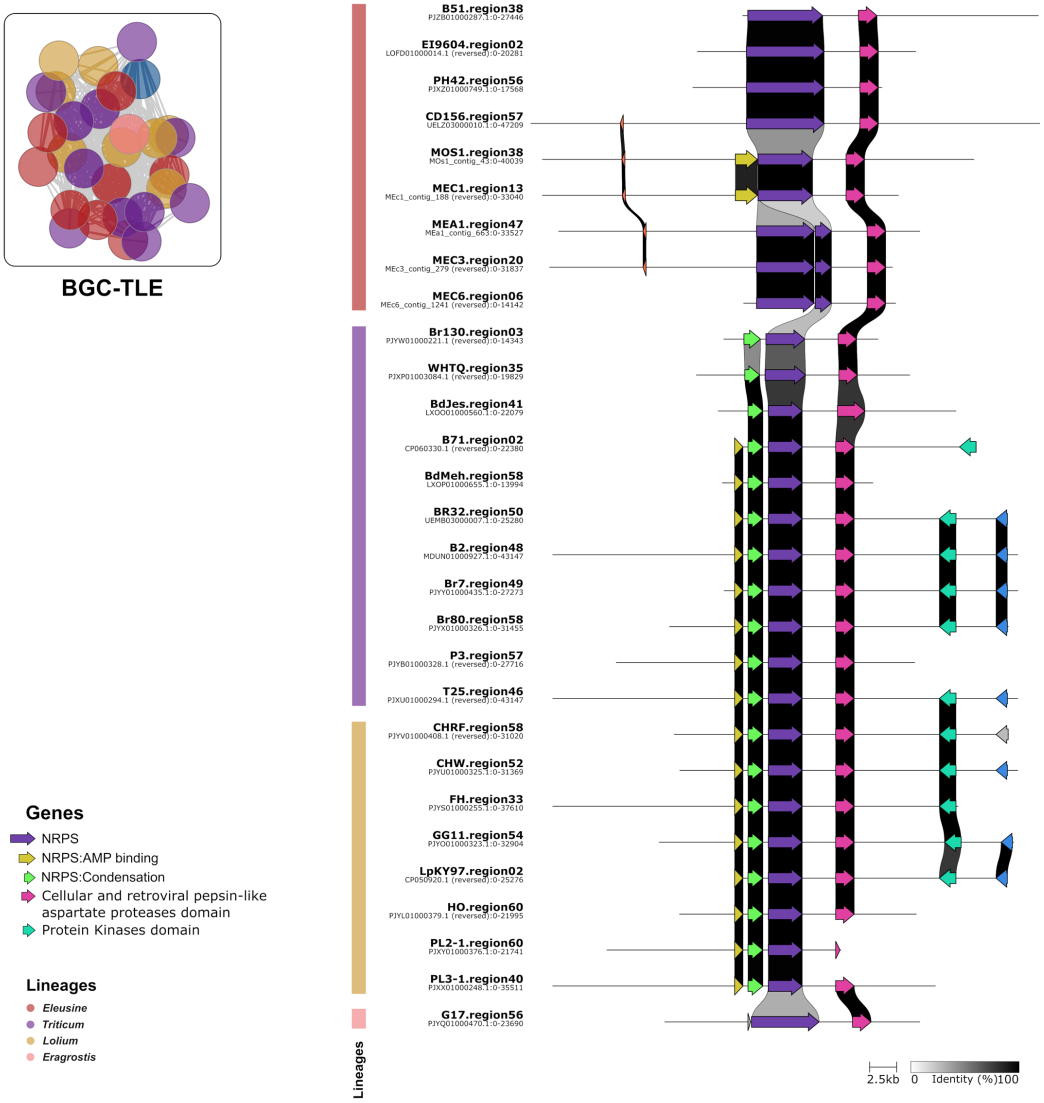

Figure 5.12: BGC-TLE is predominantly present in *Triticum*, *Lolium* and *Eleusine*-specific lineages of *M. oryzae*. BiG-SCAPE analysis showing BGC-TLE GCF with all the BGCs specifically present in 12, 8, and 9 genomes from *Triticum*-, *Lolium*- and *Eleusine*-specific lineages, respectively. The shaded area between any two arrows denotes degree of homology (0 to 100%; white to black, respectively) between the two sequences.

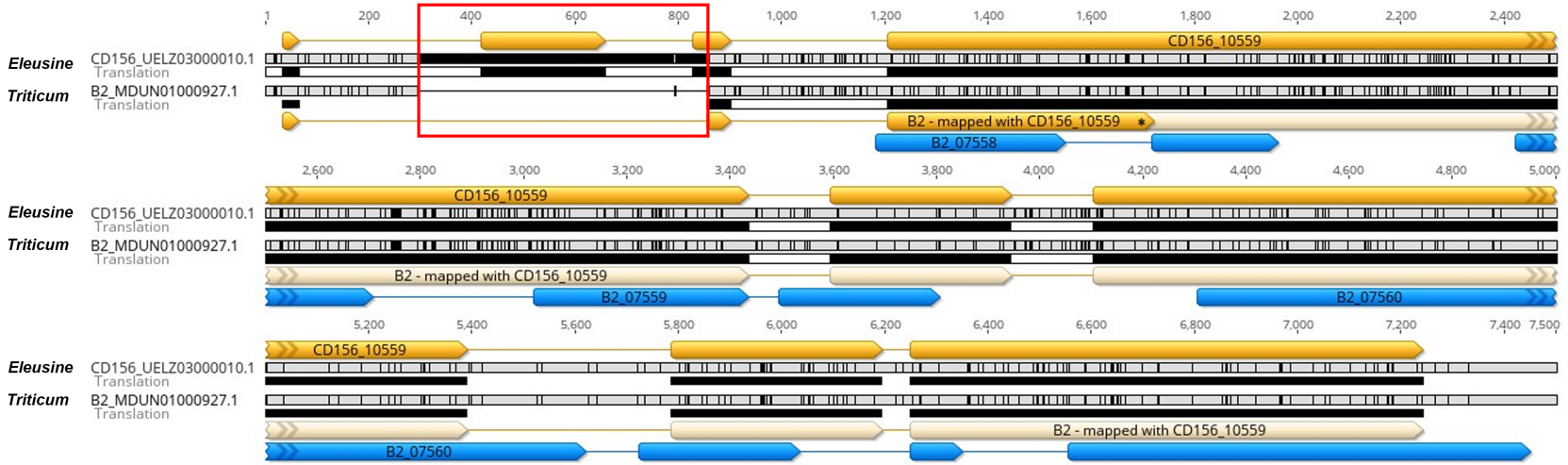

Figure 5.13: Comparison of core biosynthetic gene region of BGC-TLE shows the pseudogenization in NRPS gene belonging to *Triticum* lineage. Alignments between representative strains CD156 (*Eleusine*) and B2 (*Triticum*) shows the presence of premature stop codon depicted as * in AMP-binding domain of NRPS gene. Yellow arrows represent the gene models in B2 in accordance with the one in CD156, whereas blue arrows in B2 strain represents the earlier predicted gene model by Augustus. Red box displays the deletion of 557 bps in B2 strains, which corresponds with the second exon of CD156_10559 NRPS gene.

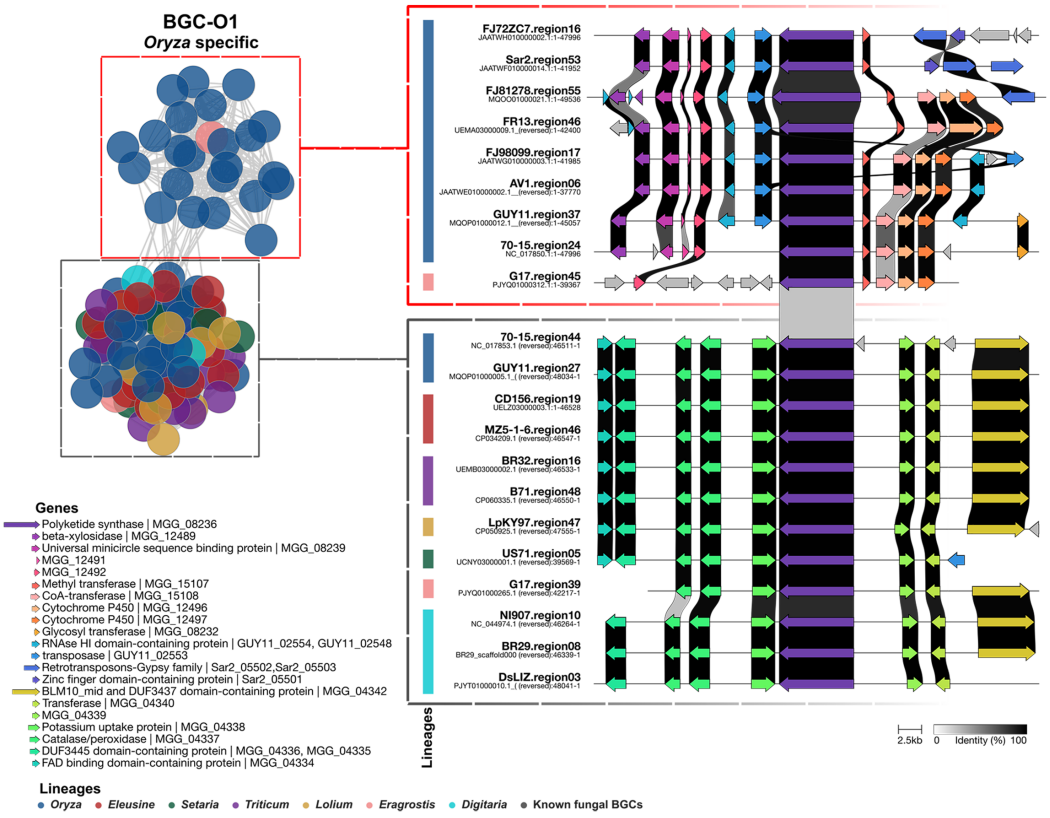

Figure 5.14: BGC-O1 is predominantly present in *Oryza* lineage of *M. oryzae*. BiG-SCAPE network showing BGC-O1 present in 23 genomes from *Oryza* and one genome from *Eragrostis* lineages. Conservation of the BGC in selected strains is depicted using Clinker. Vertical bars besides the name of strain are colored according to the lineages. Red and gray dashed boxes separate BGC-O1 and BGC-O1-like clusters, respectively.

This observation suggests that BGC-O1 represents a distinctive gene cluster likely involved in the biosynthesis of a novel secondary metabolite, associated with the specialization on the rice host.

We were curious to determine whether the specificity of the BGC-O1 cluster region was exclusive to the *Oryza* lineage, and thus consequently we assessed the conservation or variability of the flanking genomic regions surrounding BGC-O1 in different lineages. To achieve this, we aligned representative high-quality genome assemblies from each lineage against the reference genome assembly of the 70-15 strain, which belongs to the *Oryza* lineage. This allowed us to evaluate synteny at a global level. Our analysis revealed that BGC-O1 is situated in the sub-telomeric region of chromosome 2 (NC_017850.1) and is

located approximately 528 kb downstream of the *ACE1* gene cluster in the reference strain 70-15. When comparing the macrosynteny between 70-15 and GUY11, both strains from the *Oryza* lineage, we observed conservation of ~134 kb upstream and ~117 kb downstream flanking regions, encompassing the BGC-O1 cluster **(Fig. 5.15A)**. Notably, the immediate flanking regions, which span ~10 kb, exhibited a translocation between the GUY11 and 70-15 genomes (blue lines crossing over within the orange band; black arrowhead; **Fig. 5.15A**). A similar comparison between 70-15 and FR13 strains displayed partial conservation with ~14 kilobases of the BGC-O1 region retained in the FR13 strain, while an approximately 9.5 kilobase segment from the BGC-O1 region in the 70-15 strain was relocated further downstream on the same contig in the FR13 strain (red arrowhead; **Fig. 5.15A**). Furthermore, although ~102 kilobases of the upstream flanking region remained syntenic between 70-15 and FR13, the downstream flanking region exhibited several structural rearrangements **(Fig. 5.15A)**.

In contrast, when comparing with non-*Oryza* lineages, such as *Eleusine* and *Setaria*, a significant loss of synteny was observed in the sub-telomeric region of chromosome 2. While the upstream flanking region located ~97 kb upstream of BGC-O1 in 70-15 strain, exhibited synteny with a corresponding genomic locus in MZ5-1-6, CD156 and US71, the BGC-O1 cluster and downstream sequences were either missing or exhibited limited conservation in the genomes of these lineages **(Fig. 5.15B)**.

Further, in order to understand the genetic diversity in these SM-BGCs, we looked for overall synteny between the two genomes (70-15 and MZ5-1-6) with chromosome-level assemblies. The orthologous gene-pairs, among all the genes present in SM-BGCs, were investigated using a Bi-directional Best Blast Hit approach, where we identified 529 orthologous SM gene pairs between the two isolates **(Table 5.2)**. Whereas, we identified 638 orthologous gene pairs, when SM genes of MZ5-1-6 were compared with total genes of 70-15. This suggests that the orthologs of 109 SM-BGC genes of MZ5-1-6 are likely located in the regions outside of the predicted SM-BGCs in the 70-15 genome, likely due to differential rearrangements in these two genomes.

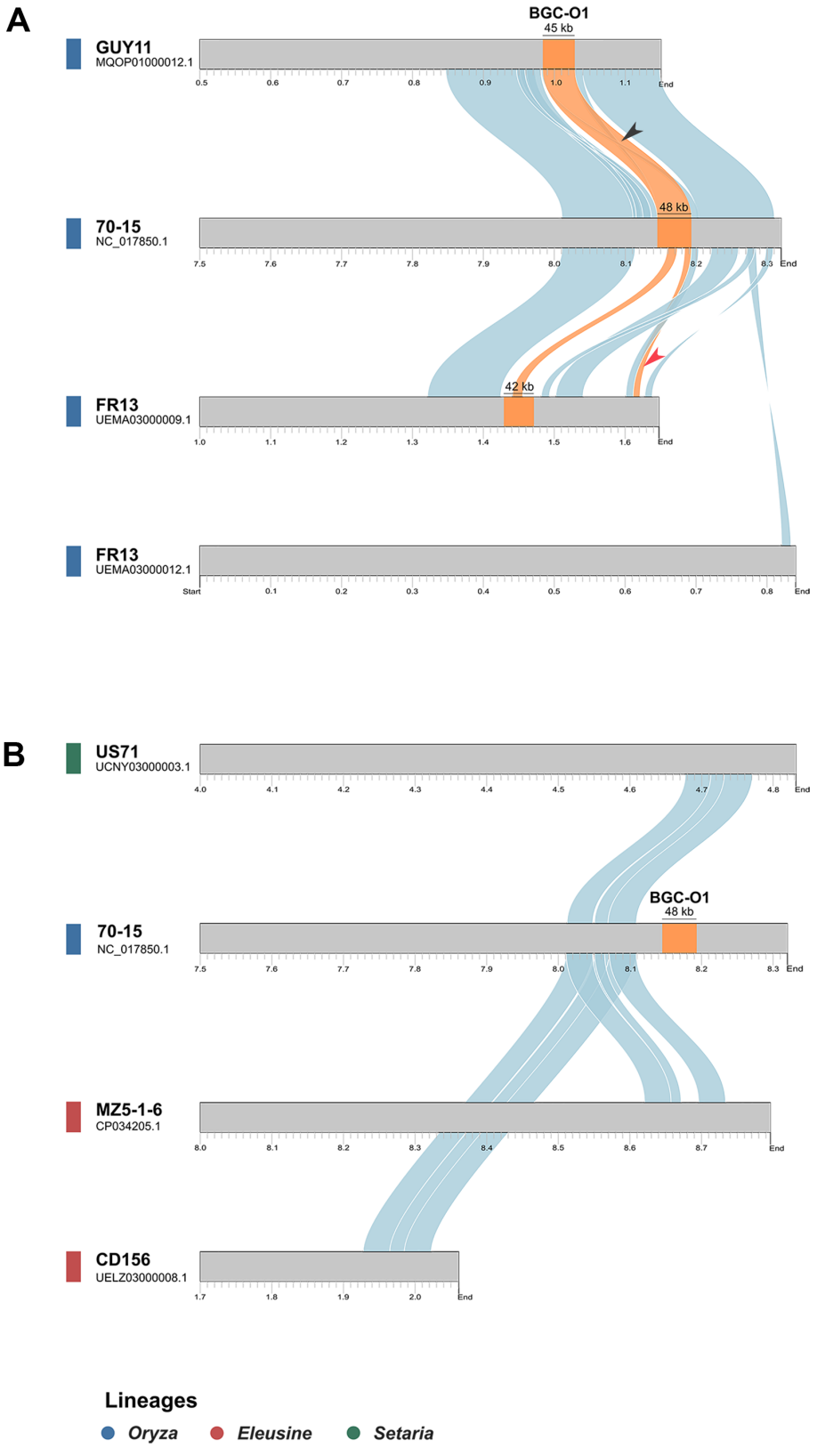
A
BGC-O1
45 kb
GUY11
MQOP01000012.1
0.5
0.6
0.7
0.8
0.9
1.0
1.1
End
48 kb
70-15
NC_017850.1
7.5
7.6
7.7
7.8
7.9
8.0
8.1
8.2
8.3
End
42 kb
FR13
UEMA03000009.1
1.0
1.1
1.2
1.3
1.4
1.5
1.6
End
FR13
UEMA03000012.1
Start
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
End
B
US71
UCNY03000003.1
4.0
4.1
4.2
4.3
4.4
4.5
4.6
4.7
4.8
End
BGC-O1
48 kb
70-15
NC_017850.1
7.5
7.6
7.7
7.8
7.9
8.0
8.1
8.2
8.3
End
MZ5-1-6
CP034205.1
8.0
8.1
8.2
8.3
8.4
8.5
8.6
8.7
End
CD156
UELZ03000008.1
1.7
1.8
1.9
2.0
End
Lineages
Oryza
Eleusine
Setaria

Figure 5.15: Genomic localization of BGC-O1 and synteny between the *Oryza* and non-*Oryza* lineages. A) Synteny analysis using pairwise comparison within *Oryza* lineage. Genomes of GUY11 and FR13 aligned individually with reference genome 70-15. The syntenic BGC-O1 locus (orange) and the flanking regions (blue) on chromosome 2 of 70-15 are depicted. The differential lengths of BGC-O1 in different isolates are marked with a bar and corresponding length in kilobases. The black arrowhead depicts genomic rearrangement (swapping) in the flanking ~10 Kb region. The red arrowhead marks the rearrangement of ~9 Kb region of the BGC-O1 in FR13. B) Synteny analyses of representative genome assemblies belonging to *Setaria* (US71), *Eleusine* (MZ5-1-6 and CD156) and *Oryza* (70-15) lineages. The BGC-O1 locus (orange) and flanking region (blue) sequences from US71, MZ5-1-6 and CD156 aligned with that of 70-15 are depicted.

The Circos plots were constructed to determine the synteny of these orthologous SM-genes between the genomes of the rice isolate (70-15) and finger millet isolate (MZ5-1-6) **(Fig. 5.16)**. While, most of the genes were highly syntenic, a total of 36 SM-gene pairs located on to different chromosomes in the two genomes. Notably, 30 out of 36 SM-genes, which located on the chromosome 6 of MZ5-1-6, were rather found on chromosome-1 of 70-15. Such large chromosomal translocation events have been reported earlier for MZ5-1-6 and 70-15 (Luciano et al. 2019). Thus, the differential genomic rearrangement, especially with respect to the genes involved in secondary metabolism, in the rice and millet isolates might have some evolutionary role in adaptation to a specific host plant.

Table 5.2: Summary of ortholog pairs between the genomes of millet isolate (MZ5-1-6) and rice isolate (70-15).

Pairs	**Total Genes MZ5 vs 70-15**	**MZ5 SM-genes vs 70-15 total genes**	**MZ5 SM-genes vs 70-15 SM- genes**
Total number of orthologs	10813	638	529
Total number of orthologs rearranged	671	84	36
Total number of orthologs rearranged between MZ5_Chr6 and 70-15_Chr1	549	50	30

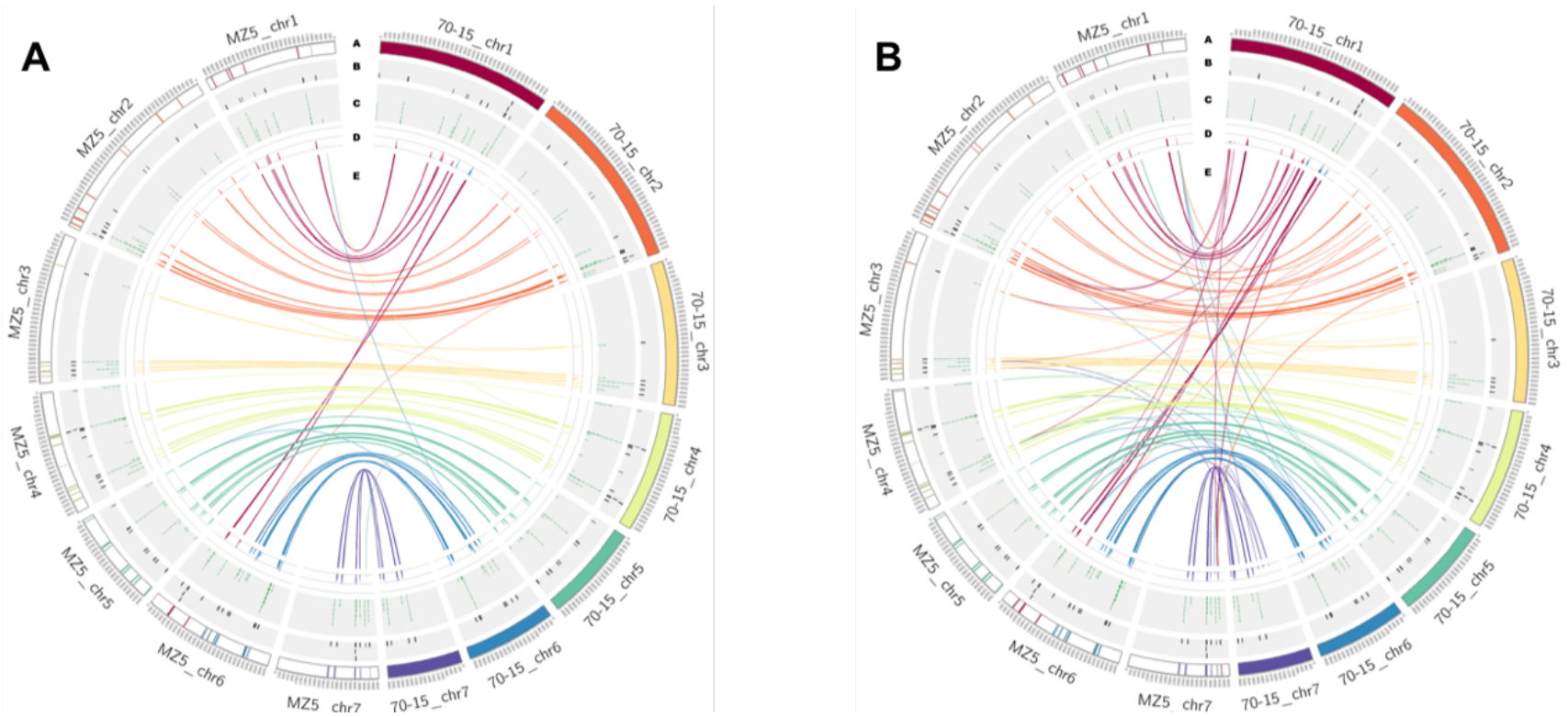

Figure 5.16: Circos plots showing synteny among SM related genes between representative strains from *Oryza* and *Eleusine* lineages. The orthologous gene pairs represented by connecting links between (A) SM-genes of both MZ5-1-6 and 70-15 and (B) SM-genes of MZ5-1-6 and total genes of 70-15. In each plot, individual tracks (A to E) denote different characteristics – Track A, chromosomes scaled according to their length; Track B, candidate clusters positioned accordingly on the chromosomes; Track C, core biosynthetic genes with respect to their positions

on the chromosomes; Track D, histogram showing orthologous gene-pairs between the two genomes, with color codes according to those of the chromosomes of 70-15; and Track E, links representing orthologous gene-pairs between the two genomes, with color codes same as those for the 70-15 chromosomes.

5.5 Evolutionary history of the novel polyketide synthase gene

In our quest to unravel the evolutionary history of BGC-O1, we embarked on a search for orthologous and closely related counterparts of the MGG_08236 rPKS protein within the *Pezizomycotina* group from the MycoCosm repository (Grigoriev et al., 2014;). Interestingly, only two closely resembling homologues were identified from the fungus *Colletotrichum eremochloae*. Employing a phylogenetic analysis, it became evident that the rPKS 670826 from *C. eremochloae* is a true orthologue of MGG_08236 rPKS, being in the adjacent clade to the *M. oryzae* lineage **(Fig 5.17A)**. Meanwhile, the other paralogue in *C. eremochloae*, 679399, falls within a clade restricted to the *Colletotrichum* genus and constitutes a sister clade to the MGG_08236 rPKS clade. Upon conducting a comparative analysis of the genomic loci encompassing both homologues in *C. eremochloae* and BGC-O1 in *M. oryzae,* it was revealed that these two clusters shared the rPKS and two cytochrome P450 genes with BGC-O1. However, the methyltransferase and Co-A transferase genes were found as single copies elsewhere within the *C. eremochloae* genome **(Fig. 5.17B)**. Notably, the orthologous BGC in *C. eremochloae* shares an average nucleotide identity of 76% when compared to the *M. oryzae* BGC. Intriguingly, both BGCs in *C. eremochloae* share only 50% nucleotide identity with each other.

5.6 BGC-O1 genes are expressed specifically during host invasion

In our study, we conducted PCR targeting the open reading frame (ORF) region of the rPKS gene MGG_08236, using genomic DNA extracted from *M. oryzae* strains from rice and finger millet host plants **(Fig. 5.18A)**. The MGG_08236 gene was found to be present in all the examined *Oryza* strains studied, except for MOS3, which was stood out as the sole *Oryza* lineage strain lacking the BGC-O1 as investigated by our in-silico analysis.


A
Tree scale: 0.1
jgi|Colgr1|2442|GLRG 11878T0
jgi|Coler1|679399|MIX14804 34 33
jgi|Colso1|621565|MIX19503 23 27
jgi|Colsu1|687791|estExt Genemark1.C 10058
jgi|Colzo1|733567|estExt Genemark1.C 3150005
jgi|Gloci1|1841336|e gw1.15.201.1
jgi|Color1|8761|CORC01 00678T0
jgi|Colph1|455038|fgenesh1 pg.33 28
jgi|Colgo1|716101|estExt fgenesh1 pg.C 230218
jgi|Gloac1|1398789|estExt Genewise1Plus.C 520057
jgi|Colfi1|281529|mRNA6349
jgi|Colac2|797178|estExt Genemark1.C 1740014
jgi|Colsi1|979924|mrna 8044
jgi|Colny1|1022492|mrna 14212
jgi|Colsp1|6070|mrna 3976
jgi|Colme1|1833|mrna 1833
jgi|Colpa1|2449|mrna 12266
jgi|Collup1|10609|mrna 10609
jgi|Collu1|410842|e gw1.30.48.1
jgi|Colcu1|7861|mrna 7861
jgi|Colta1|1723|mrna 14004
jgi|Colab1|4797|mrna 4797
jgi|Colco1|15579|mrna 8055
jgi|Coler1|670826|MIX6231 187 8
IB49.region51 L0 CDS3
MOS5.region03 L0 CDS2
FJ98099.region18 L0 CDS9
ML33.region50 L0 CDS2
FJ98099.region17 L0 CDS7
Sar2.region53 L0 CDS7
FJ72ZC7.region16 L0 CDS7
FJ98099.region18 L0 CDS16
PH14.region61 L0 CDS5
AV1.region06 L0 CDS5
jgi|MagorTH3 1|1461275
FJ81278.region55 L0 CDS10
RMgDI.region02 L0 CDS3
FR13.region46 L0 CDS5
P131.region20 L0 CDS1
70-15.region24 L0 CDS6 MGG_08236
IE1K.region54 L0 CDS1
MG01.region08 L0 CDS1
jgi|MagorGY11 1|1536307
Y34.region17 L0 CDS1
GUY11.region37 L0 CDS6
IB33.region53 L0 CDS2
IA1.region54 L0 CDS1
jgi|MagorFR13 1|2134543
B157.region05 L0 CDS1
MOS2.region14 L0 CDS2
IC17.region53 L0 CDS2
jgi|Magor1|5062
G17.region45 L0 CDS8
B
C. eremochloae
Coler1.region048
scaffold_5:1-46229
Coler1|679399
50%
C. eremochloae
Coler1.region029
scaffold_178 (reversed):28187-1
Coler1|670826
76%
P. oryzae
70-15.region24
NC_017850.1:1-47996
MGG_08236
P. oryzae
Magor1.region020
Chromosome_2:1-47996
Magor1|5062
P. oryzae
MagorGY11_1.region032
scaffold_24:1-48366
MagorGY11_1|1536307
2.5kb
0
Identity (%)
100
Genes:
Polyketide synthase
Methyl transferase
Co-A transferase
beta-xylosidase
Cytochrome P450
Cytochrome P450
Glycosyl transferase

Figure 5.17: Evolutionary origin of the reducing polyketide synthase (PKS) MGG_08236 gene. A) Maximum-likelihood phylogenetic tree using protein sequences of reducing PKS genes from BGC-O1 and related gene cluster families, as well as homologues retrieved from MycoCosm repository. Tip labels depicting sequences from *Oryza* and *Eragrostis*-specific clade are marked with blue and pink background, respectively. The *Colletotrichum eremocloae* ortholog (jgi.p_Coler1_670826), closer to MGG_08236, is denoted with pink label; whereas the more distant paralogue (jgi.p_Coler1_679399) is shown in orange. Gray shaded triangles denote collapsed clades with distant sequences. Branches were supported by > 95% Bootstrap values indicated with gray circles at the nodes. The tree is rooted at midpoint. B) Comparative analysis, using Clinker tool, of the BGC-O1 or BGC-O1-like loci from three *Oryza*-specific and *C. eremochloae* genomes.

The MGG_08236 ORF was notably absent in all the *Eleusine* strains investigated, as well as in the MOS1 and MOS4 strains, both of which are placed outside of *Oryza* lineage according to our phylogenetic analysis **(Fig. 5.1B and 5.18A)**.

Subsequently, we delved into the investigation of the expression profiles of the rPKS genes (MGG_08236) and three associated tailoring genes, namely methyl transferase (MGG_15107), CoA-transferase (MGG_15108) and cytochrome P450 (MGG_12496), all of which are members of the BGC-O1 locus. This was achieved through semi-quantitative RT-PCR analysis conducted at different stages of infection. Total RNA was isolated from fungal vegetative mycelia cultured in complete medium, barley leaves inoculated with *M. oryzae* strain and incubated for different time intervals, and uninoculated barley leaves as mock samples. These assays were conducted to assess the transcript levels of the aforementioned genes relative to those of β-Tubulin, serving as an endogeneous control. Our finding indicated that the expression of the core rPKS gene was either undetected or barely seen during vegetative growth – mycelium and pre-invasive stage - 12 hours post inoculation (hpi), respectively. However, transcript accumulation commenced during the progression of pathogenesis, with a substantial increase in expression at 24 hpi, followed by a consistent, albeit gradual, rise in expression levels until 72 hpi **(Fig. 5.18B and 5.18C)**. The expression patterns of MGG_15107 and MGG_15108 tailoring genes displayed partial correlation with that of rPKS, albeit with a lower expression level **(Fig. 5.18B and 5.18C)**. In contrast, MGG_12496 exhibited a different expression pattern with its highest expression levels observed in mycelium and at 48 hpi during infection **(Fig. 5.18B and 5.18C)**.

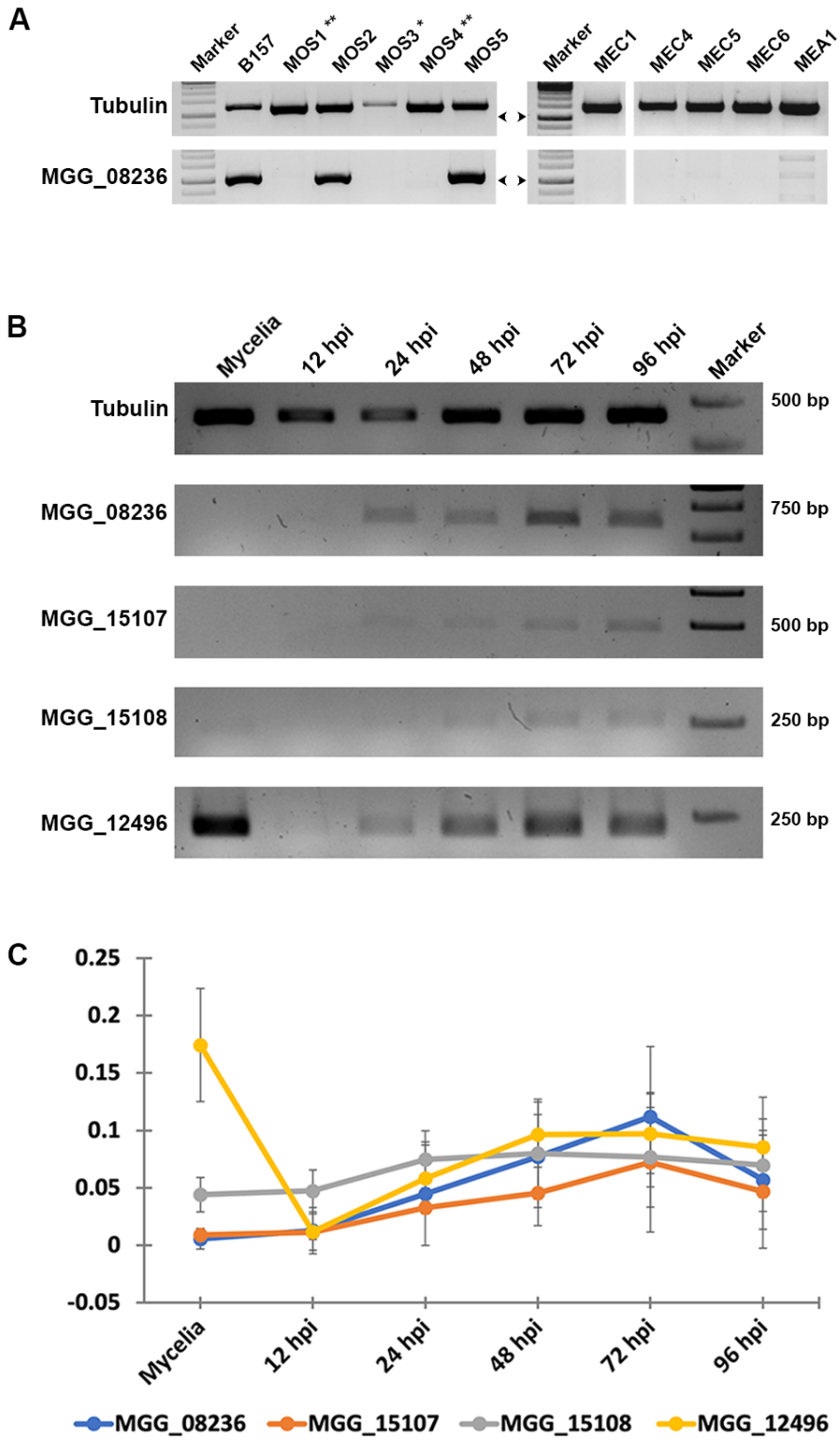
A
Marker
B157
MOS1 **
MOS2
MOS3 *
MOS4 **
MOS5
Marker
MEC1
MEC4
MEC5
MEC6
MEA1
Tubulin
MGG_08236
B
Mycelia
12 hpi
24 hpi
48 hpi
72 hpi
96 hpi
Marker
Tubulin
500 bp
MGG_08236
750 bp
MGG_15107
500 bp
MGG_15108
250 bp
MGG_12496
250 bp
C
0.25
0.2
0.15
0.1
0.05
0
-0.05
Mycelia
12 hpi
24 hpi
48 hpi
72 hpi
96 hpi
MGG_08236
MGG_15107
MGG_15108
MGG_12496

Figure 5.18: BGC-O1 genes are specifically expressed during pathogenic development. A) Gel electrophoresis of PCR products displaying presence or absence of ORF region of rPKS MGG_08236 from genomic DNA of the indicated *M. oryzae* strains. B157, MOS2, MOS3 and MOS5 belong to the *Oryza* lineage; whereas, the MOS1, MEC1, MEC4, MEC5, MEC6 and MEA1 belong to the *Eleusine* lineage. * - the only *Oryza* strain that lacked the BGC-O1 in in-silico analysis. ** - the *Oryza* strains placed outside *Oryza* lineage. Arrowheads corresponds to the 1 Kb size of band from Marker. B) RT-PCR gel depicting expression of genes MGG_08236 (Polyketide synthase), MGG_15107 (Methyl transferase), MGG_15108 (Co-A transferase) and MGG_12496 (Cytochrome P450) in *Oryza*-specific strain B157 at different stages of barley infection (12 to 96 hpi) and during vegetative growth (mycelium) in complete medium. C) Quantification of the expression of BGC-O1 genes. The intensity of each band was measured using ImageJ, and relative gene expression was calculated relative to that of β-tubulin (MGG_00604) as an endogenous control. The data on expression of MGG_08236 represents mean ± standard deviation of mean (SDM) from three independent biological experiments. Data on expression of the tailoring genes (MGG_15107, MGG_15108 and MGG_12496) represent observation from a single experiment.

These results indicate that the MGG_08236 PKS gene is specifically expressed during pathogenesis and potentially has a key role to play during host colonization. Additionally, our findings suggest that the BGC-O1 likely comprises only two co-regulated tailoring genes.

Altogether, our in-silico analyses identified a novel PKS gene cluster in the *Oryza*-specific lineage, which likely played a key role in shaping specialization of the blast fungus to rice host.