Chapter 4

Results:

In silico analyses of known effectors of the blast fungus in representative Indian field strains of *M. oryzae*

4.1 Whole genome sequencing using Illumina next generation sequencing platform

In this study, we sequenced 15 representative field isolates of *M. oryzae* that were collected from infected rice, finger millet and foxtail millet plants grown in different parts of India. The next generation sequencing was performed for whole genome using a paired-end sequencing approach at >50X depth on Illumina HiSeq2500 platform at AgriGenome Labs Pvt. Ltd., India. Libraries for Illumina paired-end sequencing were prepared using NEBNext® Ultra DNA Library Prep Kit. The raw reads obtained from the sequencing facility for the quality using FastQC for various parameters (per base sequence quality, per tile seq quality, per seq quality scores, per base seq content, per seq GC content, per base N content, seq length distribution, seq duplication levels, overrepresented seq, adapter content, Kmer content). The raw reads were processed using Trimmomatic at a threshold for the minimum read length of 80 bp and the quality score 20. The raw sequence data statistics are summarized in **Table 4.1**.

4.2 Genome-wide variations in host-specific lineages of *M. oryzae*

To assess the genomic relatedness among the different strains, variant calling was conducted. Single Nucleotide Polymorphisms (SNPs) and Insertion/Deletions of bases (InDels) are the footprints of the genetic variations, and also the important mechanisms to generate genetic divergence across the pathogen population. A phylogenetic tree was constructed using a total of 544,643 SNP sites identified from the coding regions of the genomes, from all the strains used in this study. The topology of the tree suggested the population divergence into three lineages. Interestingly, these lineages were not diverged based on the geographic location, from where the strains have been isolated, but are corelated to their host of origin **(Figure 4.2)**. These lineages are referred as *Oryza, Eleusine* and *Setaria* based on their host preferences. Principal component analysis (PCA) based on genome-wide SNPs was also performed to further assess the population structure of the different host-specific strains. The interesting observation made was that the strains are clustered in majorly three groups, which correlate with the three host plant species **(Figure 4.3)**. This finding is supported with various other reports suggesting the divergence of *M.*

oryzae species is host-driven (Chiapello et al., 2015; Gladieux, Condon, et al., 2018; Zhong et al., 2016).

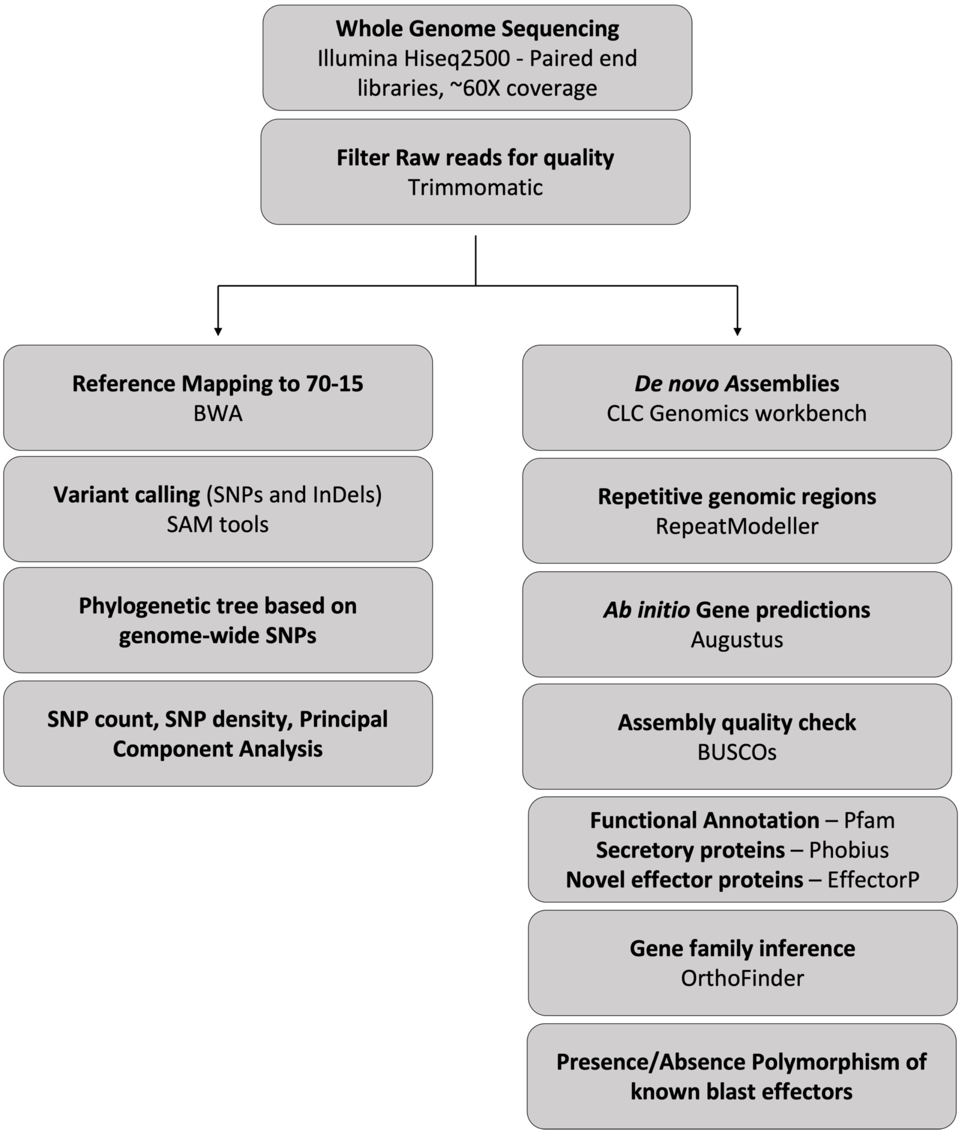
Whole Genome Sequencing
Illumina Hiseq2500 - Paired end libraries, ~60X coverage
Filter Raw reads for quality
Trimmomatic
Reference Mapping to 70-15
BWA
Variant calling (SNPs and InDels)
SAM tools
Phylogenetic tree based on genome-wide SNPs
SNP count, SNP density, Principal Component Analysis
***De novo* Assemblies**
CLC Genomics workbench
Repetitive genomic regions
RepeatModeller
***Ab initio* Gene predictions**
Augustus
Assembly quality check
BUSCOs
Functional Annotation – Pfam
Secretory proteins – Phobius
Novel effector proteins – EffectorP
Gene family inference
OrthoFinder
Presence/Absence Polymorphism of known blast effectors

Figure 4.1: Workflow depicting the analyses performed on whole genome sequencing data

Table 4.1: Summary of Raw reads obtained from Illumina sequencing method.

Strain	Synonyms	Read orientation	Mean Read Quality	Number of Reads	% GC	% Q < 10	% Q 10-20	%Q 20-30	% Q > 30	Number of Bases (Mb)	Mean Read Length (bp)	% Trimmed Paired Reads
MOS1	OSBC1	R1	35.74	13745493	48.69	1.38	1.65	5.42	91.55	1374.55	100.00	74.84
		R2	34.22	13745493	48.77	3.50	3.34	8.19	84.96	1374.55	100.00	
MOS2	GR11BC1	R1	35.74	11857430	49.04	1.33	1.64	5.47	91.56	1185.74	100.00	75.05
		R2	34.20	11857430	49.12	3.48	3.36	8.28	84.88	1185.74	100.00	
MOS3	MO-EI-23	R1	39.22	20102877	50.38	0.00	1.43	2.87	95.69	2010.29	100.00	50.66
		R2	32.92	20102877	50.70	0.05	13.53	14.08	72.33	2010.29	100.00	
MOS4	OS-KK-L1.1A	R1	39.24	33345753	51.14	0.00	1.43	2.86	95.71	3334.58	100.00	66.4
		R2	35.26	33345753	51.02	0.01	8.86	9.97	81.16	3334.58	100.00	
MOS5	OS-ULNSK-N2.3	R1	39.21	24196135	51.79	0.00	1.44	2.91	95.65	2419.61	100.00	51.41
		R2	32.95	24196135	52.17	0.05	13.40	14.09	72.46	2419.61	100.00	
MOS6	OS-GWSK-N2.1	R1	39.26	39184263	50.75	0.00	1.39	2.76	95.85	3918.43	100.00	70.05
		R2	35.94	39184263	50.76	0.02	7.10	9.22	83.66	3918.43	100.00	
MEC1	PR202-B1.1	R1	35.56	15231230	48.93	1.38	1.72	5.99	90.90	1523.12	100.00	75.02
		R2	34.23	15231230	48.97	3.28	3.29	8.46	84.97	1523.12	100.00	
MEC2	GPU48-C2.2	R1	39.12	37755314	50.56	0.00	1.60	3.03	95.37	3775.53	100.00	67.08
		R2	35.79	37755314	50.44	0.02	7.42	9.56	83.00	3775.53	100.00	
MEC3	EC-GKVK-L3	R1	38.91	18867499	50.12	0.00	1.85	3.56	94.59	1886.75	100.00	50.89
		R2	33.07	18867499	50.16	0.01	13.37	14.09	72.54	1886.75	100.00	
MEC4	EC-GKVK-F1.1	R1	38.91	19573969	51.58	0.01	1.79	3.55	94.66	1957.40	100.00	53.23
		R2	33.04	19573969	51.76	0.02	13.46	13.54	72.98	1957.40	100.00	

MEC5	EC-DWSK-F3.a	R1	37.43	25119154	50.29	0.00	3.92	6.53	89.55	2511.92	100.00	50.82
		R2	33.75	25119154	50.27	0.28	11.41	12.81	75.50	2511.92	100.00	
MEC6	EC-SSK-N8.1	R1	38.99	25172795	51.27	0.00	1.71	3.40	94.88	2517.28	100.00	49.23
		R2	32.39	25172795	51.36	0.01	14.95	15.09	69.95	2517.28	100.00	
MEA1	WR-L1.1	R1	36.10	13466070	50.73	1.09	1.25	4.52	93.14	1346.61	100.00	80.34
		R2	34.93	13466070	50.76	2.71	2.54	6.63	88.12	1346.61	100.00	
MSI1	FXM1-L3.1.1.a	R1	39.10	31725340	50.24	0.00	1.63	3.06	95.31	3172.53	100.00	57.06
		R2	34.12	31725340	50.22	0.03	10.59	12.62	76.76	3172.53	100.00	
MSI2	FXM3-L2.2	R1	35.40	14262686	50.40	1.69	2.02	6.04	90.25	1426.27	100.00	75.12
		R2	34.18	14262686	50.43	3.54	3.32	8.22	84.93	1426.27	100.00	

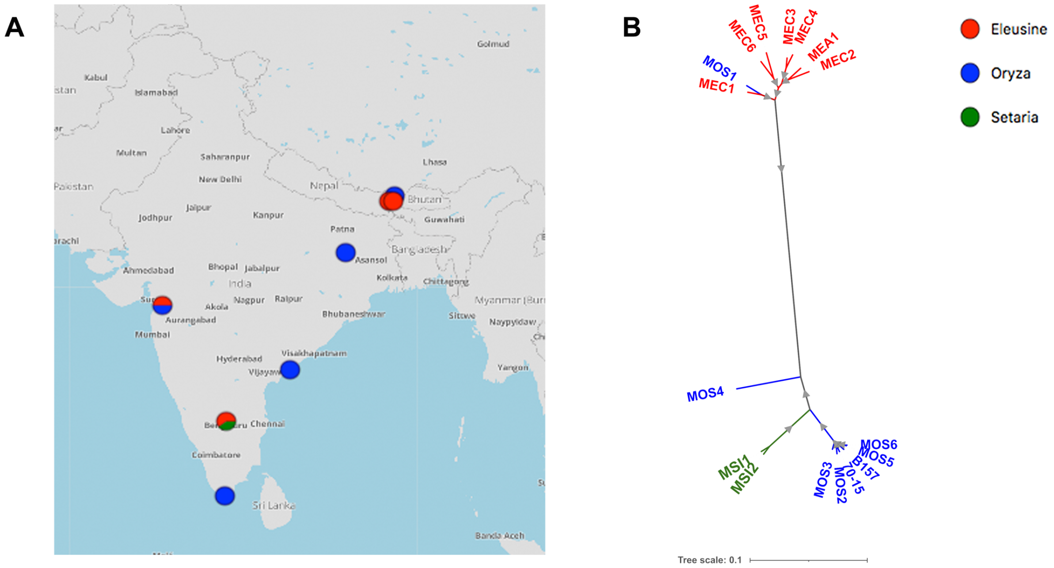

Figure 4.2: Genetic divergence of Indian population of *M. oryzae*. (A) Geographic distribution of the *M. oryzae* strains sequenced in this study. (B) Phylogenetic tree based on the total SNPs identified from each genome. Strains belonging to the different lineages are color coded accordingly. Branches with bootstrap value 100 are shown with grey triangles.

Exceptionally, MOS1, which is originally isolated from rice host, was found to be grouped with the isolates from *Eleusine* (finger millet) lineage. The whole-plant infection assay using the MOS1 strain showed differential virulence pattern, where inoculated finger millet and rice plants developed highly-susceptible and moderately-resistant lesions, respectively **(Fig. 3.17A)**. This indicates that MOS1 is most likely an *Eleusine*-infecting/adapted strain; but was isolated from infected rice leaf. Similarly, MOS4 strain was also isolated from infected tissue of another rice plant grown in field; however, it induced moderately-resistant lesions on the rice cultivar CO39 in our whole plant infection assays. Intriguingly, this strain was neither placed with *Oryza* lineage in the phylogenomic tree nor did show any genetic similarity with any of the other host-specific lineages.

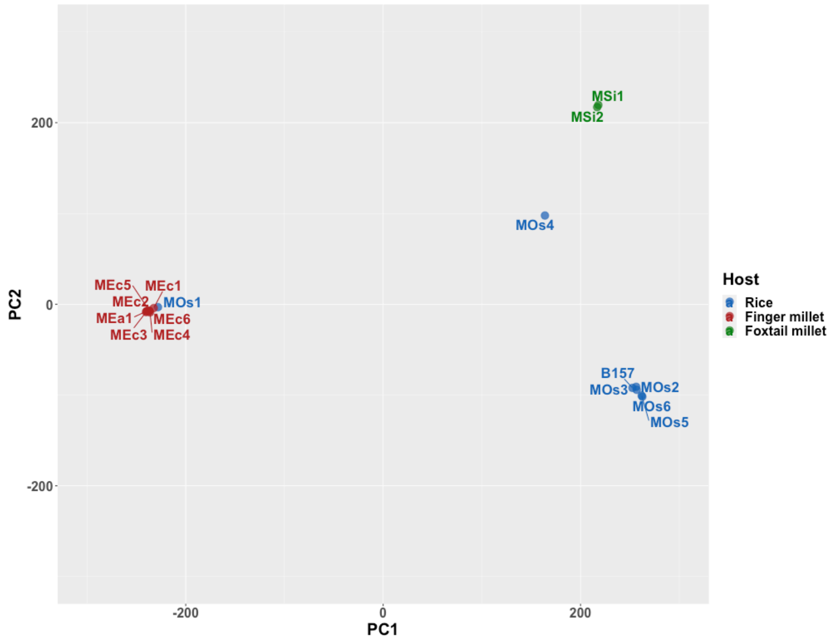

Figure 4.3: Principal component analysis (PCA) based on total SNPs identified from various *M. oryzae* strains. Strains are color-coded according to their host-of-origin.

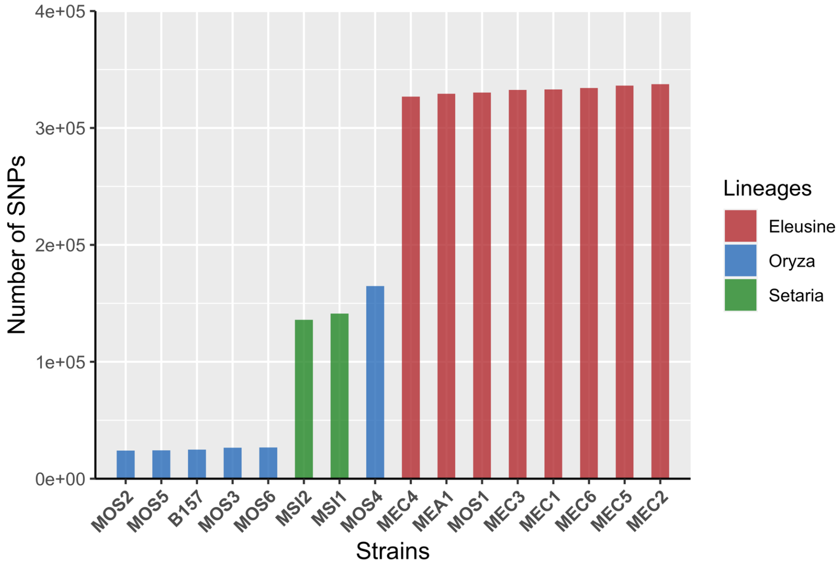

Figure 4.4: Genomic comparison of total SNPs number in genomes belonging to different lineages. X-axis represents the strains compared and Y-axis represents the total count of SNPs in each genome. Genomes are color-coded according to the lineages they belong to.

The multi-sample variant calling yielded an average SNPs count of 25294, 332452 and 138615 for *Oryza, Eleusine* and *Setaria* lineages*,* respectively. Similarly, the analysis yielded average InDels 5417, 39120 and 18247 for the three respective host-specific lineages, *Oryza, Eleusine* and *Setaria*. It is clearly demonstrated that the *Eleusine* lineage has the highest number of variants **(Figure 4.4)**, followed by *Setaria* lineage, these results are concordant with the earlier reported variant analysis (Shirke et al., 2016; Zhong et al., 2016).

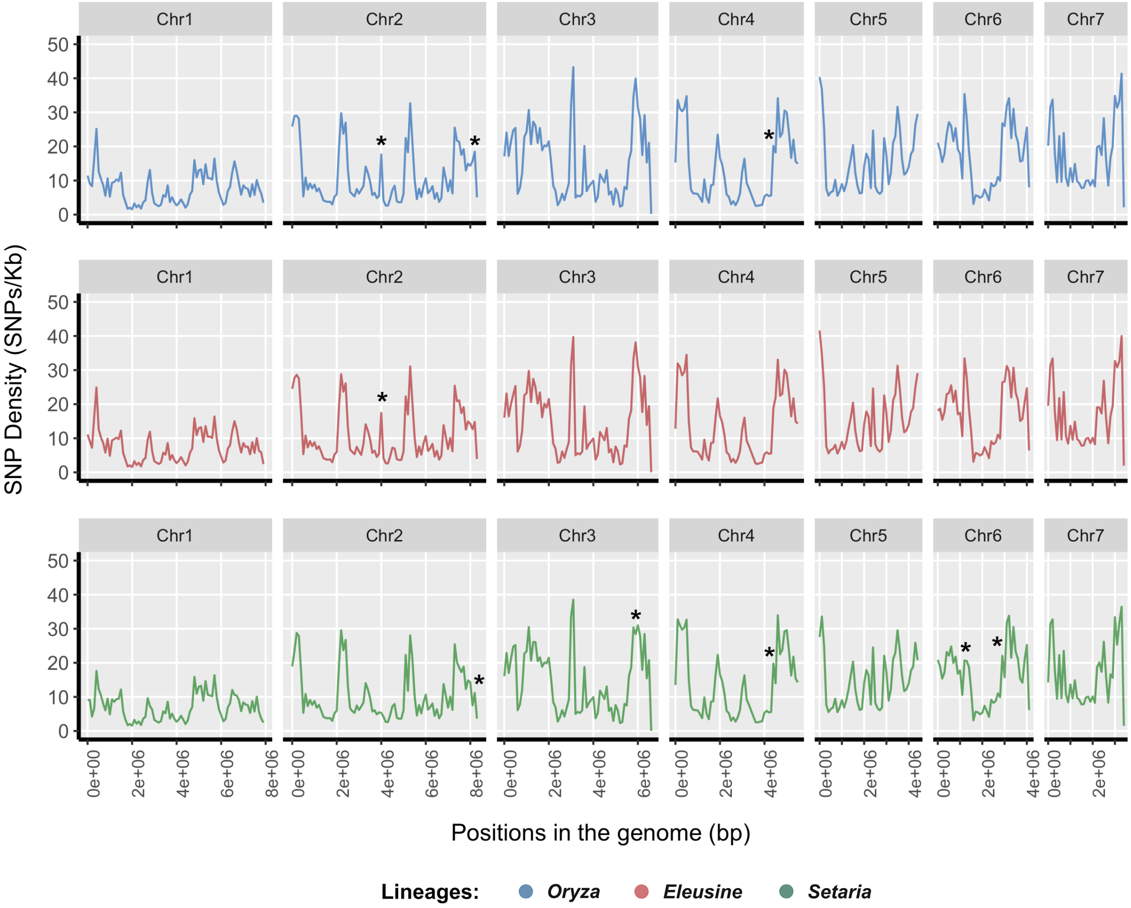

Figure 4.5: Whole genome distribution of SNP density shown in the different chromosomes. X-axis represents the positions on different chromosomes of reference genome 70-15 and Y -axis represents the SNP density calculated per 100 Kb window.

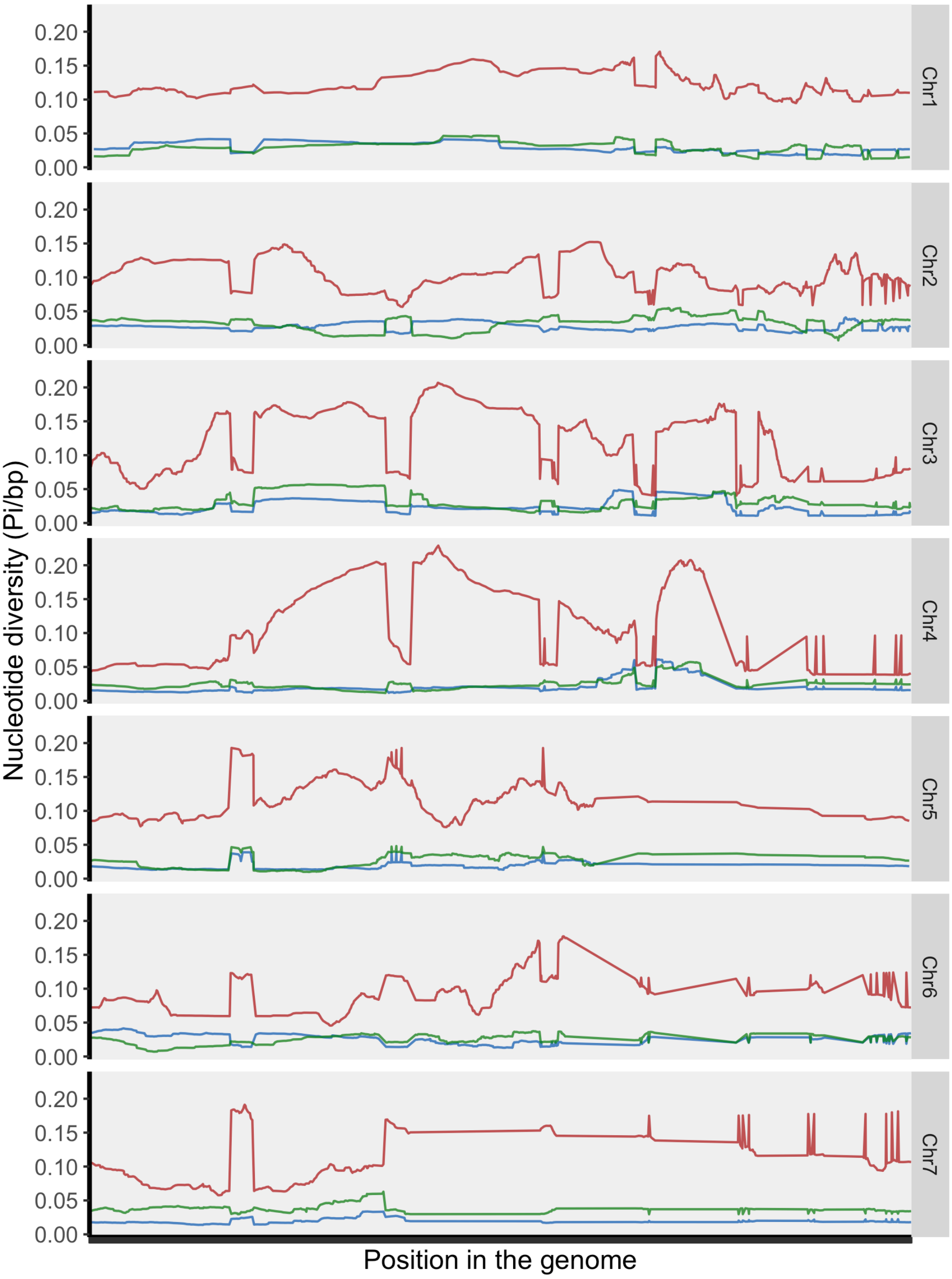

Figure 4.6: Whole genome distribution of Nucleotide diversity measured as Pi per bp, shown along the length of different chromosomes. X-axis represents the positions on different chromosomes of reference genome 70-15 and Y -axis represents the Nucleotide diversity calculated per 100 Kb window.

Although, We did not find any significant differences in the SNP density when compared among different lineages. The SNP density (SNPs per Kb) was measured with sliding window of 100 Kb across the lengths of all the chromosomes. Only a few genomic loci showed differences in the SNP density (marked with *, **Figure 4.5**).

Pi is a measure of nucleotide diversity and is defined as the average number of nucleotide differences per site between two DNA sequences in all possible pairs in the sample population (Nei & Li, 1979). Interestingly, the average nucleotide diversity (Pi per bp) showed significant differences - 0.024, 0.111 and 0.030 for *Oryza, Eleusine* and *Setaria*, respectively - which correlated with the SNP count data. Measure of nucleotide diversity was significantly higher in *Eleusine* as compared to *Oryza* and *Setaria* lineages **(Figure 4.6)**, which in-turn supports the larger genetic distances in the phylogenetic tree, and suggests that rice-infecting *Oryza* lineage originated as a result of a host shift from *Setaria* lineage, during crop-domestication (Gladieux, Ravel, et al., 2018; Zhong et al., 2016).

Overall, variant calling suggest how some of the SNPs or InDels could be one of the factors altering the functionality of the virulence or host-specific genes. A detailed analysis of such variants in pathogenesis related genes might be useful in studying the evolutionary mechanism of host jump and in identification of genetic markers to assess various *Magnaporthe* pathotypes in field conditions.

4.3 Genome assembly and gene predictions

Trimmed reads were processed for the de novo assembly. De novo sequence assemblies were constructed using CLC Genomics Workbench. The assembly statistics for each strain sequenced in this study is summarized in **Table 4.2**. The number of scaffolds ranged from 1727 to 2959. The N50 ranged from 41-91 Kb across all the strains indicating good quality of assemblies. The average genome sizes of the strains belonging to *Oryza, Eleusine* and *Setaria* host-specific lineages were found to be 38.83, 40.16 and 38.18 Mb respectively.

The resulted de novo assemblies were screened for repeats present in the each genome. Repeats usually are masked in the genome in order to avoid spurious gene predictions. The masking is of two kind – hard mask, where each nucleotide of the identified repeats is transformed to ‘N’ or ‘X’, and soft mask where each nucleotide is transformed to a lower case (a, t, c, g). We used RepeatModeller to identify and classify the repeats in all the genome assemblies and soft-masked using RepeatMasker (Smit et al., 2015; Smit & Hubley, n.d.). The summary of repetitive sequences obtained summarized in **Table 4.3**.

The genome sizes of all 15 strains ranges from ~37.9 Mb to ~40.7 Mb, the differences are mainly due to variation in content of repetitive sequences, which is 5.93% in MEC1 and 2.53% in MOS2 **(Figure 4.7)**. It seems that the repeat content is on higher range in the genomes from finger millet (*Eleusine*) host plants **(Table 4.3)**. These repeat percentages are comparable to previously reported genomes from *M. oryzae* for Illumina reads (Chiapello et al., 2015; Gómez Luciano et al., 2019; Gowda et al., 2015; Shirke et al., 2016). Assemblies achieved using short-read sequencing technology such as Illumina, tend to break in repeat-rich regions, so it is possible to identify a less proportions of regions comprising repeats using such assemblies.

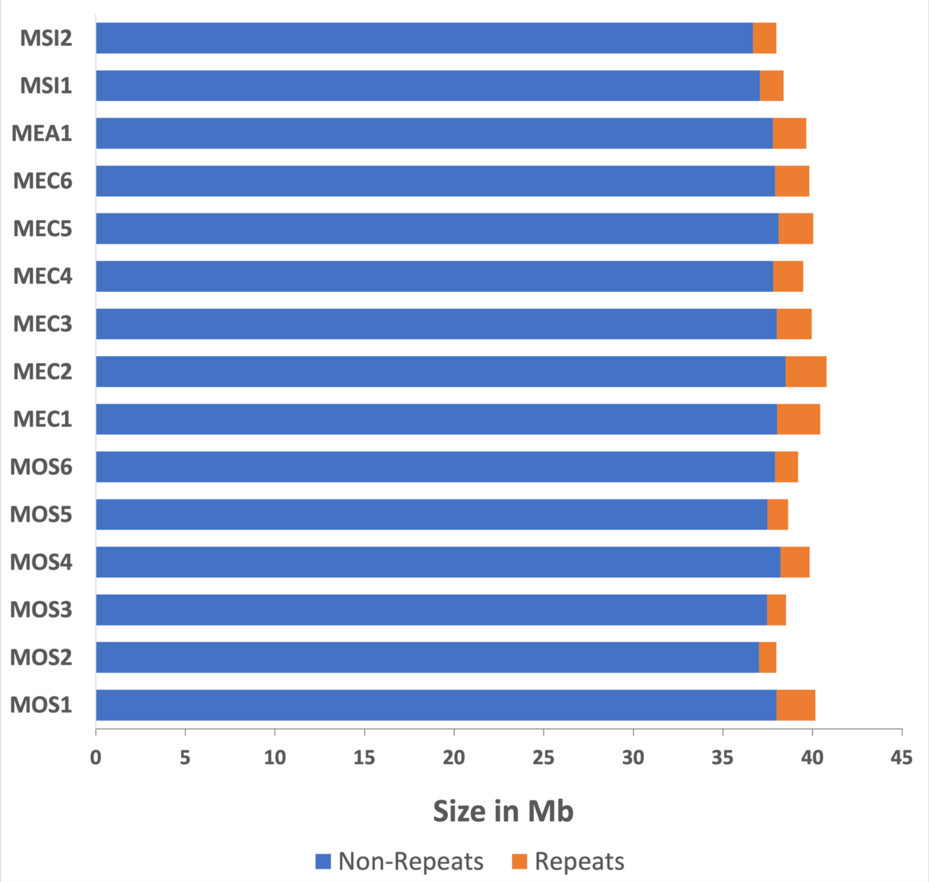

Figure 4.7: Content of repetitive elements in the de novo assembled genomes from each strain. The solid blue portion denotes the total size of non-repetitive region in each genome, whereas the orange portion denotes the repetitive DNA sequences in the each genome.

The major proportion of repetitive sequences consists of transposable elements (TEs) in *M. oryzae*. The significant TEs of blast fungus are LTR retro transposons such as Pyret, MGLR3, MAGGY, Grasshopper; DNA transposons Pot2 and Occan; and SINE like elements Mg-SINE (Dobinson et al., 1993; Farman et al., 1996; Kachroo et al., 1994, 1995; Kang, 2001; Kito et al., 2003; Nakayashiki et al., 2001). *Oryza* lineage seemed to have overall highest copy numbers of TEs in their genomes (**Figure 4.8**). Mg-SINE found to be in low copy number in other lineages as compared to *Oryza* lineage (with an average of 133 copies) with exception in MOS3 strain, contrary to the finding that Mg-SINE found to be present in ~100 copies in both rice and non-rice infecting isolates of *M. oryzae* (Kachroo et al., 1995). Pot3 was also found to be present significantly in high copies in strains from *Oryza* lineage as reported earlier (Couch, 2005). Grasshopper TE found to be absent in most strains (except for MOS4 and MOS6) belonging to *Oryza* lineage. Grasshopper identified to be exclusively present in multiple copies in *M. oryzae* strains infecting finger millet and goosegrass (*Eleusine* sp.) and has been suggested to be acquired after the evolution of *Eleusine* host-specific lineage (Dobinson et al., 1993).

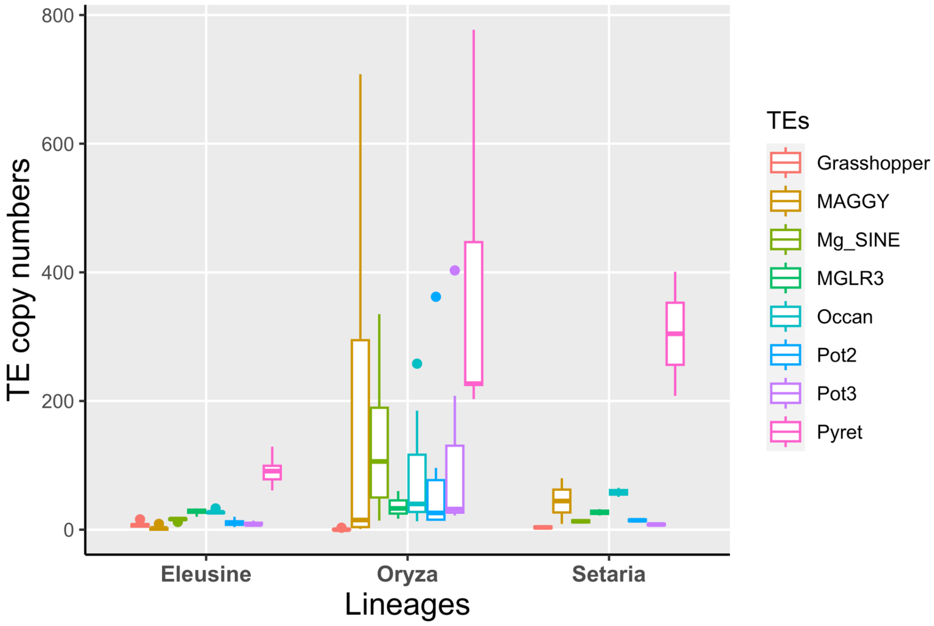

Figure 4.8: Distribution of *M. oryzae* specific transposable elements in various host-specific lineages. Box-plot shows the range of copy numbers present in the different strains belonging to each lineage.

Table 4.2: De novo assembly statistics and gene prediction content of *M. oryzae* strains sequenced in this study.

Host of isolation	**Strains**	**Tissue**	**Illumina PE reads (millions)**	**Coverage**	**Number of scaffolds**	**Assembly size (Mb)**	**Largest scaffold (bp)**	**N50 (bp)**	**N75 (bp)**	**BUSCO (%)**	**Repeat (%)**	**GC (%)**	**Number of genes**
Rice	**MOS1**	Leaf	18208538	51x	1904	40.16	281161	65340	33132	97	5.41	50.65	10709
	MOS2	Leaf	15478674	47x	2047	37.97	367257	72410	29985	97	2.53	51.52	10545
	MOS3	Leaf	17205136	53x	2767	38.52	267001	50541	21849	96.9	2.73	51.4	10811
	MOS4	Leaf	38912430	110x	2355	39.84	560096	91955	35479	97.1	4.1	51.03	10889
	MOS5	Neck	21798984	64x	2719	38.63	339207	52164	22677	97	2.94	51.41	10864
	MOS6	Neck	46532296	140x	2781	39.20	510671	74458	29003	96.9	3.27	51.24	10767
Finger millet	**MEC1**	Leaf	20560638	56x	2071	40.42	286458	33770	17282	96.9	5.93	50.61	10707
	MEC2	Leaf	44918978	124x	2959	40.78	2525283	51730	24248	96.9	5.6	50.55	10869
	MEC3	Leaf	17427662	48x	2342	39.95	396207	48482	23772	96.8	4.86	50.75	10744
	MEC4	Finger	18641988	53x	2220	39.48	291154	50680	24788	96.8	4.22	50.88	10705
	MEC5	Finger	22619870	63x	2879	40.04	241515	41163	20564	96.4	4.8	50.76	10988
	MEC6	Neck	22624154	62x	2271	39.81	294595	50102	24357	96.8	4.77	50.85	10720
	MEA1	Leaf	19463120	54x	2088	39.65	364278	57051	30038	96.9	4.67	50.86	10667
Foxtail millet	**MSI1**	Leaf	31336604	94x	2393	38.38	199508	45371	23183	96.8	3.44	51.3	10643
	MSI2	Leaf	18569022	55x	1727	37.97	318537	51176	27865	96.7	3.4	51.37	10496

Table 4.3: Proportions of different types of repetitive elements in *M. oryzae* genomes.

Strains	Repeats type										
	LINEs	**LTRs**	**DNA transposons**	**Rolling-circles**	**Unclassified**	**Total Interspersed Repeats**	**Small RNA**	**Satellites**	**Simple Repeats**	**Low Complexity**	**Total (%)**
MOS1	0.12	1.68	0.28	0	2.12	4.2	0	0	1.02	0.19	5.41
MOS2	0	0.43	0.03	0	0.89	1.36	0	0	1	0.17	2.53
MOS3	0	0.58	0.06	0	0.9	1.54	0	0.02	1	0.17	2.73
MOS4	0	1.42	0.25	0	1.29	2.95	0.01	0	0.98	0.17	4.1
MOS5	0.01	0.77	0.06	0	0.91	1.74	0.01	0	1.01	0.17	2.94
MOS6	0	0.86	0.06	0	1.15	2.07	0	0	1.03	0.18	3.27
MEC1	0.08	1.65	0.29	0	2.66	4.69	0.01	0.03	1.02	0.18	5.93
MEC2	0.01	1.84	0.37	0.07	2.1	4.32	0	0	1.03	0.18	5.6
MEC3	0.05	1.25	0.27	0	2.04	3.62	0.01	0	1.05	0.18	4.86
MEC4	0.02	1.15	0.27	0	1.56	3	0	0	1.05	0.18	4.22
MEC5	0.02	1.56	0.16	0	1.84	3.59	0	0	1.04	0.18	4.8
MEC6	0.05	1.29	0.25	0.11	1.85	3.44	0	0	1.04	0.17	4.77
MEA1	0.03	1.42	0.3	0	1.69	3.44	0.01	0	1.03	0.19	4.67
MSI1	0	1.02	0.09	0	1.1	2.21	0	0	1.06	0.18	3.44
MSI2	0	0.84	0.06	0	1.26	2.16	0	0	1.06	0.18	3.4

We carried out *ab initio* gene predictions on all the soft-masked assemblies using Augustus (Stanke et al., 2006) with *Magnaporthe_grisea* as a species model. The total number of genes predicted ranged from 10496 to 10988 in MSI2 and MEC5, respectively **(Table 4.2)**. Quality assessment of resulted assemblies were often limited to the measures like N50, although a new measure for quantitative assessment of genome assemblies and completeness of the genes predicted is based on evolutionarily informed expectation of gene content – a comprehensive datasets of Benchmarking Universal Single-Copy Orthologs (BUSCO; Simão et al., 2015). We evaluated the overall quality of each genome assembly using BUSCO v5.2.2 with the Sordariomycetes dataset. BUSCO score for each genome was found to be >96%, which indicated high-quality gene prediction and completeness of the assemblies **(Table 4.2)**. Thus, overall newly sequenced genomes have been assembled with overall completeness and these highly contiguous assemblies from Indian host-specific strains of *M. oryzae* could serve as a good base for the comparative analyses and studies on population structure.

4.4 Functional annotations of predicted genes

Predicted gene models from each genome then subjected to the functional annotation to understand their role in various biological processes. We used Pfam database to predict the functional domains of the proteins. Phobius (Käll et al., 2004) was used to identify the secretory proteins using amino acid sequences of predicted genes as input. Phobius predicts the transmembrane protein topology and signal peptide together. With the help of Phobius, a significant reduction of false classification is achieved as compared to TMHMM/SignalP (Käll et al., 2004). Average number of 1660 secretory proteins have been identified, which comprises of 15.45 % of total gene models predicted **(Figure 4.9)**. Out of these secretory proteins, average 651 novel effectors-like proteins were predicted using EffectorP **(Table 4.4)**. Major portion of these novel effector-like proteins is unique to blast fungus, and such high number suggest that the effector gene turn-over occurs frequently in blast fungus.

The homologs of known blast effector proteins were identified using a dataset comprising known effectors (Gómez Luciano et al., 2019). Resulted sets of homologs of known blast effector proteins suggested that only ~0.15 % of total protein pool consisted in the effector repertoire per genome. Large portion of the novel putative effector repertoire requires the attention and needs to be characterized, in order to understand the various mechanisms of host-specialization.

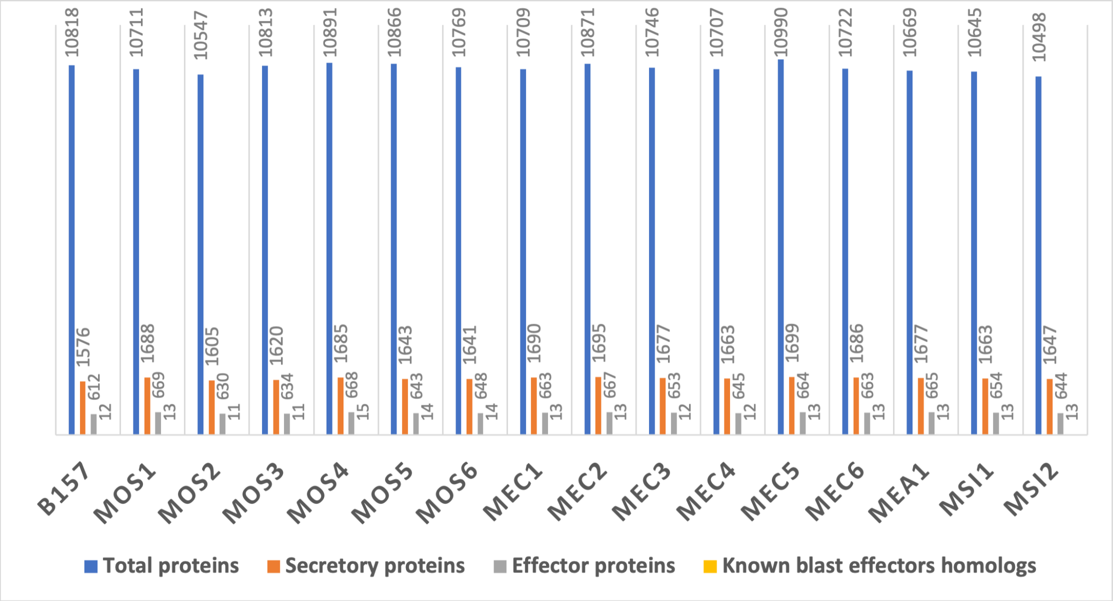

Figure 4.9: Proportions of secretory proteins and effectors proteins predicted based on functional annotation. The number of different kind of proteins are denoted as colored bars. Total proteins – blue, Secretory proteins – orange, Effector proteins – grey and Homologs of known blast effectors – yellow.

4.5 Lineage-specific gene families in the blast fungus

Clustering of genes into orthologous sets can reveal unique sets of genes that are important to one species or group of fungi that are not found in other species. The gene families were inferred by identifying orthologous groups by clustering the total proteome (171940 predicted proteins) comprised of 16 field strains of *M. oryzae*. The total number of orthogroups achieved was 11661, out of which, 26 were species-specific orthogroups. As shown in **Figure 4.10**, all the three lineages of strains have a total of 9353 orthogroups in common, thus termed as core set of genes, comprising ~90% of genes from each genome. 8933 proteins were found to belong to single-copy orthogroups. The orthologous groups

Table 4.4: Total number of novel putative effectors predicted in different field strains of *M. oryzae*

Strains	Secretory proteins	Putative Effectors	Cytoplasmic effectors	Apoplastic effectors
70-15	1949	1013	626	387
B157	1576	819	488	331
MOS1	1688	889	524	365
MOS2	1605	843	501	342
MOS3	1620	841	506	335
MOS4	1685	887	532	355
MOS5	1643	866	513	353
MOS6	1641	866	519	347
MEC1	1690	877	513	364
MEC2	1695	881	522	359
MEC3	1677	876	516	360
MEC4	1663	875	513	362
MEC5	1699	894	532	362
MEC6	1686	884	524	360
MEA1	1677	881	514	367
MSI1	1663	873	521	352
MSI2	1647	859	511	348

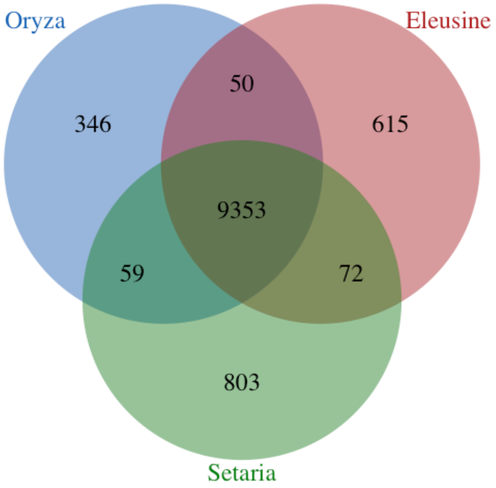

Figure 4.10: Venn diagram showing the lineage-wise clustering of proteomes. Intersection of the venn diagram depicts the sets of core proteome. Unique genes belonging to each host-specific lineage have been shown.

having more than two copies of genes in *Magnaporthe* isolates were considered as duplicated gene families. Approximately, 500 orthogroups were identified to comprise such duplications.

We identified lineage-specific orthogroups by considering single-copy orthogroups present only in that particular lineage. For *Oryza* lineage specific orthogroups, we considered all the orthogroups commonly shared by B157, MOS2, MOS3, MOS5 and MOS6 and absent in all the other isolates. Similarly, for *Eleusine* lineage specific orthogroups, all the orthogroups commonly shared by MEC1, MEC2, MEC3, MEC4, MEC5, MEC6 and MEA1 were considered. From our analysis, we identified 346, 615 and 803 genes specific to *Oryza, Eleusine* and *Setaria* lineages, respectively. Interestingly, MOS4, which is outside of *Oryza* lineage (Fig. 4.2 and 4.11), has 7 and 15 *Oryza* and *Setaria*-specific orthogroups, respectively. Whereas MOS1, although isolated from rice and found within *Eleusine* lineage, has 61 *Eleusine*-specific and only 1 *Oryza*-specific orthogroups. Most of these lineage-specific genes were found to have functions related to general metabolic processes and transmembrane transporter activities, although a few of the genes encoding enzymes involved in secondary metabolism and metalloendopeptidase activity. Especially, lineage-specific gene families are enriched with functional domains Cytochrome P450 and polyketide synthases (PKSs). Cytochrome P450 and PKSs are involved in various pathogenesis related processes, including host invasion and host-immunity modulation (Collemare, Pianfetti, et al., 2008; Huang et al., 2022; Jacob et al., 2017). Cytochrome P450 has been reported to be essential in host-adaptation as well (Durairaj et al., 2016). A well-known blast effector protein AVR-Pita belongs to Metalloproteases family, which has a role in plant–pathogen interaction in a gene-for-gene patho-system with the R protein Pita (Chuma et al., 2011). We further built a phylogenetic tree based on the presence absence variations (PAV) in a total of 1953 orthogroups detected (Fig. 4.11). The topology of this tree is very similar to the phylogenomic tree generated based on the SNPs dataset (Fig. 4.2), except for the swapping between MOS4 and *Setaria* lineage, moving the former slightly closer to the *Oryza* lineage. This observation suggests that some of the accessory genes in MOS4 are likely shared with those from the *Oryza* lineage.

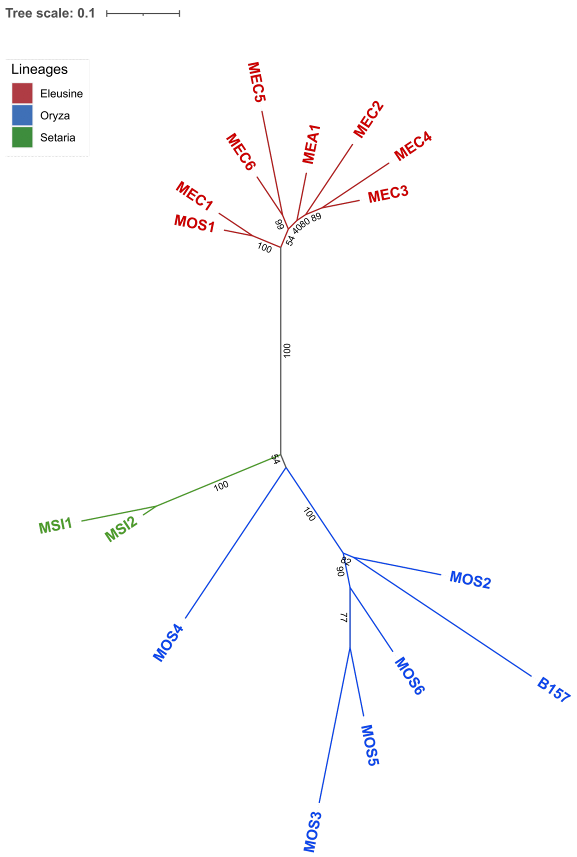

Figure 4.11: Phylogenetic tree based on PAV in accessory genes shows divergence of host-specific lineages of *M. oryzae*. The mid-point rooted phylogenetic tree is based on maximum-likelihood methods. Bootstrap values are indicated on the branches.

4.6 Distribution of cell wall degrading enzymes in various host-specific lineages

Cell wall degrading enzymes (CWDEs) are crucial virulence factors since they act on a primary barrier – plant cell wall, and thus helping fungal plant pathogens to invade the host successfully (Cosgrove, 2001). CWDEs (subclass: Endo-xylanases and cellulases) are significantly upregulated during plant infection and required for penetration and virulence in *M. oryzae* (Nguyen et al., 2011; Van Vu et al., 2012). Feruloyl esterases (Fae), a subclass of carboxylic acid esterases, also belong to one such group of CWDEs. Interestingly, the FAE gene family in *M. oryzae* is relatively expanded when compared to that in non pathogenic counterparts such as *Neurospora crassa* and *Aspergillus nidulans*, which have only one and three FAE genes, respectively (Dean et al., 2005).

A total of nine putative type B Fae sequences have been identified in *M. oryzae* and one of the Fae (Fae1) has been shown to play a crucial role, specifically in host invasion and tissue colonization (Thaker et al., 2022). Phylogenetic analysis of feruloyl estearses (FAEs) in different host-specific strains of *M. oryzae* isolated from various cereal crops such as rice (B157, MOS1, MOS2, MOS3, MOS4, MOS5, MOS6), finger millet (MEC1, MEC2, MEC3, MEC4, MEC5, MEC6, MEA1) and foxtail millet (MSI1, MSI2) was carried out using the annotated protein sequences from these isolates. Protein sequences of FAEs, in *M. oryzae* 70–15 strain, were used as the reference and *M. grisea* as an outgroup. Interestingly, our analysis showed that three *M. oryzae* Fae sequences, namely MGG_09404, MGG_09732, and MGG_08737 (Fae1), diverged likely in a host-specific manner (**Fig. 4.12;** Thaker et al., 2022).

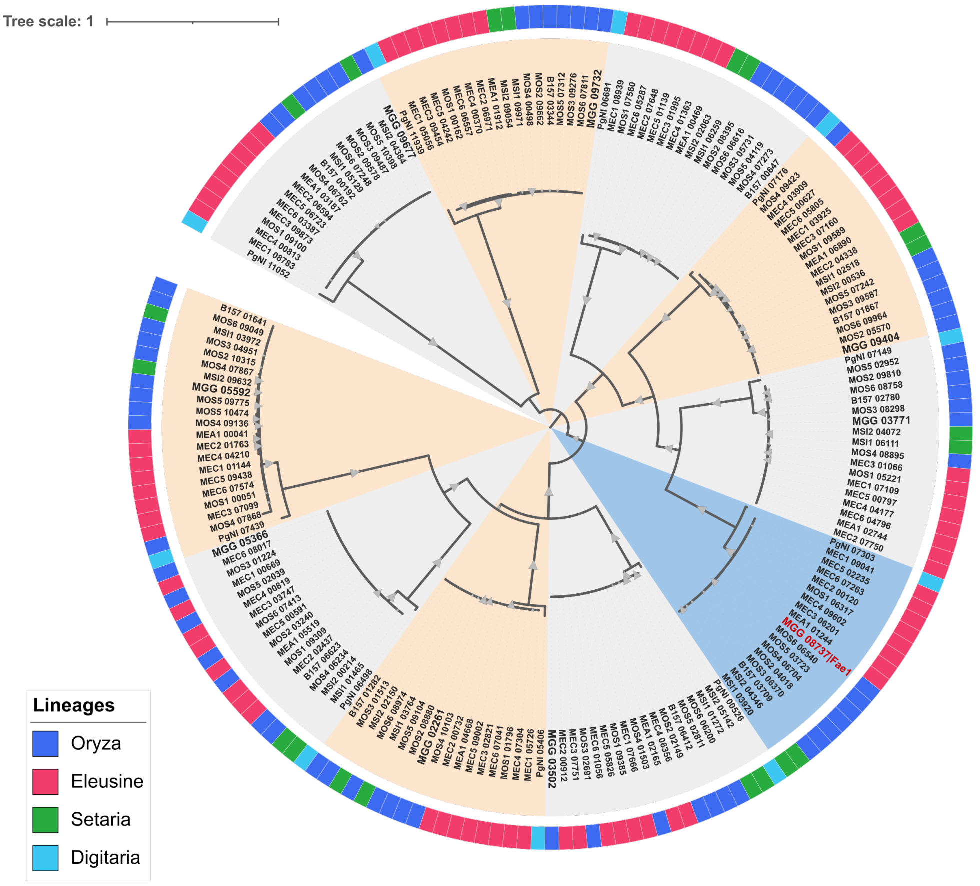

Figure 4.12: Phylogenetic analysis of FAE sequences from host-specific lineages of *M. oryzae*. The mid-point rooted phylogenetic tree is based on maximum-likelihood methods, with assessment of branch support by 1000 bootstrap replicates. Bootstrap values are indicated as grey triangles, sized according to the values (largest triangle being 100% bootstrap support). The outer circular strip is color-coded according to the lineages. The characterized Fae1 (MGG_08737) was labelled in the red color.

4.7 Distribution of known Blast Effectors in various host-specific lineages

Identification of homologs of Blat effectors was performed using Nucleotide BLAST with the earlier reported dataset comprising of 46 characterized Blast Effectors (Gómez Luciano et al., 2019). We analyzed Presence/Absence Variations (PAV) of known effectors to assess the correlation between the repertoire of effector molecules and divergence of host-specific lineages **(Figure 4.13)**. Notably, six (Ace1, Avr-Pi54, Avr-Pita3, Avr Piz-t, Avr-Pi9 and PWT3) out of 19 AVR genes found to be present in all the strains. Presence in all the strains irrespective of their infection ability indicates their potential role in the pathobiology of the fungus. Certain AVR genes (Avr-Pia, Avr-Pik, Avr-Pita1, PWL2 and PWT4) were found to be absent from all the strains of *Eleusine* and *Triticum* lineages. Avr-Rmg8, PWL3, PWL4 showed decrease in sequence identity in *Eleusine* lineage, indicating variations in those genes. Avr1-CO39 and PWL1 were found to be absent from most strains belonging to *Oryza* lineage except for MOS4 and MOS2 respectively. Loss of Avr1-CO39 has been reported in rice-infecting *Magnaporthe* strains (Couch, 2005). Interestingly, presence of AVR1-CO39 in MOS4 correlates with the finding that the placement of this strain outside the *Oryza* lineage in phylogenetic species tree, and its ability to infect rice with the moderately-resistant lesions.

Apart from the PAV analysis of known effectors, insertional mutations were also identified in a few strains. We found Tranposable Elements (TE) insertion in coding region of Avr-Pib and PWT3 genes in US71 and CD156 strains respectively. The 528 bps insertion in Avr-Pib was found to be classified as LTR-Copia1 and DNA transposon-Helitron TE types. On the other hand, PWT3 gene in CD156 carried 8961 bps of TE insertion, which consisted of various different classes of TEs such as LTR-Copia, Gypsy; DNA-hAT, Mariner, Polinton, Helitron and NonLTR-L1, Poseidon. Many such insertions have been reported in various effector genes belonging to different lineages found across a globe (Dong et al., 2015; J. Li et al., 2022; Peng et al., 2021; Wu et al., 2015).

These observation indicate that such insertion of TEs can disrupt the existing avirulence genes and lead to a functional loss. Such polymorphisms are resulted as a co-evolutionary arms-race with its host, thereby benefiting pathogen to evade the corresponding resistance mechanism, and help survive pathogen in the same host plants.

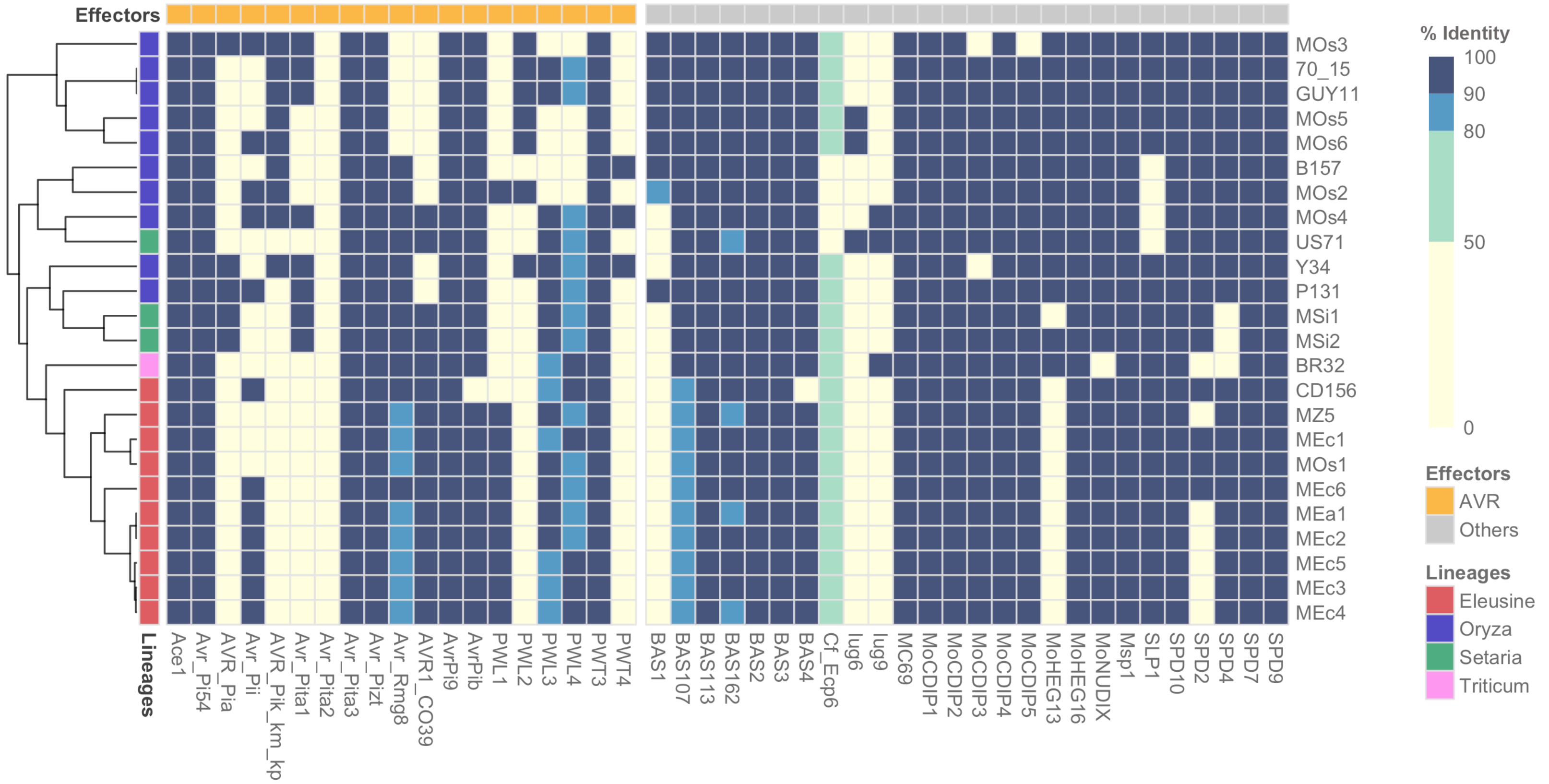

Figure 4.13: Presence/absence polymorphisms of known effector genes in various lineages of the blast fungus. Avirulence (Avr) effectors are depicted with orange bars. Color bar at right shows the % Identity to the known effector genes.