Large scale genome assemblies of *Magnaporthe oryzae* rice isolates from Italy

Joe Win[1], Adeline Harant[1], Angus Malmgren[1], Thorsten Langner[1], Ram-Krishna Shrestha[1], Sergio M. Latorre[2], Vincent Were[1], Nicholas J. Talbot[1], Hernán A. Burbano[2,3], Anna Maria Picco[4], Sophien Kamoun[1]

[1]The Sainsbury Laboratory, University of East Anglia, Norwich Research Park, Norwich, United Kingdom
[2]Research Group for Ancient Genomics and Evolution, Department of Molecular Biology, Max Planck Institute for Developmental Biology, Tübingen, Germany
[3]Centre for Life's Origins and Evolution, Department of Genetics, Evolution and Environment, University College London, London, United Kingdom
[4]Department of Earth and Environmental Sciences, University of Pavia, Pavia, Italy

We report long-range sequencing of nine rice-infecting *Magnaporthe oryzae* isolates from different rice-growing regions of Italy using Oxford Nanopore Technology. We aquired chromosome-level genome assemblies, polished with Illumina short reads, and removed mitochondrial sequences to improve the quality of the assemblies. We provide the genome assemblies to the public with open access.

Introduction

Magnaporthe oryzae (Syn. *Pyricularia oryzae*) is the number one fungal plant pathogen that poses a clear threat to global food security (Dean *et al.,* 2012; Fisher *et al.,* 2012). *M. oryzae* causes blast disease on rice, wheat and a range of other cereals including oat, finger millet and foxtail millet as well as wild grasses. *M. oryzae* is found all over the world wherever warm temperature and high humidity are common. This includes Italy which is the largest rice producing country in Europe. In Italy, *M. oryzae* is thought to be co-evolving with rice since its introduction to the country sometimes after the 15th century. We aim to test a hypothesis that *M. oryzae* has adapted to the rice cultivars and growing conditions employed in Italy, and in doing so, its genome has accumulated signatures of adaptation including genome-level structural variations. To generate the genomic data that would enable such analyses, we sequenced nine isolates of *M. oryzae* using a long-range sequencing strategy based on Oxford Nanopore Technology. We deposited these chromosome-level genome assemblies in public databases for open access.

Results

We sequenced nine *M. oryzae* isolates collected from different regions of Italy using PromethION flow cells (Table 1). We obtained N50 read lengths ranging from 23.4 to 32.3 kbp and total base counts of ~4.3 to 7.5 Gbp (~100x to 180x genome coverage). The sequence reads were assembled using Canu software (Koren *et al*., 2017). The number of contigs varied from 17 to 38 with the largest contigs in the ~10-11 Mbp range (Table 2).

To improve the quality of the assemblies, we acquired short sequence reads (Table 3) from the same fungal isolates using Illumina technology and polished the genomes using Pilon (Walker *et al.*, 2014) and Racon (Vaser *et al*., 2017) programs. We then removed the contigs in the assemblies that had high similarities to mitochondrial sequences (Table 4). We re-ordered the contigs in our assemblies to follow the chromosome structure of *M. oryzae* rice isolate 70-15 (Dean *et al*., 2005; Okagaki *et al*., 2015) by aligning our contigs to the 70-15 chromosomes using Mauve (Darling et al., 2004).

To assess the quality of the assemblies, we compared the assemblies of Italian isolates to the chromosome quality assembly of *M. oryzae* rice isolate 70- 15 by whole-genome sequence alignment using MUMmer 3 package (Kurtz *et al*., 2004). We noted significant co-linearities

between our Italian assemblies and that of 70-15 indicating that our nanopore assemblies are of acceptable quality (Figure 1).

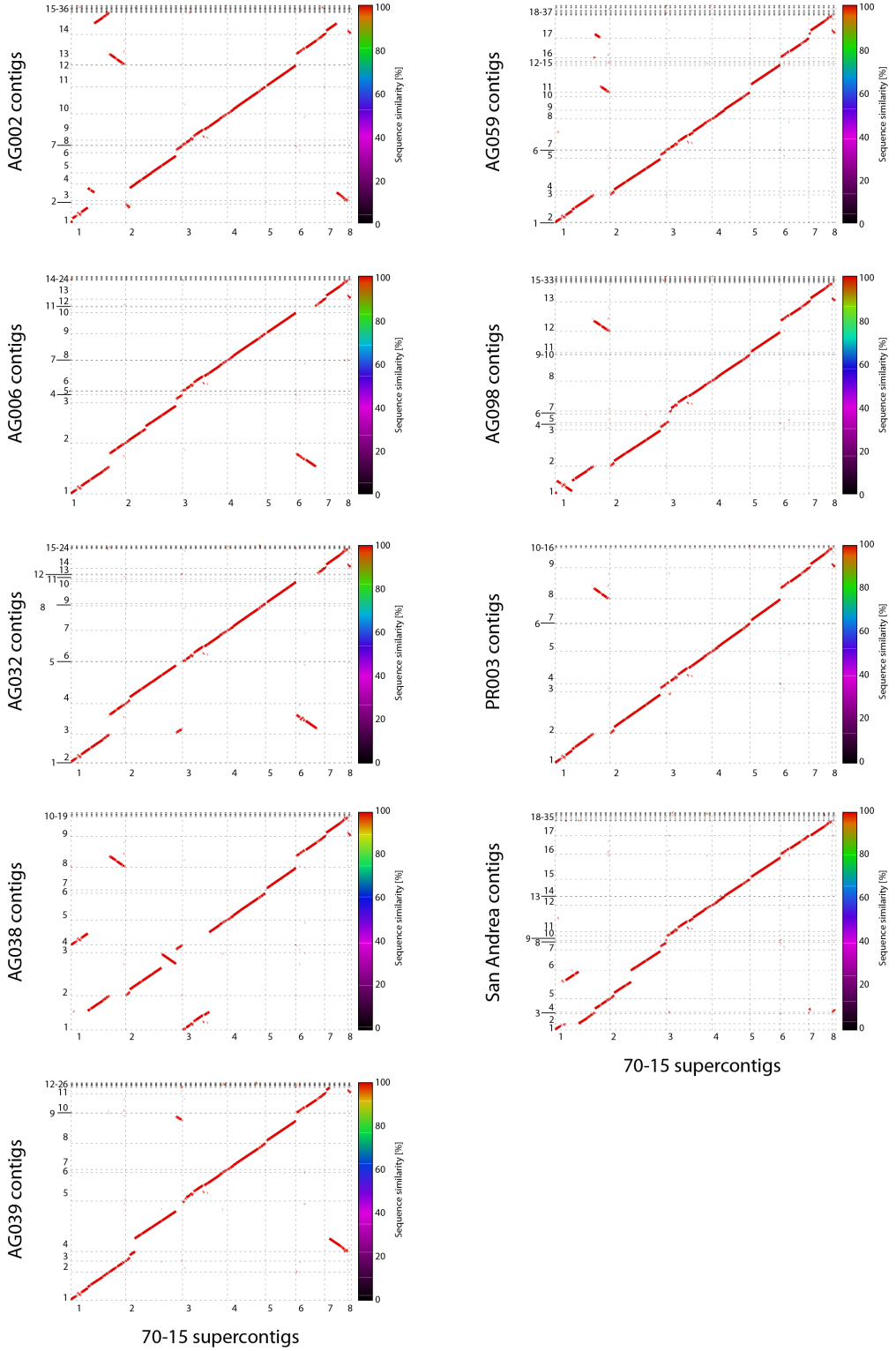

Figure 1. Genome assemblies of *Magnaporthe oryzae* Italian isolates are highly contiguous and colinear to the refererence genome of isolate 70-15. Colinearity plot of whole genome alignments between Italian rice blast isolates to the 70-15 reference genome. Contig IDs of Italian isolates are given on the Y-axis. Supercontig IDs of the reference genome are given on the X-axis. The color of the alignments shows the sequence similarity to 70-15.

To evaluate the completeness of the assemblies and any improvement in polishing process, we run Benchmarking sets of Universal Single-Copy Orthologs (BUSCO) (Seppey *et al*., 2019) using fungi_obd10 database for each genome assembly. We found that the polishing step improved the completeness of the assemblies from 72.2-83.4% before polishing to 97.7-98.8% after polishing, and gene fragmentation from 3.4-7.4% to 0-0.3% (Tables 5 and 6).

The final polished genome assemblies of *M. oryzae* we built in this project will give us useful insights into how this fungus evolves and adapts to Italian environmental conditions at the genome structure level. This project is part of the OpenRiceBlast community on Zenodo: https://zenodo.org/communities/openriceblast. The data can also be accessed via the OpenRiceBlast portal http://openriceblast.org.

Conclusions

The total lengths of the polished genome assemblies (44.6 to 48.5 Mbp) indicate near-complete sequencing of the *M. oryzae* isolates shown in Table 4. This is complemented by near complete BUSCO score (97.7-98.8%) in polished assemblies (Table 6) for all the isolates indicating that the fungal orthologous gene space is highly represented. The individual contigs were also long enough to enable structural variation studies and appear to be of acceptable quality judging from the level of collinearity observed between our assemblies and the reference 70-15 genome. We ensured open access to these genome data to inspire community involvement in analyzing these data and tackling the blast disease of rice, wheat and other crops using cutting-edge genomic tools.

Materials and methods

Single-spore cultures of *M. oryzae* isolates (Table 1) from different regions of Italy were used for genomic DNA extraction and sequencing.

High molecular weight genomic DNA from *M. oryzae* was extracted from mycelia of 7-day old cultures by following the method described by (Schwessinger and Rathjen, 2017). Genomic DNA was quantified on a TapeStation (Agilent) and treated with DNAse-free RNAse. RNAse-treated DNA was sheared using either a gTUBE or a 22 Gauge needle. Sheared DNA was captured using AMPure beads (Beckman Coulter, Indianapolis, US) and eluted in 45 µl water and used for library construction following the 1D protocol from Oxford Nanopore. Sequencing runs were performed using PromethION platform (Oxford Nanopore Technologies, Oxford, UK). Sequence reads were assembled into contigs using Canu software (v1.8) (Koren *et al.,* 2017). Simultaneously, Illumina sequencing libraries were prepared from the same genomic DNA samples by following a NextTera protocol (Caruccio, 2011) modified by Karasov *et al*., (2018), and sequence reads were obtained using the Illumina HiSeq 3000 platform located at the Genome Center facility at the Max Planck Institute for Developmental Biology.

Canu assemblies were subjected to two rounds of polishing with Illumina short sequence reads generated from the same isolates using Pilon version 1.22 (Walker *et al.*, 2014) and Racon version 1.3.2 (Vaser *et al*., 2017). To remove mitochondrial sequences, we performed blastn (Altschul *et al*., 1990) search against our assemblies using the mitochondrial sequences from the reference genome 70-15 and removed any contigs that had high similarity to the mitochondrial sequences. We ran BUSCO v4.0.6 (Seppey *et al*., 2019) using fungi_obd10 database with default parameters.

Whole genome alignments between each Italian isolate genome assembly and the reference genome of isolate 70-15 were performed using the nucmer function of the MUMmer 3 package (Kurtz *et al.*, 2004). The resulting alignments were filtered for the alignments that

span over >10 kb using the “delta-filter” function with the option “-l 10000”and colinearity plots were generated using mummerplot with the “--color” option.

Acknowledgments

This research was funded by The Gatsby Charitable Foundation, The Max Planck Society, and the European Research Council BLASTOFF project. Nanopore sequencing and assembly services were sourced from Future Genomics Technologies (Leiden, The Netherlands). Illumina sequencing was provided by the Sequencing Centre of the Max Planck Institute for Developmental Biology (Tübingen, Germany). *M. oryzae* isolates were collected within the “Risinnova Project”, AGER Foundation.

References

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J Mol Biol* **215:**403-410.

Caruccio N. 2011. Preparation of next-generation sequencing libraries using Nextera™ technology: Simultaneous DNA fragmentation and adaptor tagging by in vitro transposition. In: Kwon Y., Ricke S. (eds) *High-Throughput Next Generation Sequencing. Methods in Molecular Biology* **733:**241-255. Humana Press, New York, NY.

Darling AC, Mau B, Blattner FR, Perna NT. 2004. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res* **14:**1394-403.

Dean RA, Van Kan JAL, Pretorius ZA, Hammond-Kosack K.E., Di Pietro A, Spanu PD, Rudd JJ, Dickman M, Khamann R, Ellis J, Foster GD. 2012. The Top 10 fungal pathogens in molecular plant pathology. *Mol Plant Pathol* **13**:414-430.

Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H *et al.* 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**:980-986.

Fisher M, Henk D, Briggs C, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ. 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature* **484**:186–194.

Karasov TL, Almario J, Friedemann C, Ding W, Giolai M, Heavens D, Kersten S, Lundberg DS, Neumann M, Regalado J, Neher RA, Kemen E, Weigel D. 2018. *Arabidopsis thaliana* and *Pseudomonas* pathogens exhibit stable associations over evolutionary timescales. *Cell Host Microbe.* **24:**168-179.e4.

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res* **27**:722-736.

Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. *Genome Biol* **5**:R12.

Okagaki LH, Nunes CC, Sailsbery J, *et al.* 2015. Genome sequences of three phytopathogenic species of the Magnaporthaceae family of fungi. *G3*. **5**:2539-2545.

Schwessinger B, Rathjen JP. 2017. Extraction of high molecular weight DNA from fungal rust spores for long read sequencing. *Methods Mol Biol* **1659**:49-57.

Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing genome assembly and annotation completeness. In: Kollmar M. eds. *Gene Prediction. Methods in Molecular Biology*. **1962:**227-245. Humana Press, New York, NY.

Vaser R, Sović I, Nagarajan N, Šikić M. 2017. Fast and accurate *de novo* genome assembly from long uncorrected reads. *Genome Res* **27:**737-746.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, Earl AM. 2014. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9:**e112963.

Table 1. *M. oryzae* isolates used in this study and summary statistics for their nanopore sequencing runs

***M. oryzae* isolate**	**Host**	**Source**	**Year collected**	**Flowcell ID**	**Sequencing platform[1]**	**Basecaller (version)**	**No. Reads**	**Base count (bp)**	**Read N50**	**ENA accession**
AG002	*Oryza sativa*	Sozzago, Italy	2010	PAD04813	PromethION, ONT	Guppy (2.1.3)	469,224	7,522,559,914	24,130	ERR4704598
AG006	*Oryza sativa*	Vigevano, Italy	2011	PAD04813	PromethION, ONT	Guppy (2.1.3)	321,713	5,198,920,186	29,066	ERR4704596
AG032	*Oryza sativa*	Ferrara, Italy	2011	PAD04813	PromethION, ONT	Guppy (2.1.3)	333,608	5,596,762,033	28,813	ERR4704597
AG038	*Oryza sativa*	Olcenengo, Italy	2011	PAD04813	PromethION, ONT	Guppy (2.1.3)	284,715	5,438,810,944	32,347	ERR4704599
AG039	*Oryza sativa*	Vercelli, Italy	2011	PAD04813	PromethION, ONT	Guppy (2.1.3)	305,056	5,380,738,368	30,721	ERR4704592
AG059	*Oryza sativa*	Siziano, Italy	2011	PAD04813	PromethION, ONT	Guppy (2.1.3)	417,916	6,602,278,461	23,457	ERR4704594
AG098	*Oryza sativa*	Oristano, Italy	N/D	PAD04813	PromethION, ONT	Guppy (2.1.3)	314,015	4,665,230,724	28,118	ERR4704600
PR003	*Oryza sativa*	Dossena, Italy	2003	PAD04813	PromethION, ONT	Guppy (2.1.3)	245,744	4,274,949,588	30,333	ERR4704595
San andrea	*Oryza sativa*	N/D	2001	PAD04813	PromethION, ONT	Guppy (2.1.3)	360,461	5,562,064,787	26,468	ERR4704593

[1]Sequencing was performed by Future Genomics Technologies, Leiden; ONT = Oxford Nanopore Technology
N/D = No data; bp = base pair; ENA = European Nucleotide Archive

Table 2. Summary statistics for *M. oryzae* genomes assembled from nanopore reads

***M. oryzae* isolate**	**Assembly Software (version)**	**No. contigs**	**Assembly length (bp)**	**N50 (bp)**	**Max length (bp)**	**Min length (bp)**	**GenBank Accession**
AG002	Canu (1.8)	38	45,963,345	4,529,404	6,354,330	1,902	GCA_905067045.1
AG006	Canu (1.8)	26	46,895,299	6,579,242	10,942,893	51,636	GCA_905067025.1
AG032	Canu (1.8)	25	45,799,422	5,926,258	8,797,420	1,932	GCA_905067055.1
AG038	Canu (1.8)	20	46,144,830	5,640,447	9,106,624	45,956	GCA_905067005.1
AG039	Canu (1.8)	29	47,340,277	6,212,358	10,940,842	1,858	GCA_905067035.1
AG059	Canu (1.8)	38	47,498,635	5,859,208	7,099,302	42,191	GCA_905066965.1
AG098	Canu (1.8)	34	47,650,771	6,089,662	7,827,513	1,894	GCA_905067015.1
PR003	Canu (1.8)	17	44,525,684	6,029,236	8,528,881	40,948	GCA_905067075.1
San andrea	Canu (1.8)	37	48,327,069	4,099,541	6,266,075	13,134	GCA_905067085.1

Table 3. Illumina short sequence reads used to polish the genome assemblies

M. oryzae Isolate	Sequencing Platform[1]	Library Layout	Base count (bp)	No. reads	ENA accession
AG002	HiSeq 3000, Illumina	Paired	1,818,725,400	3,031,209	ERR4757123
AG006	HiSeq 3000, Illumina	Paired	962,429,400	1,604,049	ERR4757124
AG032	HiSeq 3000, Illumina	Paired	1,493,714,400	2,489,524	ERR4757125
AG038	HiSeq 3000, Illumina	Paired	1,252,386,000	2,087,310	ERR4757126
AG039	HiSeq 3000, Illumina	Paired	1,256,038,200	2,093,397	ERR4757127
AG059	HiSeq 3000, Illumina	Paired	1,258,650,600	2,097,751	ERR4757128
AG098	HiSeq 3000, Illumina	Paired	1,345,979,400	2,243,299	ERR4757129
PR003	HiSeq 3000, Illumina	Paired	994,322,400	1,657,204	ERR4757130
San andrea	HiSeq 3000, Illumina	Paired	1,111,573,200	1,852,622	ERR4757131

[1]Sequencing was performed by Max Planck Institute, Tübingen

Table 4. Polished genome assemblies of Italian *Magnaporthe oryzae* isolates without mitochondrial contigs

M. oryzae isolate	Software (version)	No. contigs	Assembly length (bp)	N50 (bp)	Max length (bp)	Min length (bp)	GenBank accession
AG002	Pilon (1.22), Racon (1.3.2)	36	46,086,469	4,555,161	6,389,779	1,896	GCA_905067045.2
AG006	Pilon (1.22), Racon (1.3.2)	24	47,005,811	6,618,557	11,008,917	51,574	GCA_905067025.2
AG032	Pilon (1.22), Racon (1.3.2)	24	45,916,910	5,961,411	8,843,348	1,925	GCA_905067055.2
AG038	Pilon (1.22), Racon (1.3.2)	19	46,291,169	5,673,907	9,163,141	46,127	GCA_905067005.2
AG039	Pilon (1.22), Racon (1.3.2)	26	47,495,958	6,249,570	11,002,933	1,852	GCA_905067035.2
AG059	Pilon (1.22), Racon (1.3.2)	37	47,743,121	5,900,770	7,152,446	42,369	GCA_905066965.2
AG098	Pilon (1.22), Racon (1.3.2)	33	47,810,826	6,094,221	7,877,299	500	GCA_905067015.2
PR003	Pilon (1.22), Racon (1.3.2)	16	44,615,198	6,063,740	8,582,813	41,659	GCA_905067075.2
San andrea	Pilon (1.22), Racon (1.3.2)	35	48,509,714	4,122,582	6,303,665	53,723	GCA_905067085.2

Table 5. BUSCO[1] scores of the genome assemblies before polishing

Assembly name	DB[2]	Mode	Completed	SingleCopy	Duplicated	Fragmented	Missing	Total
AG002	fungi_odb10	genome	632 (83.4%)	630 (83.1%)	2 (0.3%)	26 (3.4%)	100 (13.2%)	758
AG006	fungi_odb10	genome	599 (79.0%)	599 (79.0%)	0 (0.0%)	42 (5.5%)	117 (15.5%)	758
AG032	fungi_odb10	genome	613 (80.9%)	613 (80.9%)	0 (0.0%)	28 (3.7%)	117 (15.4%)	758
AG038	fungi_odb10	genome	616 (81.3%)	616 (81.3%)	0 (0.0%)	29 (3.8%)	113 (14.9%)	758
AG039	fungi_odb10	genome	610 (80.5%)	603 (79.6%)	7 (0.9%)	32 (4.2%)	116 (15.3%)	758
AG059	fungi_odb10	genome	547 (72.2%)	545 (71.9%)	2 (0.3%)	56 (7.4%)	155 (20.4%)	758
AG098	fungi_odb10	genome	610 (80.4%)	609 (80.3%)	1 (0.1%)	30 (4.0%)	118 (15.6%)	758
San andrea	fungi_odb10	genome	627 (82.7%)	621 (81.9%)	6 (0.8%)	29 (3.8%)	102 (13.5%)	758

[1]BUSCO = Benchmarking Universal Single-Copy Orthologs; Seppey *et. al*., 2019. *Methods in Molecular Biology*. 1962: 227-245
[2]DB = Database used in BUSCO

Table 6. BUSCO[1] scores of the genome assemblies after polishing

Assembly name	DB[2]	Mode	Completed	SingleCopy	Duplicated	Fragmented	Missing	Total
AG002	fungi_odb10	genome	747 (98.6%)	744 (98.2%)	3 (0.4%)	0 (0.0%)	11 (1.4%)	758
AG006	fungi_odb10	genome	745 (98.3%)	744 (98.2%)	1 (0.1%)	2 (0.3%)	11 (1.4%)	758
AG032	fungi_odb10	genome	745 (98.3%)	743 (98.0%)	2 (0.3%)	0 (0.0%)	13 (1.7%)	758
AG038	fungi_odb10	genome	745 (98.3%)	744 (98.2%)	1 (0.1%)	1 (0.1%)	12 (1.6%)	758
AG039	fungi_odb10	genome	741 (97.8%)	736 (97.1%)	5 (0.7%)	0 (0.0%)	17 (2.2%)	758
AG059	fungi_odb10	genome	745 (98.3%)	744 (98.2%)	1 (0.1%)	0 (0.0%)	13 (1.7%)	758
AG098	fungi_odb10	genome	749 (98.8%)	747 (98.5%)	2 (0.3%)	1 (0.1%)	8 (1.1%)	758
PR003	fungi_odb10	genome	742 (97.9%)	740 (97.6%)	2 (0.3%)	0 (0.0%)	16 (2.1%)	758
San andrea	fungi_odb10	genome	741 (97.7%)	737 (97.2%)	4 (0.5%)	0 (0.0%)	17 (2.3%)	758

[1]BUSCO = Benchmarking Universal Single-Copy Orthologs; Seppey *et. al*., 2019. *Methods in Molecular Biology*. 1962: 227-245
[2]DB = Database used in BUSCO