# Evidence of non-tandemly repeated rDNAs and their intragenomic heterogeneity in *Rhizophagus irregularis*

## Authors

Taro Maeda[1], Yuuki Kobayashi[1], Hiromu Kameoka[1], Nao Okuma[1], Naoya Takeda[2], Katsushi Yamaguchi[3], Takahiro Bino [3], Shuji Shigenobu[3,4*], Masayoshi Kawaguchi[1,4*]

1 Division of Symbiotic Systems, National Institute for Basic Biology, Japan

2 School of Science and Technology, Kwansei Gakuin University, Japan

3 Functional Genomics Facility, National Institute for Basic Biology, Japan

4 The Graduate University for Advanced Studies [Sokendai], Japan

*Corresponding authors

# Abstract

Arbuscular mycorrhizal fungi (AMF) are one of the most widespread symbionts of land plants. Our substantially improved reference genome assembly of a model AMF, *Rhizophagus irregularis* DAOM-181602 (total contigs = 210, contig N50 = 2.3Mbp) facilitated discovery of repetitive elements with unusual characteristics. *R. irregularis* has only ten to eleven copies of the complete 45S rDNA, whereas the general eukaryotic genome has tens to thousands of rDNA copies. *R. irregularis* rDNAs are highly heterogeneous and lack a tandem repeat structure. These findings provide evidence for the hypothesis of concerted evolution that rDNA homogeneity depends on its tandem repeat structure. RNA-seq analysis confirmed that all rDNA variants are actively transcribed. Observed rDNA/rRNA polymorphism may modulate translation by using different ribosomes depending on biotic and abiotic interactions. The non-tandem repeat structure and intragenomic heterogeneity of AMF rDNA may facilitate adaptation to a broad host range despite lacking a sexual life cycle.

# Introduction

The arbuscular mycorrhizal fungus (AMF) is an ancient fungus at least from the early Devonian[1,2] and forms symbiotic networks with most land plant species[3,4]. AMF colonizes plant roots and develops highly branched structures called arbuscules in which soil nutrients (phosphate and nitrogen) are efficiently delivered to the host plant[5]. The mycelial network formed by various AMF species contributes to plant biodiversity and productivity within the terrestrial ecosystem[6]. The distinctive features of AMF have made it an important model in ecology and evolution[7,8]; these include coenocytic mycelium[4], nutrition exchange with the plant and classification as an obligate biotroph[9], signal crosstalk during mycorrhiza development[10] and extremely high symbiotic ability[4,11].

The hyphae of AMF form a continuous compartment through which nuclei can migrate, such that the body is constructed with coenocytic cells which contain multiple nuclei per cell[4,12]. Since Sanders *et al* (1995)[13], many studies have indicated intracellular polymorphisms of rDNA (ITS) in various AMF species[14-16]. *In situ* DNA-DNA hybridization using rDNA detected the coexistent of genetically different nuclei (heterokaryosis) in a cell of an AMF, *Scutellospora castanea*[12]. Hijri and Sanders (2005) confirmed the internuclear variation using a single-copy gene (i.e., Pol1-like sequence: PLS)[17]. Other experiments on various genes and species further supported heterokaryosis in AMF[18-20]. No sexual stages have been observed in AMF, leading to much discussion on the evolutionary significance of heterokaryosis in the absence of sexual recombination[21-25]. However, heterokaryosis in AMF was challenged recently by multiple studies. Sanger-sequencing of isolated spore and nuclei from *Claroideoglomus etunicatum* indicated gene-duplication of PLS, and suggested variation among the paralogs of PLS and rDNA (=intragenomic heterogeneity)[26]. Illumina-sequencing of an isolated haploid single nucleus of a model strain of AMF, *Rhizophagus irregularis* DAOM-181602 also indicated heterozygous "single nucleotide polymorphisms" (SNPs) on rDNAs, suggesting intragenomic heterogeneity of rDNA copies[27]. Another genomic study found dikaryon-like heterokaryosis from multiple wild *R. irregularis*, implying that AMF have a sexual reproductive stage and alter nucleus heterogeneity within their life cycle[28].

Recently, multiple genome projects have advanced understanding of AMF. Genomic data have been provided from *R. irregularis* DAOM-181602[29], *Gigaspora rosea*[30], *Rhizophagus clarus*[31], and wild *R. irregularis*[28]. These studies revealed potential host-dependent biological pathways[29,30], and candidate genes for plant infection and sexual reproduction[28,30,31]. Gene-duplication of a subset of marker genes for heterokaryosis is also suggested within the genome of *R. irregularis*[27].

Although these studies advanced the field, a fragmented genome assembly was a barrier to further molecular biological studies of AMF[32]. The first published genome sequence of *R. irregularis* DAOM-181602[29] contained 28,371 scaffolds and an N50 index of 4.2 kbp (Table S3). The second sequence by Lin et al. 2014[27] (GCA000597685.1) contained 30,233 scaffolds with an N50 of 16.4 kbp (Table S3). The quality of genomic sequence data for other AMF species did not surpass that of DAOM-181602[30,31]. In contrast, many of other

fungus than AMF contains less than several hundred scaffolds and N50 lengths over 1Mbp[33]. For example, a genomic sequence of an asymbiotic fungus closely related to AMF, *Rhizopus delemar* (GCA000149305.1), was constructed from 83 assemblies and an N50 of 3.1 Mbp[34]. Fragmented genome sequences limit our ability to analyze repetitive structure and to distinguish between orthologous and paralogous genes[32]. Thus, we present here an improved whole-genome sequence of *R. irregularis* DAOM-181602 to facilitate examination of the genomics underlying specific features of AMF. In this paper, we focused on rDNA which is a key component used in many AMF studies including heterokaryosity[6,12,26,27], phylogeny[35-37], ecology[8,37,38], and evolution[39].

# Results & Discussion

## A highly contiguous and complete reference genome of DAOM-181602 generated by PacBio-based *de novo* assembly

We used primarily single-molecule, real-time (SMRT) sequencing technology for sequencing and assembling the *R. irregularis* genome. We generated a 76-fold whole-genome shotgun sequence (11.7 Gbp in total) (Table S1) using the PacBio SMRT sequencing platform from genome DNA isolated from a spore solution of a commercial strain of *R. irregularis* DAOM-181602. A total of 766,159 reads were generated with an average length of 13.1 kbp and an N50 length of 18.8 kbp (Table S1). We assembled these PacBio reads using the HGAP3 program[40] (149.9 Mb composed of 219 contigs). To detect erroneous base calls, we generated 423M of 101bp-paired-end Illumina whole-genome sequence data (Table S1) and aligned them to the HGAP3 assembly. Through variant calling, we corrected 3,032 single base call errors and 10,841 small indels in the HGAP3 assembly. Nine contigs were almost identical to carrot DNA sequences deposited in the public database (Table S2), and these were removed based on the assumption that they were contaminants derived from a host plant used by the manufacturer in the cultivation of *R. irregularis*. We evaluated completeness of the final assembly using CEGMA[41]; of the 248 core eukaryotic genes, 244 genes (98.4%) were completely assembled (Tables 1, S3). Consequently, we obtained a high-quality reference genome assembly of *R. irregularis* DAOM-181602, which is referred to as RIR17.

Compared with previous assemblies[27,29], the new assembly RIR17 represents a decrease in assembly fragmentation (ca. 30,000 to 210) and 140-fold improvement in contiguity using the N50 contig length as a metric (Tables 1, S3). The total size of the assembly increased by 9-59 Mbp from previous versions, reaching 97.24% coverage of the whole genome, assuming a genome size of 154 Mb as estimated with flow cytometry[29] (Tables 1, S3). The new assembly contains no ambiguous bases (N-bases), whereas previous assemblies had 30 kbp or 270 kbp of ambiguous sites (Table 1, S3).

## New gene annotation for DAOM-181602 confirms gene family expansion, loss of conserved fungal genes, heterokaryosis, and repeat-richness

Using the RIR17 assembly together with newly generated RNA-seq data ("Rir_RNA_SS" in Table S1), we built a new set of 41,572 gene models (43,675 transcripts) (Table S4). New models showed >20% longer ORFs on average and more coverage of "Benchmarking Universal Single-Copy Orthologs" (BUSCO) (Table S4), indicating an improvement in gene prediction. Of the 41,572 genes predicted, 27,860 (67.0%) had either RNA-seq expression support, homology evidence or protein motif evidence (Table S5).

*R. irregularis* has the largest number of genes among fungi, to our knowledge (Figure S1). The inflation of the gene number was caused by lineage-specific expansions of gene families, not by whole genome

duplications (Table S6). This gene expansion is consistent with the previously suggested "highly paralogous" gene composition of the *R. irregularis* genome[25,27,29]

Analysis of AMF genes based on the previous *R. irregularis* genomic assemblies suggested the loss of several categories of genes from symbiosis with host plants[27,29,30]. Our ortholog analyses using the improved RIR17 genome assembly and gene models confirmed the loss of genes involved in degradation of plant cell walls such as cellobiohydrolase (GH6 and GH7 in CAZy database), polysaccharide lyases (PL1 and PL4), proteins with cellulose-binding motif 1 (CBM1), and Lytic polysaccharide monooxygenases (Table S7), and nutritional biosynthetic genes including fatty acid synthase (FAS) and the thiamine biosynthetic pathway (Table S8). Given that fatty acid and thiamin are essential nutrients for fungi[42,43], *R. irregularis* should take up those essential nutrients from a host plant without digestion of the plant cell wall. Several recent papers have described the transport of lipids from plants to AMF[44-46].

Despite pervasive gene family expansions in the *R. irregularis* genome, many marker genes that have been used for evaluation of heterokaryosis were single-copy in RIR17. The marker genes included "POL1-like sequence" (PLS)[17], "heat shock 70kDa protein 5" ("Binding Protein: BIP " in Kuhn et al 2001)[12], "40S ribosomal protein S2" (40S-riboprot/rps2)[18], and "translation elongation factor Tu" (Ef-tu)[18]. Our RIR17 indicated a single copy of PLS, rps2 and Ef-tu genes, but showed triplication of the BIP gene (Table S9). The gene number for PLS and BIP are consistent with estimates from previous genomes[27]. Our exhaustive gene sets supports heterokaryosis based on PLS[17], rps2[18] and Ef-tu[18], and suggests an overestimation of polymorphism based on BIP[47].

Analysis of repeat elements in the RIR17 assembly indicated that the AMF genome is repeat-rich with a unique repeat profile. The RepeatModeler[48] and RepeatMasker[48] pipeline identified 64.4 Mb (43.03%) as repetitive elements, with 39.84% classified as interspersed repeats (Table S10). Although the interspersed repeats were classified as DNA transposons (21.74%), LINEs (10.42%) and LTRs (5.00%), the majority (62.83%) of the interspersed repeats could not be categorized with known repeat classes (Table S10), indicating that the AMF genome accumulated novel classes of interspersed repeats. Although the SINE repeat is common in closely related fungi (Mucorales), we could not find any SINE repeats in the AMF (Table S10). A Pfam search identified 810 gene models containing a transposase domain (Table S11).

## *R. irregularis* has the lowest rDNA copy number among eukaryotes

The general eukaryotic genome has tens to thousands of rDNA copies[49] (Figure 1a). However, the RIR17 genome assembly contained only ten copies of the complete 45S rDNA cluster composed of 18SrRNA, ITS1, 5.8SrRNA, ITS2, and 28SrDNAs (Figure 1b, Table S12). To confirm that no rDNA clusters were overlooked, we also estimated the rDNA copy number based on read depth of coverage. Mapping the Illumina reads of the genomic sequences (Rir_DNA_PE180) onto the selected reference sequences indicated that the coverage depth of the consensus rDNA was 11-8 times deeper than the average of the single-copy genes (Figure 1c, Table S13),

indicating the number of rDNA copies is approximately 10, and the RIR17 assembly covers almost all of the rDNA copies.

This rDNA copy number is one of the lowest among eukaryotes[50] (Table 2), and has relevance for understanding the translation system in AMF. For instance, wild-type yeast (*Saccharomyces cerevisiae* NOY408-1b) has about 150 rDNA copies (Table 2). An experimental decrease of rDNA copy number could not isolate any strain having <20 copies, which is considered the minimum number to allow yeast growth[51]. The doubling time of a yeast with 20 rDNA copies (TAK300) was 20% longer than that of the wild type[51]. In DAOM-181602, successive cultivation has been widely observed while under an infected state with a plant host suggesting that this exceptionally small rDNA copy number is enough to support growth. The multinucleate feature of AMF would increase the rDNA copy number per cell and thereby may supply enough rRNA to support growth.

## *R. irregularis* rDNAs are heterogeneous and completely lack a tandem repeat structure

Interestingly, none of the RIR17 rDNAs form a tandem repeat structure, in contrast to most eukaryotic rDNA comprising tens to hundreds of tandemly repeated units. Most of the rDNA clusters in RIR17 were placed on different contigs; a single copy of rDNA was found in "unitig_311", "_312", "_35", "_356", "_4", and "_52", and two copies were found in "unitig_39" and "_62" (Figure 1b, Table S12). In the case where two rDNA clusters were found, the two copies resided apart from each other and did not form a tandem repeat; the distance between the clusters was over 70 kbp (76,187 bp in unitig_62 and 89,986 bp in unitig_39, Figure 1b, Table S12), the internal regions contained 31 and 42 protein-coding genes, and the two clusters were placed on reverse strands from each other (Figures 1b, Table S12). Since all rDNA copies are located over 28 kbp away from the edge of each contig (Figure 1b, Table S12), it is unlikely that the observed loss of tandem repeat structure is an artifact derived from an assembly problem often caused by highly repetitive sequences.

We examined polymorphism among the 45S rDNA clusters. Pairwise comparisons of the ten rDNA copies detected 27.3 indels and 106.1 sequence variants with 98.18 % identity on average (Table S15), whereas the sequences of rDNA clusters at c311-1 and c52-1 were found to be identical. There was no sequence similarity up- or down-stream of ten 48S rDNA paralogs. Polymorphisms were distributed unevenly throughout the rDNA; percent identities were 99.91% in 18SrDNA, 97.93% in 28SrDNA, 96.65% in 5.8SrDNA, 93.45% in ITS1, and 90.28% in ITS2 (Table S14, S15, Figure 2). The number of polymorphic sites in *R. irregularis* rDNAs reached 4.07 positions per 100bp, much higher than other fungi with polymorphic sites of 0.04-1.97 positions per 100bp (Table 2).

These results can be explained by “concerted evolution model” of rDNA. The concerted evolution model suggests that multiple types of DNA conformations (such as recombinational repair and intragenomic gene amplification) maintain rDNA homogeneity, and tandemly repeated rDNAs behave as templates or binding sites

for other repeats[52]. This model has supported by *Arabidopsis thaliana* which has one pseudogenic rDNA (lacking 270bp of the important helix as rRNA[53]) besides the main tandem repeated rDNA arrays[54]. In this study, we reported the absence of tandem repeat structure of rDNA from the AM genome. Such a complete lack of rDNA tandem repeats have never been detected in wild or artificial eukaryote except malaria-causing *Plasmodium* parasites[55,56]. The rDNA polymorph was also confirmed in *Plasmodium*[55,56]. Our results on a fungus evolutionary distinct from *Plasmodium* (Alveolata) and *Arabidopsis* (Archaeplastida) reinforce the generality of the hypothesis that rDNA homogeneity depends on its tandem repeat structure. Although rDNA heterogeneity has been reported in various fungi (e.g., pathogenic fungi[57,58], and other AMFs including *C. etunicatum* and *R. intraradices*[26]), the repeat structures of these rDNAs have never been revealed. Pawlowska and Taylor (2004) predicted that rDNA heterogeneity is caused by relaxation of concerted rDNA evolution in Glomerales including *Rhizophagus*[26]. Here, we propose a hypothetical mechanism for the "relaxation" in AMF; we suggest that loss of tandem repeat structure precludes DNA conformations associated with the maintenance of homogeneity and thereby inhibits the general homogenization process in eukaryotic rDNA.

Our phylogenetic analysis suggests that AMF has a different type of weak concerted rDNA evolution system. Two rDNA pairs on the same contig (c39-1 and c39-2, c62-1 and c62-2) had higher similarity than other paralogs (Figure 2f) and were in the opposite direction from each other (Figure 1b). This inverted repeat structure is often observed in chloroplast rDNAs and sequence identity is maintained by gene conversion[59]. Furthermore, we found no orthologous rDNA genes from other *Rhizophagus* species (Figure S2). Previously observed rDNA heterogeneity in Glomerales[26] suggests that the concerted evolution relaxed before the diversification of *Rhizophagus* species. When the rDNA duplicated before speciation, each of the duplicated genes formed a clade with the orthologs in other species[20]. Our tree suggests that the observed rDNAs in *R. irregularis* either expanded after speciation, or a weak sequence homogenization system assimilated the copies after speciation.

## Impact to translation and the biological significance of non-tandemly repeated rDNAs

To confirm the transcriptional activity of each rDNA, we conducted a total-RNAseq. Illumina sequencing of a modified library for rRNA sequencing ("Rir_RNA_rRNA" in Table S1) produced 18,889,290 reads (read length = 100-301bp) from DAOM-181602. We mapped the reads to all gene models from RIR17 (43,675 protein-encoding isoforms and ten 48S rDNA paralogs) and estimated the expression levels of each gene by eXpress software. All rDNAs paralogs were over 5,000 FPKMs (Fragments Per Kilobase of exon per Million mapped fragments) (Table 3) and multiple reads were matched to the specific region of each paralog, indicating that the ten rDNA copies are transcriptionally active. In general, some eukaryotes change the transcribed sequences by "RNA editing"[60] and normally eukaryotes will silence a part of the rDNA copies[61]. These mechanisms were not detected in the AMF, and the rRNA were as polymorphic as the rDNA. These results show that DAOM-181602 have multiple types of ribosomes containing different rRNA. Additionally, we

determined highly duplicated ribosomal protein genes (e.g., Ribosomal protein S17/S11) (Table S6) and tRNA genes indicating unknown amino acid isotypes, which may also account for the heterogeneity of ribosomes (Table S16).

The ribosomal heterogeneity sheds light on the novel adaptation strategy of AMF to a broad host range. The heterokaryosity in AMF has been proposed to drive variable genetic combinations of mycelium in the absence of sexual recombination[24]. Recent genomic studies, furthermore, discovered many signatures of sexual reproduction within the dikaryon-like stage[25,28]. Our rDNA research here appends another hypothetical strategy to enhance phenotypic plasticity. Loss of tandem repeat structure accelerates the accumulation of mutations in each rDNA copy resulting in polymorphism of rRNA and ribosomes, and consequently increasing the rate of adaptation by different translation activities within the same species. In various eukaryotes (e.g. yeast, mice, and *Arabidopsis*), RNA editing and switching of ribosomal protein paralogs produces heterogeneous ribosomes and subsequently alters phenotypes[62]. Furthermore, the malaria parasite with heterogeneous rDNA also produces functionally different multiple rRNA and changes the expression level of each paralog depending on the host species (mammalian or insects)[56,62]. AMF is like the malaria parasite in that they both infect distantly related host species. The relationship between the diversity of host organisms and rDNA polymorphism will be an important area for further research. Our hypothesis does not exclude current theories for the genetic and phenotypic plasticity of AMF (heterokaryosis and sexual reproduction), but proposes a multi-layered diversification mechanism leading to the widespread distribution of AMF.

## Minor genomic variants in DAOM-181602

To find sequence polymorphs associated with heterokaryosis in the strain DAOM-181602, we compared our ten rDNA paralogs with previous partial rDNAs obtained using single-spore cloning-based sequencing[20,63]. Although many of the cloning-based rDNA sequences were almost identical (unmatched sites < 2) with any of our rDNA paralogs in RIR17, seven cloning-based sequences indicated over two un-matched sites (Table S17). Our coverage depth research indicated that one *R. irregularis* genomic dataset contains around 11 rDNA copies and our reference data covered ten copies. Hence, observation of seven un-referential paralogs indicates that this strain contains many more genotypes then the copy number in the genome, supporting heterokaryosis or the inter-spore variation of rDNA genotype sets (mutation during the passage) in this strain.

Heterokaryosis was also supported by mapping DNA sequence data back to the non-repetitive regions in RIR17. We prepared two types of Illumina short-reads: one dataset from multiple spores ("Rir_DNA_PE180") and 28 datasets from the isolated spore (ERR1135012-ERR1135040). From mapping of those reads to the repeat-masked RIR17 (85,305,802bp, Table S10), 6851 SNPs were detected in the multiple-spores dataset, and 38,944 SNPs were detected in the single-spore dataset (Table S19). The two types of datasets shared 837 SNP sites (Tables S19, S20). Among the shared SNPs, 834 sites were heterozygous SNPs in the isolated spore datasets (Table S20), indicating that the single-spore dataset contained the non-referential rDNA genotypes in addition to the referential rDNAs.

Multiple studies have detected intra-mycelial genomic diversity from isolated nuclei or spores of AMF, and DAOM-181602 has been considered a highly homokaryotic strain[27,29]. Induced errors during whole genome amplification made it difficult to assess if this strain is absolutely homokaryotic as other eukaryotes[25]. Our study obtained hundreds of SNPs among the different isolates of DAOM-181602, suggesting the existence of minor polymorphs independent of amplification errors, and supporting the presence of weak heterokaryosis in DAOM-181602. *In vitro* propagation for years might have reduced the genetic variation in this strain because wild *R. irregularis* have more heterokaryotic polymorphisms[64]. Overall, we confirmed heterokaryosis in DAOM-181602 and suggested that the previously investigated intracellular rDNA polymorphism in *R. irregularis*[18,20] would arise from both heterokaryosis and intragenomic heterogeneity of rDNAs.

## Conclusion

We here report a substantially improved version of the genome assembly and the gene models of *R. irregularis* DAOM-181602. The new assembly had the largest gene number among fungi via lineage-specific expansions of particular genes but lacked some genes for essential metabolic processes. The evolutionary reason for this unbalanced gene composition is an important area for future studies of AMF symbiosis.

Our improved genome revealed that common concepts of eukaryotic rDNA are not applicable to AMF. The lack of rDNA tandem repeat structure was an unexpected finding in this study, because the tandem repeat has been conserved throughout most eukaryotes[26,32,65,66]. The dispersed heterogeneous rDNAs supported the association between tandem repeat structure and homogeneity of eukaryotic rDNAs.

The ribosome is a core component for protein synthesis. Our RNAseq data showed that AMF has a heterogeneous composition of ribosomes. AMF may modulate gene expression patterns by using multiple types of ribosomes from polymorphic rDNAs, depending on the environment (Figure 3). Although predicting the functional effect of observed rDNA mutations remains to be solved, it should be noted that the middle area of 28SrDNA (4,450 - 4,500bp on c62-1) had a higher mutation rate than ITS regions (Figure 2b). Because the mutation rate of the ITS region (encoding non-functional RNA) will vary under neutral mutation rates among paralogs, the accumulated variants in the middle-28S region may have functional effects favored by natural selection (via diversifying selection). This region is a useful target for future functional analyses of AMF rRNA.

The rDNA polymorphisms observed in the RIR17 covered most of the polymorphism previously reported in this species, providing incentive to review previous molecular ecological results[7,8] of AMF. rDNA has been widely used for phylogenic analysis[36,67] and species/strain identification in AMF[37,68,69]. Incorporating our research into reassessment of previous work may help to clarify studies on the diversity of wild AMF and the compatibility between AMF and plants. The degree of intragenomic variation was not high enough to disrupt species-level identification, but was sufficient to cause erroneous assumption of *R. irregularis* strains (Figure S2).

# Materials and Methods

## PacBio-based assembling

### DNA preparation

The DNA sample for the PacBio and Illumina sequencing was extracted from a commercial strain of *R. irregularis* DAOM-181602 (MYCORISE® asp, Premier Tech Biotechnologies, Canada). The DNA extraction followed the method of Fulton et al., 1995,[70] with some modification as described below. Purchased spore solution (including about 1,000,000 spores) was centrifuged (4500rpm, 20min), and washed three times with distilled water. Precipitated spores were frozen with liquid nitrogen, ground with pestle, and dispersed in extraction buffer (100mM Tris-HCl pH 8.0, 20mM EDTA, 0.75% sarkosyl, 0.1% PVP, 0.75% CTAB, 0.13M sorbitol, 0.75M NaCl, 0.1 mg/ml proteinase K). After incubation at 37°C for 10 min, the aqueous phase was centrifuged (15000 rpm, 4min) and the pellet was discarded. An equal volume of phenol/chloroform (1:1, vol/vol) was added, gently mixed, and centrifuged (15000rpm, 2 min). The aqueous phase was collected, and an equal volume of chloroform was added, mixed, and centrifuged (15000rpm, 2 min). The aqueous phase was collected again, and 1/10 vol of sodium acetate and 0.7x vol of isopropanol were added, mixed and centrifuged (12000 rpm, 20 min). The resulting pellet was washed twice with 70% EtOH, and eluted with TE buffer. Extracted DNA was purified with Genomic-tip (Qiagen, Germany) following the manufacturer's instructions.

### PacBio sequencing

Long-read sequences were generated with a PacBio RS II sequencer (Pacific Biosciences, Menlo Park, CA, USA) using DNA/Polymerase Binding Kit P6 v2 (Pacific Biosciences) and DNA Sequencing Reagent Kit 4.0 (Pacific Biosciences). The library was prepared according to the 20-kb Template Preparation Using BluePippin™ Size-Selection System (Sage science, MA, USA). A total sequence of 11.7 Gb in 955,841 reads (76× coverage of the genome, assuming a genome size of 154Mb) was obtained from 29 SMRT cells (Table S1). The N50 length of the raw reads was 13,107 bp.

### PacBio-based genome assembling

The *R. irregularis* genome was assembled using the RS_HGAP_Assembly.3 protocol for assembly and Quiver for genome polishing in SMRT Analysis v2.3.0 (Pacific Biosciences). The procedure consisted of three parts involving (1) generation of preassembled reads with improved consensus accuracy; (2) assembly of the genome through overlap consensus accuracy using Celera Assembler; and (3) one round of genome polishing with Quiver. For HGAP, the following parameters were used: PreAssembler Filter v1 (minimum sub-read length = 500 bp, minimum polymerase read quality = 0.80, minimum polymerase read length = 100 bp); PreAssembler v2 (minimum seed length = 6,000 bp, number of seed read chunks = 6, alignment candidates per chunk = 10, total alignment candidates = 24, minimum coverage for correction = 6, and blasr options = 'noSplitSubreads, minReadLength=200, maxScore=1,000, and maxLCPLength=16'); AssembleUnitig v1 (genome size = 150 Mbp, target coverage = 25, overlapper error rate = 0.06,

overlapper min length = 40 bp and overlapper k-mer =14); and BLASR v1 mapping of reads for genome polishing with Quiver (maximum divergence = 30, minimum anchor size = 12). Assembly polishing with PacBio read was carried out with Quiver using only unambiguously mapped reads. The statistics of the PacBio-only assembly set and previously sequenced data (Lin14, Tis13) were evaluated using QUAST ver. 4.3 [71]. The percentage of genome coverage was estimated assuming the genome size to be 154 Mb based on Tisserant et al [29].

### Error correction with Hiseq data

An Illumina genomic library for error correction was constructed with an S220 Focused-Ultrasonicator (Covaris, MA, USA) for fragmentation, Pippin Prep (Sage Science) for size selection (Target fragment size = 180 bp), and a TruSeq DNA Sample Prep Kit (Illumina, CA, USA) for adapter ligation and library amplification. The library was sequenced (101 bp from each end) on a HiSeq 1500 platform (Illumina). A total of 423,041,682 raw reads (53.3 Gb) was obtained from the library (Table S1).

Obtained reads were mapped to PacBio-only assemblies for the error correction. Illumina technology has higher accuracy then PacBio and our Illumina library was made from the same sample as the PacBio libraries. Thus, we considered that detected “variants” from the Illumina data would be caused by sequencing errors in PacBio-only assembling. After filtering low-quality and adapter sequences, paired Illumina reads were joined into overlapping extended sequences using FLASH ver. 1.2.10 [72] with default settings. Using an input of 159,988,396 paired-end sequences, FLASH constructed 159,393,209 jointed reads (101–242 bp). The jointed reads were mapped to PacBio-only assemblies using BWA-mem ver. 0.7.1 [73], and the 152,304,272 reads (95.6%) were mapped to the assemblies. The mapped reads were realigned with IndelRealigner in GenomeAnalysisTK-3.5 [74] and the erroneous sites were called with variant callers; "samtools mpileup" and "bcftools call" (option –vm) for the variant calling and “bcftools filter” for filtering (option -i ‘%QUAL>10’, -s LOWQUAL) of the low quality calling [75]. The called “variant” sites were summarized with plot-vcfstats script in samtools. The discordant positions between PacBio assemblies and the mapped Illumina reads were fixed with the Illumina reads data using “bcftools consensus” [76]. We evaluated the quality of Illumina-polished assemblies with QUAST for calculating the assembly statistics, and with CEGMA ver. 2.5 [41] for estimation of the gene-level completeness. QUAST analysis was done using the same settings as PacBio-only assemblies.

### Identification of host plant contamination

After the polish using Illumina data, we eliminated the sequences derived from contaminated DNAs during the sample preparation. Blastn search of the polished assemblies against “refseq_genomic” database detected nine assemblies showing similarity with carrot sequences (query coverage per subject >95%, percentages of identical matches >90%, bit score > 1000) (Table S2), which might be used as a host plant by the manufacturer for the cultivation of *R. irregularis* samples. After elimination of the nine contaminated contigs, we submitted the assemblies to DDBJ as whole genome shotgun sequence data (RIR17) of *R. irregularis* DAOM-181602 (BDIQ01).

# Gene prediction and annotation

De novo repeat motifs were identified using RepeatModeler v 1.0.8 which combines RECON and RepeatScout programs [48]. Based on the identified motif, the repetitive region in the assemblies was masked with RepeatMasker v 4.0.5 [48]. We used the default parameters for the identification and the masking. The genes containing transposase domain were determined with Pfam in Interproscan v 5.23.62 [77].

For the gene models constructed from RIR17 assemblies, standard RNAseq data was obtained from *R. irregularis* spores and hyphae. The RNA was extracted with RNeasy Plant Mini kit (Qiagen) after the incubation of the purchased spores (MYCORISE® asp) in a minimum nutrient medium for one day. An Illumina RNAseq library was constructed with a TruSeq Stranded mRNA Library prep kit (Illumina). The library was sequenced (101 bp from each end) on a HiSeq 1500 platform (Illumina). A total of 16,122,964 raw reads (3.2 Gb) was obtained from the library (Table S1). After filtering low-quality and adapter sequences, RNAseq data was mapped to RIR17 assemblies with Tophat v2.1.1 [78] with the default setting.

Then, the RIR17 assemblies were processed through the RNAseq-based gene model construction pipeline using Augustus software (ver 3.2.1) [79]. We constructed *R. irregularis* specific probabilistic models of the gene structure, based on manually constructed 495 gene models from the longest "unitig392" sequence in RIR17. Manual gene models were made with *ab initio* Augustus analysis based on probabilistic models for *Rhizopus oryzae*, and by manual refine using the homology data with already-known genes and mapped RNA-seq data. Then, with the trained probabilistic models and the intron-hints data from the mapped RNA-seq read, 37,639 optimal gene models were constructed using the Augustus pipeline. We then confirmed if the Augustus pipeline overlooked the called genes in previous genome studies. We mapped all protein sequences obtained from previous gene modeling on Lin14 and Tis13 against our RIR17 genomic sequences with Exonerate[80] (option --model est2genome --bestn 1), resulting in the recruitment of 3,933 overlooked genes. The completeness of the constructed gene model was evaluated with BUSCO v2.0 [81]. The BUSCO analysis used "Fungi odb9" gene set (http://buscodev.ezlab.org/datasets/fungiodb9.tar.gz) as benchmarks and employed the "-m proteins" option to analyse the pre-constructed protein data without the *ab initio* gene modeling step.

The confidences of the obtained 41,572 genes models were estimated based on 1) RNA-seq expression support, 2) homology evidence, or 3) protein motif evidence. For calculation of gene expression level, we mapped our "Rir_RNA_SS" data and 32 RNA-seq data submitted on SRA database (24 data from DRP002784, and 8 data from DRP003319), and calculated gene expression level (FPKM) using featureCounts [82] with the default setting. Homology with previously known genes was determined by blast search against the orthoDB database (odb9). The protein motif was searched using Pfam analysis in Interproscan ver. 5.22-61.0 [77] (Table S3). The gene models supported by any of the confirmation methods were submitted to DDBJ as a standard gene and the models having no support were assigned as "PREDICTED" gene models.

Constructed gene models were annotated by several *in-silico* searches; blastp (BLAST ver. 2.2.31+) searching against nr, RefSeq and UniProt databases, and orthologous group searching against orthoDB. We manually selected the descriptive nomenclatures from those four searches for the gene function. The MACGs (missing ascomycete core

genes) orthologs were searched using BLAST ver. 2.2.31+ with "-evalue 0.0001" option, and the reference sequences for the MACGs search were selected from protein data from an S288C reference in the SGD database (Table S8). Genes involved in the degradation of plant cell walls were searched by BLAST with the same settings as the MA z z CGs search, and the reference sequences were selected from *Aspergillus niger* CBS 513.88 data in Genbank based on CAZY classification (Table S7).

# Detection of Ribosomal DNA and intragenomic polymorphism

Ribosomal DNA regions were detected by RNAmmer v 1.2[83] from whole RIR17 assemblies, and were manually refined based on the MAFFT v7.294b[84] alignment to the 48S rRNA on *Saccharomyces cerevisiae* S288C. The genomic positions of rDNAs were visualized with python v 3.4.0 (BasicChromosome v 1.68, and GenomeDiagram v 0.2 modules) (Figure 1b).

The number of rDNA paralogs in the genome was estimated by mean depth of coverage. We masked reportative regions (based on RepeatModeler analysis) and all rDNA regions on RIR17 except one rDNA copy (C62-1). Then, trimmed R1 Illumina reads from "Rir_DNA_PE180" library were mapped to the repeat-masked RIR17 using bowtie2 v 2.2.9[85]. The coverage depth of the rDNA region and 243 single-copy BUSCOs were obtained using bedtools v2.26.0 ("bedtools coverage" command with -d option), and the statistics of each regions were calculated and visualized by R software v3.4.2 with ggplot2 library (Figure 1c). To prevent copy number estimation from depth fluctuation due to the intragenomic heterogeneity, we confirmed the coverage depth using the consensus sequences of all ten rDNA paralogs; the joined Illumina reads (from "Rir_DNA_PE180" library) were mapped back to a consensus rDNA sequences and ten single copy BUSCO genes from RIR17, then the depth of coverages was counted by bedtools (genomeCoverageBed) (Table S13).

The difference among the rDNA paralogs was calculated from the aligned sequences by Mafft v7.309 (options: --localpair, --op 5, --ep 3, --maxiterate 1000), using the pairwise comparison with CLC Main Workbench 7.8.1 (Qiagen). The mutation type was called by eye from the alignment, and we chose the c62-1 paralog as a reference sequence for mutation calling (Figures 2a). Phylogenetic trees (Figures 2c, S2) were constructed from the Mafft alignment by the neighbor-joining method with MEGA v 7.0.21[86] under the Maximum composite likelihood model, and were tested for robustness by bootstrapping (500 pseudoreplicates).

# Heterogeneity of translation machineries

The expression level of the rDNA paralogs was examined with modified Illumina sequencing of *R. irregularis* spores and hyphae. Total RNA was extracted with RNeasy Plant Mini kit (Qiagen) after the incubation of the MYCORISE spores in a minimum nutrient medium for seven days. An Illumina RNAseq library was constructed

with a TruSeq Stranded mRNA Library prep kit (Illumina). To skip the poly-A tailing selection step in the library construction, we started from the "fragmentation step" of the standard manufacturer's instructions. The library was sequenced (301 bp from each end) on a Miseq platform (Illumina). A total of 16,122,964 raw reads (3.2 Gb) was obtained from the library (Table S1). After filtering low-quality and adapter sequences, RNAseq data was mapped to the RIR17 assembly with Tophat with the default settings. Fragments Per Kilobase of exon per Million mapped fragments (FPKM) for each gene were calculated with eXpress v1.5.1 with "--no-bias-correct" option. Transfer RNAs were determined with tRNAscan-SE v 1.3.1 [87].

# Minor genomic variants

The different sequence sites among our ten rDNA paralogs and previous cloning-based partial rDNA sequences were calculated by CLC Main Workbench (Qiagen), after the Mafft alignment. On the repetitive region in RIR17, SNPs were determined by two different methods. For our "Rir_DNA_PE180" library from multiple-spore samples, we mapped the joined Illumina read with BWA-mem software against repeat-masked RIR17 sequence, and removed the multiple-hit reads (using "grep" command against 'XT:A:U' meaning uniq-hit in BWA-mem output format) and the PCR duplicate reads (using "samtools rmdup" command). Obtained SAM formatted data including unique hit reads were converted to VCF format using "samtools mpileup" and "bcftools convert" command. After removal of sites with low mapping quality (QUAL<10) and/or a low coverage depth (DP <100), we calculated the rate of non-referential alleles from DP4 values in VCF output files and called the sites showing over 10% non-referential alleles rate as a heterozygous SNP (Table S18). For the libraries from single spore sequencing, we mapped the read by BWA-mem (without the joining of the reads by FLASH) and trimmed the multiple-hit and PCR duplicate reads using the same procedure with the "Rir_DNA_PE180" library. After conversion to VCF format by "samtools mpileup" and "bcftools convert", the SNPs were called with "bcftools call" (-m and --ploidy-file options). Although all of the AMF strains including DAOM-181602 were haploid, we set ploidy-file as diploid ("* * * * 2") to detect the heterozygous SNPs due to the heterokaryosis. After calling by bcftools, we selected SNPs observed from over two libraries as a reliable SNPs (Table S19).

# Authors contributions

T.M., S.S., M.K., conceived of and designed the experiments, T.M., Y.K., H.K., N.T., K.Y., T.B., performed the experiments, T.M., N.O., S.S., T.B., analyzed the data, T.M., Y.K., H.K., K.Y., T.B., S.S., M.K., wrote the manuscript.

# Acknowledgements

This work was supported by JST ACCEL Grant Number JPMJAC1403, Japan. We thank "Functional Genomics Facility" and "Data Integration and Analysis Facility", National Institute for Basic Biology for technical supports; Katsuharu Saito and Kohki Akiyama for discussions; and present and past members of the Kawaguchi-lab and Shigenobu-lab.

# Competing interests

The authors declare no conflicts of interest associated with this manuscript.

# References

1 Remy, W., Taylor, T. N., Hass, H. & Kerp, H. Four hundred-million-year-old vesicular arbuscular mycorrhizae. *Proceedings of the National Academy of Sciences* **91**, 11841-11843 (1994).

2 Redecker, D., Kodner, R. & Graham, L. E. Glomalean fungi from the Ordovician. *Science* **289**, 1920-1921(2000).

3 Bougoure, J., Ludwig, M., Brundrett, M. & Grierson, P. Identity and specificity of the fungi forming mycorrhizas with the rare mycoheterotrophic orchid Rhizanthella gardneri. *Mycol Res* **113**, 1097-1106 (2009).

4 Parniske, M. Arbuscular mycorrhiza: the mother of plant root endosymbioses. *Nat Rev Microbiol* **6**, 763-775 (2008).

5 Smith, S. E. & Smith, F. A. Roles of Arbuscular mycorrhizas in plant nutrition and growth: New paradigms from cellular to ecosystem scales. *Annu Rev Plant Biol* **62**, 227-250 (2011).

6 van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69-72 (1998).

7 Johnson, N. C., Graham, J. H. & Smith, F. A. Functioning of mycorrhizal associations along the mutualism-parasitism continuum. *New Phytol* **135**, 575-586 (1997).

8 Davison, J. *et al.* Global assessment of arbuscular mycorrhizal fungus diversity reveals very low endemism. *Science* **349**, 970-973 (2015).

9 Fellbaum, C. R. *et al.* Fungal nutrient allocation in common mycorrhizal networks is regulated by the carbon source strength of individual host plants. *New Phytol* **203**, 646-656 (2014).

10 Lee, J. The distribution of cytoplasm and nuclei within the extra-radical mycelia in *Glomus intraradices*, a species of arbuscular mycorrhizal fungi. *Mycobiology* **39**, 79-84 (2011).

11 Zhang, Y. & Guo, L. D. Arbuscular mycorrhizal structure and fungi associated with mosses. *Mycorrhiza* **17**, 319-325 (2007).

12 Kuhn, G., Hijri, M. & Sanders, I. R. Evidence for the evolution of multiple genomes in arbuscular mycorrhizal fungi. *Nature* **414**, 745-748 (2001).

13 Sanders, I. R., Alt, M., Groppe, K., Boller, T. & Wiemken, A. Identification of ribosomal DNA polymorphisms among and within spores of the glomales - application to studies on the genetic diversity of arbuscular mycorrhizal fungal communities. *New Phytol* **130**, 419-427 (1995).

14 LloydMacgilp, S. A. *et al.* Diversity of the ribosomal internal transcribed spacers within and among isolates of *Glomus mosseae* and related mycorrhizal fungi. *New Phytol* **133**, 103-111 (1996).

15 Hosny, M., Hijri, M., Passerieux, E. & Dulieu, H. rDNA units are highly polymorphic in *Scutellospora castanea* (Glomales, Zygomycetes). *Gene* **226**, 61-71 (1999).

16 Hijri, M. & Sanders, I. R. The arbuscular mycorrhizal fungus *Glomus intraradices* is haploid and has a small genome size in the lower limit of eukaryotes. *Fungal Genet Biol* **41**, 253-261 (2004).

17 Hijri, M. & Sanders, I. R. Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. *Nature* **433**, 160-163 (2005).

18 Boon, E., Halary, S., Bapteste, E. & Hijri, M. Studying genome heterogeneity within the arbuscular mycorrhizal fungal cytoplasm. *Genome Biol Evol* **7**, 505-521 (2015).

19 Boon, E., Zimmerman, E., Lang, B. F. & Hijri, M. Intra-isolate genome variation in arbuscular mycorrhizal fungi persists in the transcriptome. *J Evolution Biol* **23**, 1519-1527 (2010).

20 Thiery, O. *et al.* Sequence variation in nuclear ribosomal small subunit, internal transcribed spacer and large subunit regions of *Rhizophagus irregularis* and *Gigaspora margarita* is high and isolate-dependent. *Mol Ecol* **25**, 2816-2832, doi:10.1111/mec.13655 (2016).

21 Croll, D. *et al.* Nonself vegetative fusion and genetic exchange in the arbuscular

mycorrhizal fungus Glomus intraradices. *New Phytol* **181**, 924-937 (2009).

22 Croll, D. & Sanders, I. R. Recombination in *Glomus intraradices*, a supposed ancient asexual arbuscular mycorrhizal fungus. *Bmc Evol Biol* **9** (2009).

23 Sanders, I. R. Evolutionary genetics - No sex please, we're fungi. *Nature* **399**, 737-739 (1999).

24 Sanders, I. R. & Croll, D. Arbuscular Mycorrhiza: The challenge to understand the genetics of the fungal partner. *Annu Rev Genet* **44**, 271-292 (2010).

25 Corradi, N. & Brachmann, A. Fungal mating in the most widespread plant symbionts? *Trends Plant Sci* **22**, 175-183 (2017).

26 Pawlowska, T. E. & Taylor, J. W. Organization of genetic variation in individuals of arbuscular mycorrhizal fungi. *Nature* **427**, 733-737 (2004).

27 Lin, K. *et al.* Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *Plos Genet* **10** (2014).

28 Ropars, J. *et al.* Evidence for the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi. *Nat Microbiol* **1**, 16033 (2016).

29 Tisserant, E. *et al.* Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *P Natl Acad Sci USA* **110**, 20117-20122 (2013).

30 Tang, N. *et al.* A survey of the gene repertoire of *Gigaspora rosea* unravels conserved features among Glomeromycota for obligate biotrophy. *Frontiers in Microbiology* **7**, 233 (2016).

31 Toro, K. S. & Brachmann, A. The effector candidate repertoire of the arbuscular mycorrhizal fungus *Rhizophagus clarus*. *Bmc Genomics* **17** (2016).

32 Young, J. P. W. Genome diversity in arbuscular mycorrhizal fungi. *Curr Opin Plant Biol* **26**, 113-119 (2015).

33 Mohanta, T. K. & Bae, H. The diversity of fungal genome. *Biol Proced Online* **17** (2015).

34 Ma, L. J. *et al.* Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication. *Plos Genet* **5** (2009).

35 Redecker, D. & Raab, P. Phylogeny of the Glomeromycota (arbuscular mycorrhizal fungi): recent developments and new gene markers. *Mycologia* **98**, 885-895 (2006).

36 Schussler, A., Schwarzott, D. & Walker, C. A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycol Res* **105**, 1413-1421, doi:Doi 10.1017/S0953756201005196 (2001).

37 Stockinger, H., Kruger, M. & Schussler, A. DNA barcoding of arbuscular mycorrhizal fungi. *New Phytol* **187**, 461-474 (2010).

38 Sun, Y. Q. *et al.* The molecular diversity of arbuscular mycorrhizal fungi in the arsenic

mining impacted sites in Hunan Province of China. *J Environ Sci-China* **39**, 110-118 (2016).

39 VanKuren, N. W., den Bakker, H. C., Morton, J. B. & Pawlowska, T. E. Ribosomal RNA gene diversity, effective population size, and evolutionary longevity in asexual Glomeromycota. *Evolution* **67**, 207-224 (2013).

40 Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, 563-569 (2013).

41 Parra, G., Bradnam, K. & Korf, I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genornes. *Bioinformatics* **23**, 1061-1067 (2007).

42 Tehlivets, O., Scheuringer, K. & Kohlwein, S. D. Fatty acid synthesis and elongation in yeast. *Bba-Mol Cell Biol L* **1771**, 255-270 (2007).

43 Li, M. G. *et al.* Thiamine biosynthesis in *Saccharomyces cerevisiae* is regulated by the NAD(+)- dependent histone deacetylase Hst1. *Mol Cell Biol* **30**, 3329-3341 (2010).

44 Wewer, V., Brands, M. & Dormann, P. Fatty acid synthesis and lipid metabolism in the obligate biotrophic fungus *Rhizophagus irregularis* during mycorrhization of *Lotus japonicus*. *Plant J* **79** (2014).

45 Keymer, A. *et al.* Lipid transfer from plants to arbuscular mycorrhiza fungi. *Elife* **6** (2017).

46 Bravo, A., Brands, M., Wewer, V., Dormann, P. & Harrison, M. J. Arbuscular mycorrhiza-specific enzymes FatM and RAM2 fine-tune lipid biosynthesis to promote development of arbuscular mycorrhiza. *New Phytol* **214**, 1631-1645 (2017).

47 Angelard, C., Colard, A., Niculita-Hirzel, H., Croll, D. & Sanders, I. R. Segregation in a mycorrhizal fungus alters rice growth and symbiosis-specific gene transcription. *Curr Biol* **20**, 1216-1221 (2010).

48 A.F.A. Smit, R. H., P. Green *RepeatMasker*, < http://repeatmasker.org>

49 Eickbush, T. H. & Eickbush, D. G. Finely orchestrated movements: Evolution of the ribosomal RNA genes. *Genetics* **175**, 477-485 (2007).

50 Prokopowich, C. D., Gregory, T. R. & Crease, T. J. The correlation between rDNA copy number and genome size in eukaryotes. *Genome* **46**, 48-50 (2003).

51 Takeuchi, Y., Horiuchi, T. & Kobayashi, T. Transcription-dependent recombination and the role of fork collision in yeast rDNA. *Gene Dev* **17**, 1497-1506 (2003).

52 Gibbons, J. G., Branco, A. T., Godinho, S. A., Yu, S. K. & Lemos, B. Concerted copy number variation balances ribosomal DNA dosage in human and mouse genomes. *P Natl Acad Sci USA* **112**, 2485-2490 (2015).

53 Mentewab, A. B., Jacobsen, M. J. & Flowers, R. A. Incomplete homogenization of 18 S ribosomal DNA coding regions in Arabidopsis thaliana. *BMC Research Notes* **4**,

93-93 (2011).

54 Kaul, S. *et al.* Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* **408**, 796-815 (2000).

55 Gardner, M. J. *et al.* Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498-511 (2002).

56 Vembar, S. S., Droll, D. & Scherf, A. Translational regulation in blood stages of the malaria parasite *Plasmodium* spp.: systems-wide studies pave the way. *Wires RNA* **7**, 772-792 (2016).

57 Ganley, A. R. D. & Kobayashi, T. Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. *Genome Res* **17**, 184-191 (2007).

58 Simon, U. K. & Weiss, M. Intragenomic Variation of Fungal Ribosomal Genes Is Higher than Previously Thought. *Mol Biol Evol* **25**, 2251-2254 (2008).

59 Daniell, H., Lin, C. S., Yu, M. & Chang, W. J. Chloroplast genomes: diversity, evolution, and applications in genetic engineering. *Genome Biol* **17** (2016).

60 Brennicke, A., Marchfelder, A. & Binder, S. RNA editing. *Fems Microbiol Rev* **23**, 297-316 (1999).

61 Birch, J. L. & Zomerdijk, J. C. B. M. Structure and function of ribosomal RNA gene chromatin. *Biochem Soc T* **36**, 619-624 (2008).

62 Xue, S. F. & Barna, M. Specialized ribosomes: a new frontier in gene regulation and organismal biology. *Nat Rev Mol Cell Bio* **13**, 355-369 (2012).

63 Su, A. A. H. & Randau, L. A-to-I and C-to-U Editing within Transfer RNAs. *Biochemistry-Moscow+* **76**, 932-937 (2011).

64 Wyss, T., Masclaux, F. G., Rosikiewicz, P., Pagni, M. & Sanders, I. R. Population genomics reveals that within-fungus polymorphism is common and maintained in populations of the mycorrhizal fungus *Rhizophagus irregularis*. *Isme J* **10**, 2514-2526 (2016).

65 Dover, G. Molecular drive - a cohesive mode of species evolution. *Nature* **299**, 111-117 (1982).

66 Richard, G. F., Kerrest, A. & Dujon, B. Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiol Mol Biol R* **72**, 686-727, (2008).

67 Redecker, D. Molecular identification and phylogeny of arbuscular mycorrhizal fungi. *Plant Soil* **244**, 67-73 (2002).

68 Kruger, M., Stockinger, H., Kruger, C. & Schussler, A. DNA-based species level detection of Glomeromycota: one PCR primer set for all arbuscular mycorrhizal fungi. *New Phytol* **183**, 212-223 (2009).

69 Stockinger, H., Walker, C. & Schüßler, A. 'Glomus intraradices DAOM197198', a model fungus in arbuscular mycorrhiza research, is not *Glomus intraradices*. *New Phytol* **183**, 1176-1187 (2009).

70 Fulton, T. M., Chunwongse, J. & Tanksley, S. D. Microprep protocol for extraction of DNA from tomato and other herbaceous plants. *Plant Mol Biol Rep* **13**, 207-209 (1995).

71 Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072-1075 (2013).

72 Magoc, T. & Salzberg, S. L. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).

73 Li, H. & Durbin, R. Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589-595 (2010).

74 McKenna, A. *et al.* The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* **20**, 1297-1303 (2010).

75 Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987-2993 (2011).

76 Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25** (2009).

77 Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240 (2014).

78 Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105-1111 (2009).

79 Keller, O., Kollmar, M., Stanke, M. & Waack, S. A novel hybrid gene prediction method employing protein multiple sequence alignments. *Bioinformatics* **27**, 757-763 (2011).

80 Slater, G. S. & Birney, E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics* **6** (2005).

81 Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210-3212 (2015).

82 Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923-930 (2014).

83 Lagesen, K. *et al.* RNAmmer: consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Res* **35**, 3100-3108 (2007).

84 Katoh, K., Misawa, K., Kuma, K. & Miyata, T. MAFFT: a novel method for rapid

multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res* **30** (2002).

85 Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357-U354 (2012).

86 Kumar, S., Stecher, G. & Tamura, K. MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Mol Biol Evol* **33**, 1870-1874 (2016).

87 Schattner, P., Brooks, A. N. & Lowe, T. M. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* **33**, W686-W689 (2005).

88 Kwan, E. X., Wang, X. B. S., Amemiya, H. M., Brewer, B. J. & Raghuraman, M. K. rDNA copy number variants are frequent passenger mutations in *Saccharomyces cerevisiae* deletion collections and de Novo transformants. *G3-Genes Genom Genet* **6**, 2829-2838 (2016).

89 Goffeau, A. *et al.* Life with 6000 genes. *Science* **274**, 546-567 (1996).

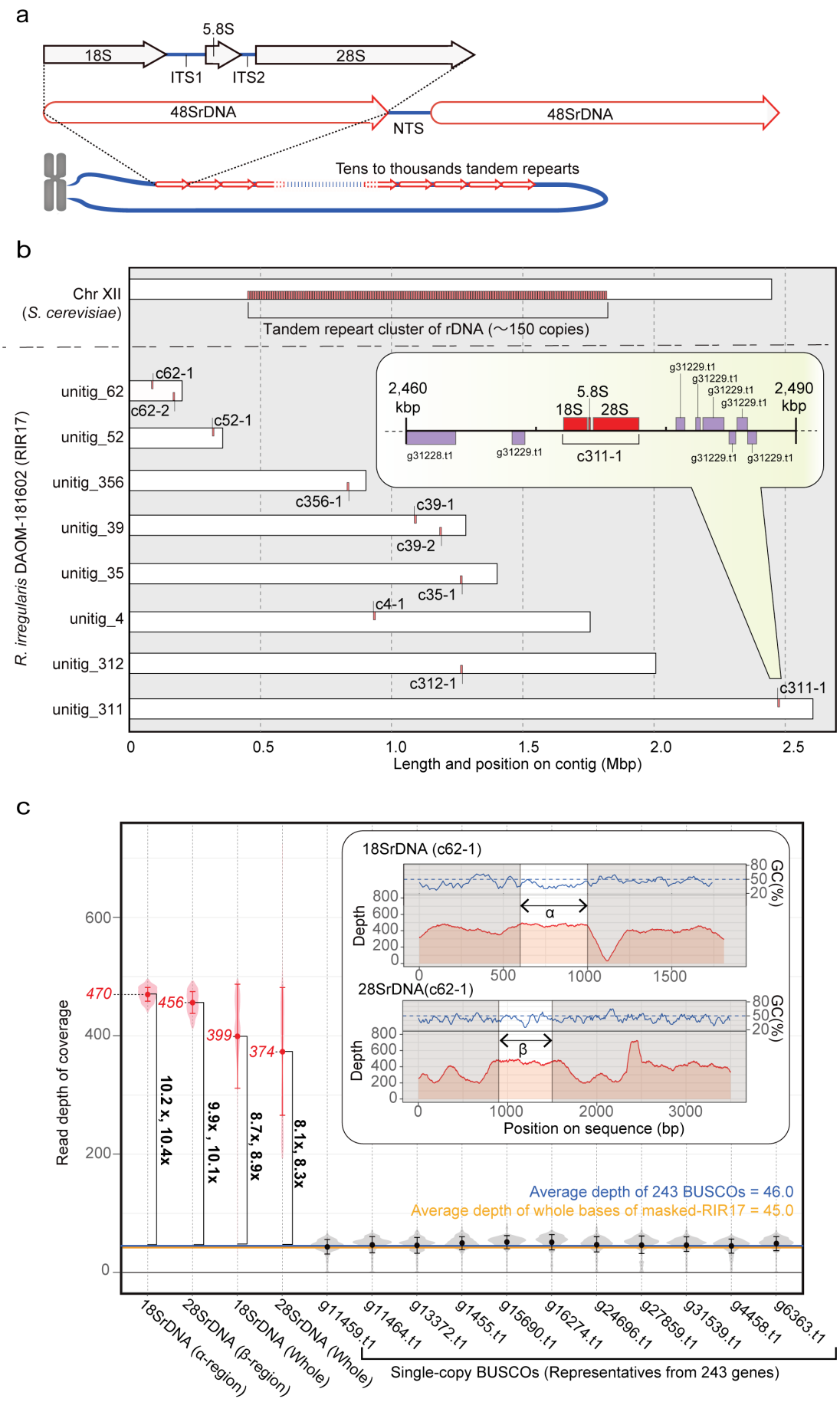


**Figure 1 Physical maps of rDNA structures and copy number in RIR17**

**(Continue)**

**Figure 1 Physical maps of rDNA structures and copy number in RIR17**

**a. A general structure of eukaryotic rDNA clusters.**

**b. Distribution of *R. irregularis* rDNA units in the genome.** Each 48SrDNA unit is represented as a red box. For comparison, rDNA clusters on *Saccharomyces cerevisiae* chromosome XII is shown[88,89]. Inset is a magnified view of a 48rDNA unit (c311-1) with nearby protein-encoding genes (purple boxes). Genes encoded on plus strand genome are depicted on the top side, and those encoded on minus strand are shown on the bottom side.

**c. Copy number of rDNA in DAOM-181602 based on the read depth of coverage.**

Averages of the "read depth of coverage" are represented as dots and italic labels. Error bars and violin plot show standard deviations and normalized coverage distribution. The depths of rDNA regions are marked with red color. For comparison, the data from representative single-copy BUSCO genes on RIR17 were shown with black color. The mean depth of means from 243 BUSCOs is described with horizontal blue line, and the mean depth of all RIR17 bases is described with orange line. The changes of the depth on rDNA regions are typed in vertical bold labels and square bracket. Adapted rDNA regions for the copy number estimation ($\alpha$- and $\beta$-region) were described in inset with the depth of coverage and the GC content on each sequence positions.

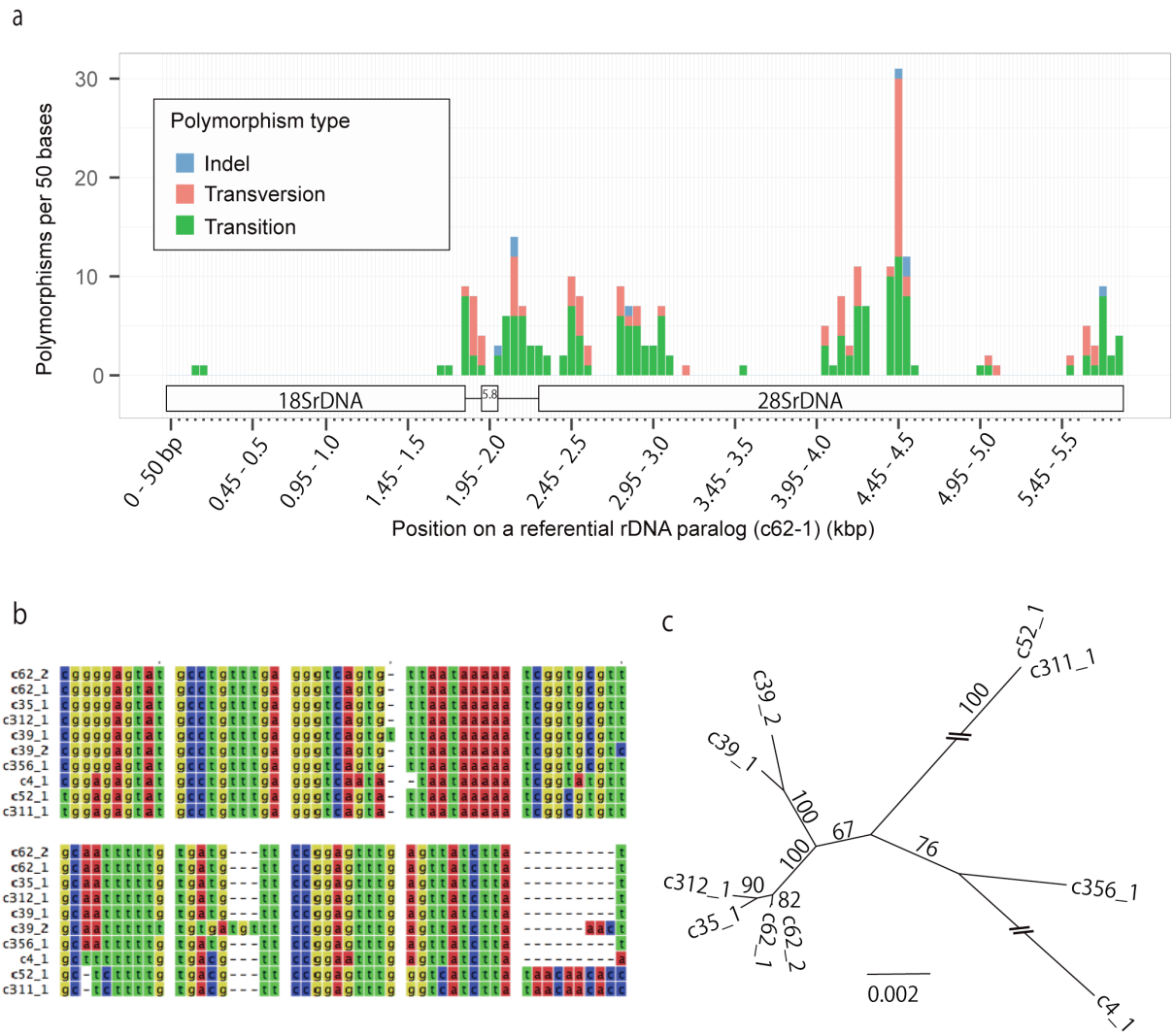


**Figure 2 Polymorphisms of 48SrDNA paralogs in RIR17**

**(Continue)**

**Figure 2 Polymorphisms of 48SrDNA paralogs in RIR17**

**a. The distribution of rDNA sequence variants within the 48SrDNA of RIR17.** The position and types of polymorphisms were called based on the paralog c62-1. **b. Alignment of a heterogeneous region among the 48SrDNA paralogs.** Partial sequences of Mafft-aligned 48SrDNAs (corresponding 2,049-2,136 bp positions on c62-1). **c. Neighbor-joining tree for phylogenetic relationships among the ten rDNA paralogs based on 5,847 aligned positions.** Bootstrap values are described on each node.

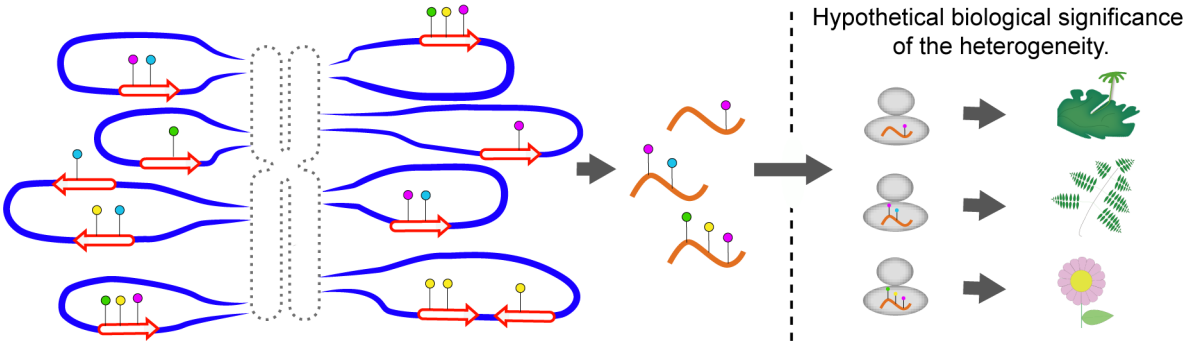


**Figure 3 The structure and expression of ribosomal DNAs/RNAs of *R. irregularis* and putative biological significance for the symbiosis**

Ten or eleven copies of rDNAs are scattered throughout the DAOM-181602 genome. The sequences of rDNAs are heterogeneous, and the transcribed rRNAs form multiple types of ribosomes. We hypothesize that *R. irregularis* uses different types of ribosomes depending on the environmental condition, which may contribute to widen the host range through the phenotype plasticity via their different translation activity among the ribosomes, like the malaria-parasite.

**Table 1 Assembly statistics of *R. irregularis* genome**

| | RIR17 |
|---|---|
| Accession number | BDIQ01000000 |
| Predicted genome size by FCM | 154 Mbp |
| Total length of contigs (% of genome) | 149,750,837 bp (97%) |
| # contigs | 210 |
| # N's bases | 0 |
| Longest contig (bp) | 5,727,599 |
| Contig N50 (bp) | 2,308,146 |
| L50 | 23 |
| GC % | 27.9% |
| CEGMA completeness for genome contigs | 98.4% |
| # Predicted genes | 41,572 |
| BUSCO completeness for gene models (DB; fungi_odb9) | 94.1% (273/290) |
| Complete single copy | 83.8% (243/290) |
| Complete duplicated | 10.3% (30/290) |
| Fragmented | 3.8% (11/290) |
| Missing | 2.1% (6/290) |

**Table 2 Numbers of intragenomic polymorphic sites in fungal rDNAs**

| Species | # polymorphic sites | Repeat unit length (bp) | # units in genome | # polymorphic sites / 100bp |
|---|---|---|---|---|
| *Rhizophagus irregularis* | 238 | 5847 | 10 | 4.07 |
| *Rhizophagus irregularis*[27] | 38 | 1563 | - | 2.43 |
| *Ashbya gossypii*[55] | 3 | 8147 | 50 | 0.04 |
| *Saccharomyces paradoxus*[55] | 13 | 9103 | 180 | 0.14 |
| *Saccharomyces cerevisiae*[55] | 4 | 9081 | 150 | 0.04 |
| *Aspergillus nidulans*[55] | 11 | 7651 | 45 | 0.14 |
| *Cryptococcus neoformans*[55] | 37 | 8082 | 55 | 0.46 |
| *Phoma exigua* var. *exigua*[56] | 27 | 1672 | - | 1.61 |
| *Mycosphaerella punctiformis*[56] | 26 | 1669 | - | 1.56 |
| *Teratosphaeria microspora*[56] | 16 | 1671 | - | 0.96 |
| *Davidiella tassiana*[56] | 33 | 1672 | - | 1.97 |

**Table 3 Transcription activity of the rDNA paralogs**

| | | Confidence interval (95%) | |
|---|---|---|---|
| Target ID | FPKM | Low | High |
| c312_1 | 28,888 | 28,672 | 29,103 |
| c39_1 | 20,719 | 20,537 | 20,901 |
| c39_2 | 20,358 | 20,177 | 20,538 |
| c62_2 | 19,431 | 19,254 | 19,608 |
| c4_1 | 19,430 | 19,255 | 19,605 |
| c52_1 | 19,054 | 18,879 | 19,228 |
| c311_1 | 19,054 | 18,879 | 19,228 |
| c356_1 | 16,151 | 15,990 | 16,311 |
| c35_1 | 10,053 | 9,927 | 10,180 |
| c62_1 | 7,656 | 7,546 | 7,766 |