The ISME Journal (2016) 10, 2514–2526

www.nature.com/ismej

OPEN

ORIGINAL ARTICLE

Population genomics reveals that within-fungus polymorphism is common and maintained in populations of the mycorrhizal fungus *Rhizophagus irregularis*

Tania Wyss[1,3], Frédéric G Masclaux[1,2,3], Pawel Rosikiewicz[1], Marco Pagni[2,4] and Ian R Sanders[1,4]

[1]*Department of Ecology and Evolution, University of Lausanne, Lausanne, Switzerland and* [2]*Vital-IT, SIB Swiss Institute of Bioinformatics, Lausanne, Switzerland*

Arbuscular mycorrhizal (AM) fungi are symbionts of most plants, increasing plant growth and diversity. The model AM fungus *Rhizophagus irregularis* (isolate DAOM 197198) exhibits low within-fungus polymorphism. In contrast, another study reported high within-fungus variability. Experiments with other *R. irregularis* isolates suggest that within-fungus genetic variation can affect the fungal phenotype and plant growth, highlighting the biological importance of such variation. We investigated whether there is evidence of differing levels of within-fungus polymorphism in an *R. irregularis* population. We genotyped 20 isolates using restriction site-associated DNA sequencing and developed novel approaches for characterizing polymorphism among haploid nuclei. All isolates exhibited higher within-isolate poly-allelic single-nucleotide polymorphism (SNP) densities than DAOM 197198 in repeated and non-repeated sites mapped to the reference genome. Poly-allelic SNPs were independently confirmed. Allele frequencies within isolates deviated from diploids or tetraploids, or that expected for a strict dikaryote. Phylogeny based on poly-allelic sites was robust and mirrored the standard phylogeny. This indicates that within-fungus genetic variation is maintained in AM fungal populations. Our results predict a heterokaryotic state in the population, considerable differences in copy number variation among isolates and divergence among the copies, or aneuploidy in some isolates. The variation may be a combination of all of these hypotheses. Within-isolate genetic variation in *R. irregularis* leads to large differences in plant growth. Therefore, characterizing genomic variation within AM fungal populations is of major ecological importance.
The ISME Journal (2016) **10,** 2514–2526; doi:10.1038/ismej.2016.29; published online 8 March 2016

Introduction

Arbuscular mycorrhizal (AM) fungi (phylum Glomeromycota) are important symbionts of plants. By colonizing roots, AM fungi extend the nutrient absorption capabilities of plants, enhancing plant mineral nutrition and health (Smith and Read, 2008). AM fungi are distributed globally (Öpik *et al.*, 2013), increase plant growth (Koch *et al.*, 2006) and drive plant diversity (van der Heijden *et al.*, 1998, 2008). They are increasingly used in agriculture, enhancing yields of globally important crops, while diminishing the need for fertilizers (Ceballos *et al.*, 2013). Despite the recognized importance of AM fungi, their evolutionary and population genetics is poorly understood (Sanders and Croll, 2010). Intraspecific genetic variation in AM fungal populations can be large (Börstler *et al.*, 2008, 2010; Croll *et al.*, 2008) and results showing differential effects on plant growth that can be as large as those caused by different AM fungal species (Munkvold *et al.*, 2004; Koch *et al.*, 2006). Therefore, genetic variation within species of AM fungi is biologically and ecologically important.

Despite the importance of AM fungi, basic knowledge on their genetics is missing. It has been hypothesized that AM fungi harbor genetically different nuclei in a common cytoplasm (heterokaryosis; Sanders and Croll, 2010). Such a state may provide these fungi with the ability to respond rapidly to environmental change (Sanders and Croll, 2010; Angelard *et al.*, 2014). Indirect experimental evidence for heterokaryosis in AM fungi exists in *Glomus etunicatum* (Hijri and Sanders,

Correspondence: IR Sanders, Department of Ecology and Evolution, University of Lausanne, Biophore Building, Lausanne 1015, Switzerland.
E-mail: ian.sanders@unil.ch
[3]These authors contributed equally to this work.
[4]These authors are joint senior authors on this work.
Received 30 April 2015; revised 22 January 2016; accepted 25 January 2016; published online 8 March 2016

2005) and in one *Rhizophagus irregularis* isolate (isolate C3; Ehinger *et al*., 2012). However, direct evidence for heterokaryosis only exists in the AM fungus *Scutellospora castanea* (Kuhn *et al*., 2001). Within-isolate genetic variation in *R. irregularis* can give rise to offspring of a given isolate that are genetically different from the parent and leads to up to fivefold changes in rice growth (Angelard *et al*., 2010). This clearly demonstrates that within-isolate genetic variation is biologically important, irrespective of whether it is located on different nuclei or not.

Whole-genome sequencing (WGS) should shed light on the state of polymorphism within an organism (Jones *et al*., 2004; Wibberg *et al*., 2013; Hane *et al*., 2014). The whole-genome sequence of the model AM fungus *Rhizophagus irregularis*, isolate DAOM 197198, was recently published (Tisserant *et al*., 2013). Although AM fungi are an ecologically important component of all soils, this single isolate is the only genome to have been sequenced from the whole Glomeromycota phylum. This haploid genome revealed a low level of within-isolate polymorphism (Tisserant *et al*., 2013). The average single-nucleotide polymorphism (SNP) density was 0.43 SNPs/kb. The cause of this variation was attributed to possible sequencing errors, gene paralogs and repeats. Another initiative sequenced the genomes of four single nuclei of the same isolate following whole-genome amplification. A very low density of SNPs occurred among four nuclei, suggesting that this isolate is homokaryotic (Lin *et al*., 2014). In contrast, another study, employing whole-genome and amplicon sequencing of targeted loci using 454 pyro-sequencing reported high within-isolate polymorphism across the genomes of *Rhizophagus* spp. isolates, finding tens of different alleles at individual loci in *R. irregularis* DAOM 197198 (Boon *et al*., 2015).

Paradoxically, results of experimental studies on *R. irregularis* are difficult to explain without invoking the heterokaryosis hypothesis. Isolates of *R. irregularis* showed differences in allele frequencies at a single locus (locus *Bg112*) after propagation from single spores (Ehinger *et al*., 2012), as a result of hybrid crosses (Angelard *et al*., 2010) or after a change in host plant species (Angelard *et al*., 2014). Such variation would not be expected if the alleles were not located on different nuclei or if the numbers of copies of the different genes varied among nuclei. Genetic variation among and within AM fungus isolates has been studied at only a very limited number of loci (Kuhn *et al*., 2001; Croll *et al*., 2008; Angelard *et al*., 2010; Ehinger *et al*., 2009, 2012; Boon *et al*., 2013).

In view of the importance of understanding genetic variation in AM fungal populations, and of apparently opposing findings on genetic variation in *R. irregularis* from WGS, 454 pyro-sequencing and experimental studies, we used two different Illumina sequencing-based techniques to genotype DAOM 197198 and nineteen *in vitro*-grown single spore isolates of *R. irregularis*, representing a population from one field (Koch *et al*., 2004; Croll *et al*., 2008). We aimed to: (1) measure how much within-isolate genetic variation exists in a population of this fungus, to what extent this differs among individuals within the population and how this compares with DAOM 197198 and (2) compare the within-isolate genetic variation in *R. irregularis* to genetic variation observed in other haploid, diploid and polyploid species, using the same techniques.

WGS is expensive with the high coverage needed to detect infrequent alleles in many isolates from a population. In restriction site-associated DNA (RAD) sequencing (Figure 1; Baird *et al*., 2008), only DNA fragments adjacent to restriction sites are sequenced, increasing the depth of coverage at an affordable cost (Emerson *et al*., 2010; Hohenlohe *et al*., 2010; Scaglione *et al*., 2012; Reitzel *et al*., 2013; Wang *et al*., 2013; Hoffman *et al*., 2014). Another technique, amplicon sequencing, involves deep sequencing of targeted loci amplified by PCR (Lange *et al*., 2014), so that even alleles in very low frequency can be more reliably detected (Figure 1; Larsen *et al*., 2013). We analyzed genetic variation in 20 isolates of *R. irregularis* with double-digest restriction site-associated DNA sequencing (ddRAD-seq), and validated a subset of polymorphic loci by amplicon sequencing.

Because *R. irregularis* is haploid (Sedzielewska *et al*., 2011; Tisserant *et al*., 2013) but could potentially be heterokaryotic, we refrained from using standard methods that are developed to study diploids (Nielsen *et al*., 2011; Catchen *et al*., 2013). We developed an approach that did not impose a predefined genetic model and called variants in what could potentially be a heterogeneous population of nuclei co-existing within an individual. To eliminate misinterpretations due to technical artifacts such as sequencing errors, we sequenced at least three biological replicates of each *R. irregularis* isolate; an approach that is almost never used in WGS and which allows a much more robust detection of true poly-allelic sites. We treated repetitions in the genome separately to avoid overestimation of within-isolate polymorphism caused by misalignment of repeated DNA fragments. Finally, to confirm the reliability of our analyses, we included DNA samples from reference species with different ploidy levels, including *Saccharomyces cerevisiae* (*n*), *Schizosaccharomyces pombe* (*n*), *Candida albicans* (2*n*) and *Betula* spp. (2*n* and 4*n*).

Materials and methods

Source of data

We performed ddRAD-seq on the DAOM 197198 reference isolate and 19 isolates from a field in Switzerland (Supplementary Table 1; Croll *et al*., 2008). Nomenclature of the Swiss isolates follows Koch *et al.* (2004). In parallel, we also performed

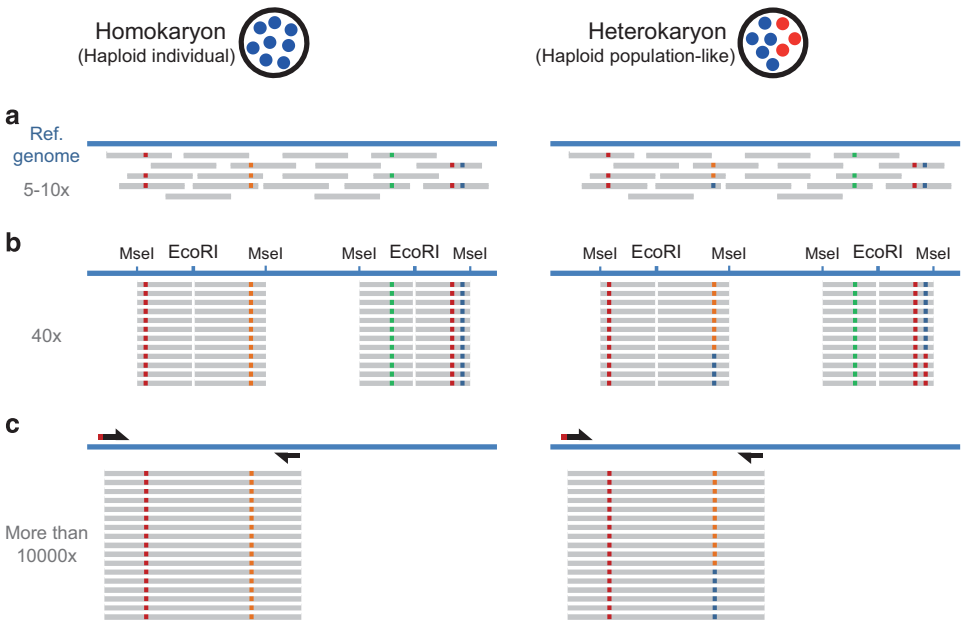

Figure 1 Sequencing techniques to discover within-fungus polymorphism. Blue lines represent the assembled genome. Gray lines represent sequence reads. Polymorphism is detected by aligning sequence reads to the reference genome and calling variants (represented as small colored bars). × = coverage. (**a**) Whole-genome sequencing: intra-genomic polymorphic sites cannot easily be detected because of low coverage. (**b**) RAD sequencing: thousands of sites are sequenced, allowing the detection of intra-genomic polymorphism. (**c**) Amplicon sequencing: target loci are sequenced, yielding very high coverage. The diagram shows the expected results from a homokaryon and a heterokaryon.

ddRAD-seq on *Saccharomyces cerevisiae* (haploid strain S288C), *Schizosaccharomyces pombe* (haploid strain 972 h-), *Candida albicans* (diploid strains DSY294 and SC5314; Sanglard *et al.*, 1998) and mixes of DNA of two *R. irregularis* isolates, B1 and C4 (Supplementary Table 2 and Supplementary Methods). Culture and DNA extraction of all samples are described in Supplementary Methods. The *Betula* spp. RAD sequencing data are described elsewhere (Wang *et al.*, 2013) and obtained from NCBI (BioProject Accession: PRJEB3322).

The reference genome of *R. irregularis* (DAOM 197198) was a single nucleus genome assembly named N6 (Lin *et al.*, 2014). Reference genomes of *S. cerevisiae*, *S. pombe* and *Betula nana* were obtained from Genbank (GCA_000146045.2, GCA_000002945.2 and CAOK01000000, respectively). The reference genome of *C. albicans* was previously described (A21-s02-m08-r09, Arnaud *et al.*, 2013).

ddRAD-sequencing paired-end library construction and sequencing

A ddRAD-seq protocol was used (Parchman *et al.*, 2012; Peterson *et al.*, 2012). Full details of the protocol are given in Supplementary Methods.

In silico *prediction of ddRAD-seq fragments in genomes*

Genomes were digested *in silico* with *Eco*RI and *Mse*I and only fragments containing both *Eco*RI and *Mse*I cut sites and longer than 50 bp were retained. Fragments were aligned to the genome with Novoalign (Novocraft-Technologies, 2014) and fragments that could not be re-aligned were identified (Supplementary Table 3). Only sequence reads falling in re-aligning predicted fragments were considered in the subsequent analyses.

In silico *characterization of the N6 genome and of predicted ddRAD-seq fragments*

Coding regions were predicted by GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008). Since misalignment of repeated regions could lead to detection of false within-isolate polymorphism, repeated regions were defined with two complementary approaches. First, we used RepeatModeler Open-1.0 (Smit and Hubley, 2008) to perform *de novo* repeat family prediction using each genome as an input, and RepeatMasker Open-3.0 (Smit *et al.*, 1996) to annotate repeats. Second, the *in silico* predicted ddRAD-seq fragments were submitted to pairwise comparisons using ggsearch36, a global pairwise alignment algorithm from the package fasta-36.3.5e (Pearson and Lipman, 1988) to identify globally similar predicted fragments. Results from RepeatMasker and ggsearch36 approaches were combined together to define the subset of predicted fragments that we referred to as 'repeated'. Each site in any predicted ddRAD-seq fragment had the characteristics of being either coding or non-coding and either repeated or non-repeated. More details are given in Supplementary Methods and Supplementary Table 3.

ddRAD-seq data analysis
Quality filtering of raw reads and de-multiplexing is described in Supplementary Methods. We aligned ddRAD-seq reads against the N6 genome with the aligner Novoalign V3.02.00 (Novocraft-Technologies, 2014). A ddRAD-seq locus was defined as a group of identical or very similar reads that align only once against the same locus in the N6 genome. Samtools version 0.1.19+ was used to manipulate these alignments and calculate depth of coverage (Li *et al*., 2009). To call variants, we used the program Freebayes (Garrison and Marth, 2012). Freebayes can be used on genomes of any ploidy, pooled samples or mixed populations. Freebayes was set up to detect SNPs, insertions and deletions (indels) and multiple-nucleotide polymorphisms (MNPs) in each replicate at sites with a minimum depth of coverage of 10×. To avoid missing rare alleles, we recorded alleles with frequencies greater or equal to 0.1. Only sites with a phred-scaled quality score greater or equal to 30 were conserved. Freebayes was run individually for each replicate. All variant files were then post-processed to establish the list of sites that were found to be variable in at least one of the isolates, giving 84 211 sites in the N6 genome. For every replicate, and at each site, we reported whether there was a single allele or more than one allele (poly-allelic site) recorded. A missing value (NA) was assigned if the depth of coverage was below 10×.

Data from the other species were treated identically. The *Betula* spp. reads first were digested *in silico* to simulate the same ddRAD-seq protocol as the one used for *R. irregularis*, and aligned to the *Betula nana* genome.

Bayesian phylogeny reconstruction
Only sites containing SNPs and for which information was present in all replicates of each isolate were considered. Only the major allele at poly-allelic sites was retained to simulate a homokaryotic state. Data were concatenated to produce a multiple alignment of 5060 bp sequences. Bayesian phylogeny was constructed with MrBayes 3.2 (Ronquist and Huelsenbeck, 2003). The evolutionary model was set to the GTR substitution model with gamma-distributed rate variation across sites and a proportion of invariable sites. Two independent chains of Markov Chain Monte Carlo analysis, with an incremental heating temperature of 0.25, were run for 300 000 generations, resulting in the standard deviation of the split frequencies below 0.01. The output tree was edited and arbitrary mid-point rooted with FigTree v1.4.0 (Rambaut, 2012).

Hierarchical clustering based on poly-allelic sites
To investigate whether poly-allelic sites were conserved among isolates and if they were meaningful at a phylogenetic level, we calculated the proportion of common poly-allelic sites between pairs of samples. The distance between isolates A and *B* was computed using the Jaccard similarity coefficient as:

$$d_{\text{Jaccard}}(A,B) = 1 - \frac{\left|\{a_i\} \cap \{b_j\}\right|}{\left|\{a_i\} \cup \{b_j\}\right|} \quad (1)$$

where $\{a_i\}$ and $\{b_j\}$ are the filtered sets of poly-allelic sites reported by Freebayes for isolates *A* and *B*, respectively. The Jaccard similarity coefficient measures the proportion of sites where poly-allelic SNPs, MNPs or indels are common between two samples. The hierarchical clustering followed Ward's minimum variance method included in the hclust function in R. A cladogram was drawn using the APE package (Paradis *et al*., 2004).

Hierarchical clustering based on allelic composition
An initial filtering step of 84 211 sites, including SNPs, MNPs and indels, resulted in 13 184 sites (N_1) where there was no missing information. Distance between two isolates was then computed as a sum of dot products,

$$d_{\text{composition}}(A,B) = \frac{1}{N}\sum_{i=1}^{N_i} 1 - \frac{\boldsymbol{A}_i \cdot \boldsymbol{B}_i}{\|\boldsymbol{A}_i\| \, \|\boldsymbol{B}_i\|} \quad (2)$$

where $\boldsymbol{A}_i$ and $\boldsymbol{B}_i$ are frequency vectors of observed alleles at position *i* for isolates *A* and *B*, respectively. The hierarchical clustering followed Ward's minimum variance method. We computed node support values from 100 subsets of 5000 randomly chosen sites. Support of consensus tree bipartitions was computed using prop.clades from the R package APE (Paradis *et al*., 2004). This counts the number of times the bipartitions in the initial tree were present in the series of 99 remaining trees.

Measurement of within-isolate polymorphism
Within-isolate polymorphism was measured as the density of poly-allelic sites, which is expressed as the number of poly-allelic variants/total kb of sequenced regions. The density of poly-allelic sites was calculated for each replicate of each isolate. The number of poly-allelic sites was counted among the 84 211 sites (N_2). This was divided by the number of sites from predicted ddRAD-seq fragments that were sequenced with a coverage greater or equal to 10. A second independent method, entropy, was also used as a measurement of within-fungus polymorphism (see Supplementary Methods).

Haplotype characterization and non-synonymous mutations
Variant calling can reveal the existence of within-isolate polymorphism in *R. irregularis*, but this cannot provide information about the genetic complexity in fungal isolates. To further characterize within-isolate polymorphism, we reconstructed haplotypes using a conservative method which links

genetic variants present on individual sequence reads in the ddRAD-seq data (see Supplementary Methods). In addition, we calculated the frequency of non-synonymous codons in transcript regions where two haplotypes were found (see Supplementary Methods).

Amplicon sequencing

To confirm polymorphic loci in *R. irregularis,* we amplified nine loci from isolates DAOM 197198, C2, C3 and C5 (Supplementary Tables 1 and 4). On the basis of previous research and the ddRAD-seq data, three of the loci were expected to be monomorphic and seven expected to be polymorphic in at least one of the four isolates. Five loci were located in non-repeated regions in the N6 genome (*Bg196*, *Bg235*, *Bg348*, ddRAD-seq loci *3*, *6*, *8*, *10* and *11*), and one locus (ddRAD-seq locus *7*) was located in a repeated region. To minimize the risk of amplifying more than one locus, we verified by BLAST that primer pairs and PCR product sequences matched only once in the N6 *R. irregularis* genome (for amplification and library construction protocols, see Supplementary Methods). We processed sequencing reads for quality as described for the ddRAD-seq analysis. Demultiplexed paired-end reads were joined with the software FLASH (Magoč and Salzberg, 2011). Amplicon sequences were only kept when both primers could be identified. We counted unique sequences, aligned the 20 most frequent with MAFFT version 7 (Katoh and Standley, 2013), and filtered out low frequency sequences. Because every unique sequence corresponded to a different haplotype, we calculated the proportions of haplotypes for each locus based on the occurrence of unique amplicons per locus.

Results

Analysis of ddRAD-seq data

Double-digest RAD sequencing allowed us to generate millions of reads for each isolate of *R. irregularis*, covering thousands of positions (see Supplementary Results and Supplementary Table 5 for read numbers and coverage). A total of 84 211 variable sites were detected in the *R. irregularis* population (74 416 SNPs, 13 324 indels and 60 MNPs; Supplementary Table 5). Mean global densities of mono-allelic SNPs ranged from 0.27 SNP/kb between D1 and D3 to 5.68 SNP/kb between A2 and C2 (SNP densities in coding regions among individual replicates shown in Supplementary Table 6). In addition, we detected many sites in each isolate that harbored more than one allele, that is, poly-allelic sites, distributed across the genome (Supplementary Tables 5 and 7, Supplementary Figure 1).

Consistency among replicates and proportion of alleles observed a small number of times

Mono-allelic sites (SNPs) were consistent among biological replicates of Swiss isolates as only 0.2–0.6% of sites were inconsistent (Supplementary Results and Supplementary Table 7), showing that variants called at sites with a minimum coverage of 10× and minimum allele frequency threshold of 0.1 were reliable. Very good consistency among biological replicates was also found for poly-allelic sites, where only an average of 13% of sites were inconsistent among replicates (Supplementary Table 7). Most of the inconsistency was due to missing alleles, which can happen in low coverage sites (Supplementary Results). At poly-allelic SNPs, alleles only found once or twice represented a very small proportion of the total number of alleles (1% of all alleles were found once and 2.7% of all alleles were found twice; Supplementary Table 8). On average, nearly 80% of alleles that were consistent among the replicates of an isolate were found more than 10 times (Supplementary Figure 2). Finally, allele frequency distributions observed when mixing known proportions of DNA of two *R. irregularis* isolates reflected the proportions of each DNA (Supplementary Results).

Bayesian phylogeny

Bayesian phylogeny reconstruction assigned the isolates to 10 genetically different groups. The phylogenetic tree comprised four main branches (Supplementary Figure 3). All replicates per isolate clustered strictly together or within the same group.

Hierarchical clustering based on poly-allelic sites

By only using poly-allelic sites to determine the phylogeny among isolates, we were able to confirm that poly-allelic sites were consistent among replicates and that closely related isolates shared similar poly-allelic sites. We identified nine distinct groups separated into three main branches (Supplementary Figure 4). All replicates clustered strictly together in the same group, showing that distribution of poly-allelic sites was not random and followed the Bayesian phylogeny.

Hierarchical clustering based on allelic composition

We also used a method that considered allelic composition and frequency of alleles at every site shared by all samples (Figure 2). This included mono- and poly-allelic sites with SNPs, indels and MNPs. The resulting tree was almost identical to the Bayesian phylogeny and similar to previous clustering (Croll *et al*., 2008).

Density of poly-allelic sites and entropy

When only considering SNPs, the density of poly-allelic sites was higher in repeated regions than in non-repeated regions (Figure 3a). This difference was also associated with a higher coverage in repeated regions (Figure 3a). In non-repeated-coding regions, the mean poly-allelic site density was 1.73 sites/kb,

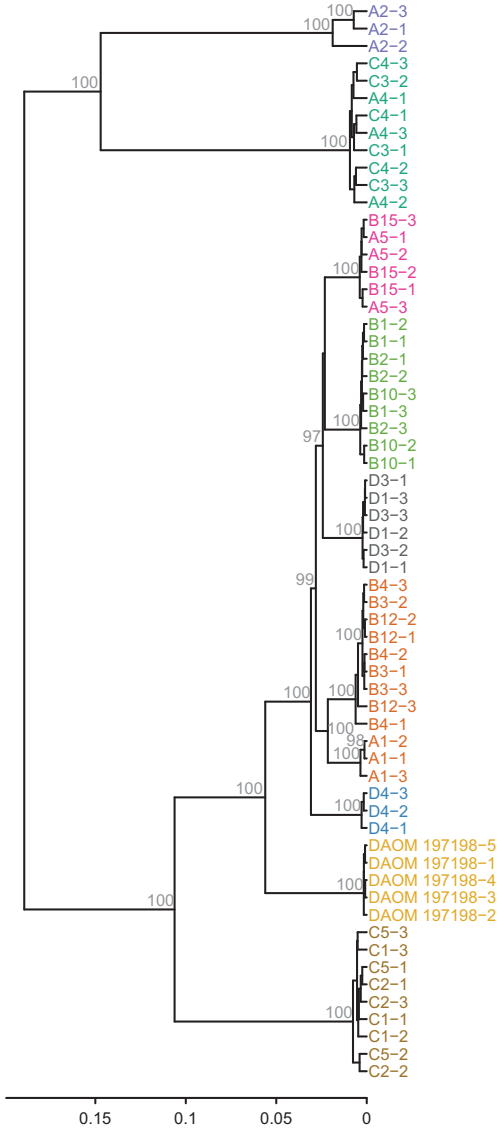

Figure 2 Hierarchical clustering of 20 isolates of *Rhizophagus irregularis* based on allele composition. Isolate DAOM 197198 was isolated in Canada, while all other isolates were isolated from a field in Switzerland, with 3–5 biological replicates per isolate. The distance between isolates was computed on 13 184 sites, based on the dot products of pairs of allele frequency vectors. Numbers associated with each node represent random sampling support of consensus tree bipartitions. Only support values greater than 95% are displayed.

and was significantly different among isolates, ranging from 0.78 sites/kb in DAOM 197198 to 3.01 sites/kb in C3 (one-way ANOVA, $F_{3,10} = 38.85$, $P < 0.001$; Figure 3b). The mean density of poly-allelic sites was 1.48 sites/kb (range 0.55–3.32 sites/kb), 4.98 sites/kb (range 2.20–9.01 sites/kb) and 7.94 sites/kb (range 4.20–11.07 sites/kb) in non-repeated-non-coding regions, repeated-non-coding regions and repeated-coding regions, respectively (Supplementary Figures 5A–C). When a higher coverage threshold was applied (30×), poly-allelic SNP density values were in a similar range than when a threshold of 10× was applied (Supplementary Results). The poly-allelic indel density was lower than the density of SNPs but the pattern was similar (Supplementary Figures 6A–E).

In haploid *S. cerevisiae* and *S. pombe,* we did not detect any sites with more than one allele (Figure 4, Supplementary Table 5). In the diploid *Candida albicans* DSY294 and SC5314, we observed 3.73 and 2.83 poly-allelic sites/kb, respectively. In the diploid *Betula nana*, we observed 3.37 and 3.10 poly-allelic sites/kb, while the tetraploid *Betula* spp. displayed poly-allelic site densities of >8 sites/kb (Figure 4).

The observed distribution of allele frequencies of poly-allelic SNPs peaked at 0.5 in the diploids *C. albicans* and *B. nana*, and peaked at 0.25, 0.5 and 0.75 in the apparently tetraploid *B.* x *intermedia*, as expected (Figure 5). Allele frequencies of most *R. irregularis* isolates did not follow a typical diploid or tetraploid distribution (Figure 5 and Supplementary Figure 7). We verified that the discrepancy between the clarity of peaks of allele frequencies in the diploids and tetraploids and the lack of clear peaks in *R. irregularis* was not due to differences in coverage. We artificially reduced coverage in the diploids and tetraploids by randomly removing reads. We obtained the same allele frequency distributions as with the full set of reads (data not shown).

Measurements of entropy, an independent method, gave the same results as poly-allelic SNP densities (Supplementary Results and Supplementary Figure 8).

Minimum haplotype characterization

In non-repeated-coding regions, *R. irregularis* isolates exhibited between 3.5% and 18.9% of loci with at least two haplotypes (Figure 6a). Tetraploid species had the highest proportion of loci with one, two or three haplotypes and a small proportion of loci with four haplotypes, as expected (Figure 6b). Diploids exhibited a majority of ddRAD-seq loci with one or two haplotypes, as expected (Figure 6c). Additional information is given in Supplementary Results, Supplementary Figure 9 and Supplementary Table 9.

Proportions of synonymous and non-synonymous substitutions in poly-allelic sites

We calculated the frequency of non-synonymous codons in transcript regions where more than two haplotypes were found, based on the ddRAD-seq data. Isolates DAOM 197198, A1, A2 and C2 had significantly lower proportions (ranging from 1% to 4%) of non-synonymous codon pairs than the diploid *B. nana* 097-10 (chi-squared test, $P < 0.001$). Isolate C3 comprised 9% of non-synonymous codon pairs, which was significantly greater than in DAOM 197198, A1, A2, C2 and the diploid *B. nana*, but not significantly different from *C. albicans* strains (Figure 7). Isolates C3, A4 and C4 showed the greatest proportions of non-synonymous codon pairs (ranging from 6% to 12%; Supplementary Figure 10, Supplementary Results).

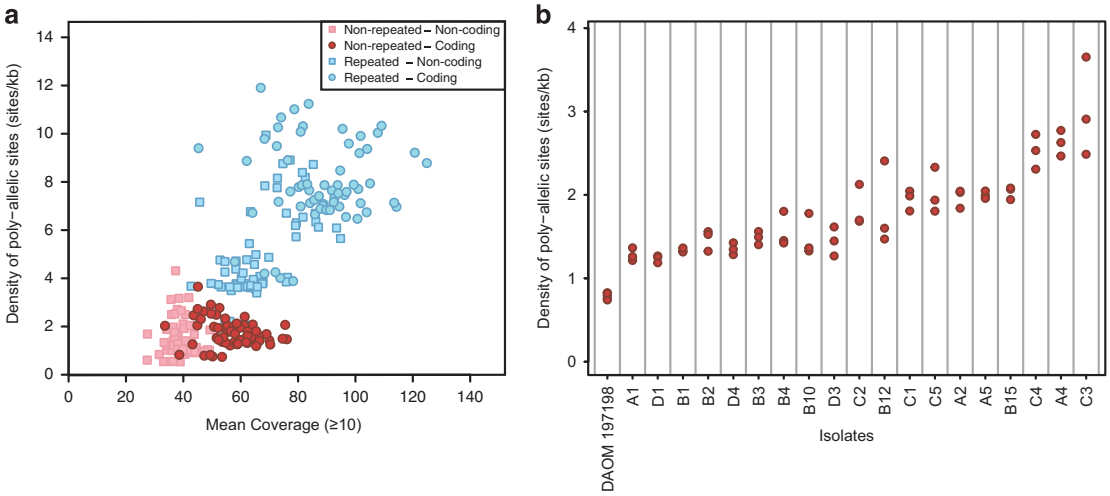

Figure 3 Density of sites with more than 1 allele in the 20 isolates of *Rhizophagus irregularis.* (**a**) Mean density of poly-allelic sites (considering only sites with SNPs) *versus* mean depth coverage (excluding ddRAD-seq loci with coverage lower than 10×). Biological replicates of each isolate are shown as separate dots. Each isolate is represented by four dots falling into four categories (non-repeated/non-coding, non-repeated/coding, repeated/non-coding, repeated/coding). (**b**) Density of poly-allelic sites (SNPs) categorized as non-repeated-coding for each biological replicate of each isolate. The red dots are the same as in (**a**), but separated according to isolate, without coverage information. Isolates are ordered in increasing mean density of poly-allelic sites from the left to the right.

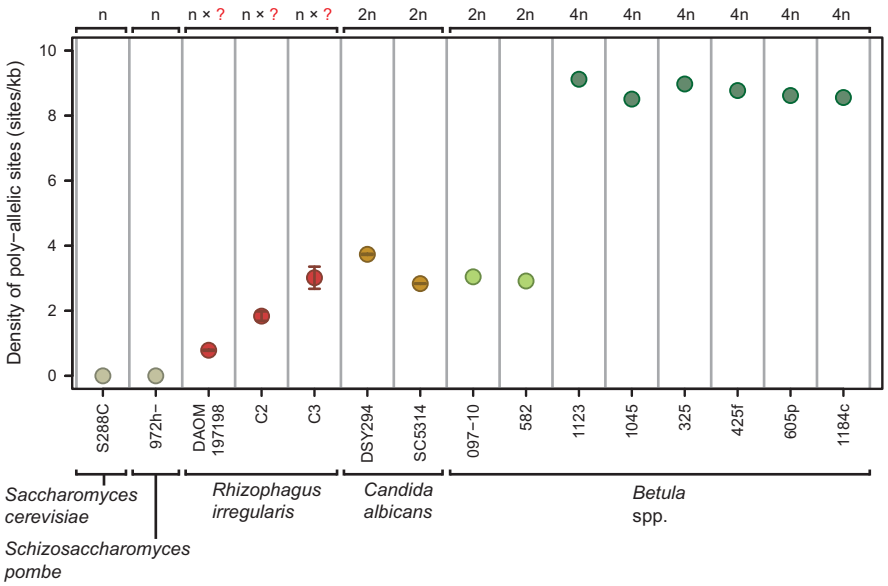

Figure 4 Density of poly-allelic sites in *Rhizophagus irregularis* compared to reference species. The reference species include: the haploid species *Saccharomyces cerevisiae* (S288C) and *Schizosaccharomyces pombe* (972 h-); *Rhizophagus irregularis* (DAOM 197198, C2, C3); the diploid species *Candida albicans* (DSY294, SC5314), and *Betula nana* (097-10, 582); the tetraploid hybrids *Betula* x *intermedia* (1123, 1045, 325) and *Betula pendula* (1148c); the tetraploid species *Betula pubescens* (425f, 605p). The density of poly-allelic sites is expressed as the number of sites/kb ± s.e. (when error bars are present) in non-repeated-coding regions.

Amplicon sequencing

Using amplicon sequencing, we observed single haplotypes in the loci *Bg348*, *Bg196* and *Bg235,* as expected (Figure 8). As expected, ddRAD-seq loci *3*, *6*, *8*, *10* and *11* presented several haplotypes per locus in at least one of the four isolates tested. Isolate C3 harbored between two and four haplotypes in ddRAD-seq loci *3*, *6*, *10* and *11*, while DAOM 197198, C2 and C5 only harbored one haplotype at these loci. ddRAD-seq locus *8* was polymorphic in C2, C3 and possibly in C5. In all loci except ddRAD-seq locus *3*, isolates C2 and C5 shared identical haplotypes that differed from haplotypes found in DAOM 197198 and C3 (Figure 8). The polymorphism observed in ddRAD-seq locus *7* was expected because it is located in a repeat. Amplicon sequencing confirmed that the same variants detected with ddRAD-seq were also detected using amplicon sequencing (Supplementary Figure 11, Supplementary Table 10).

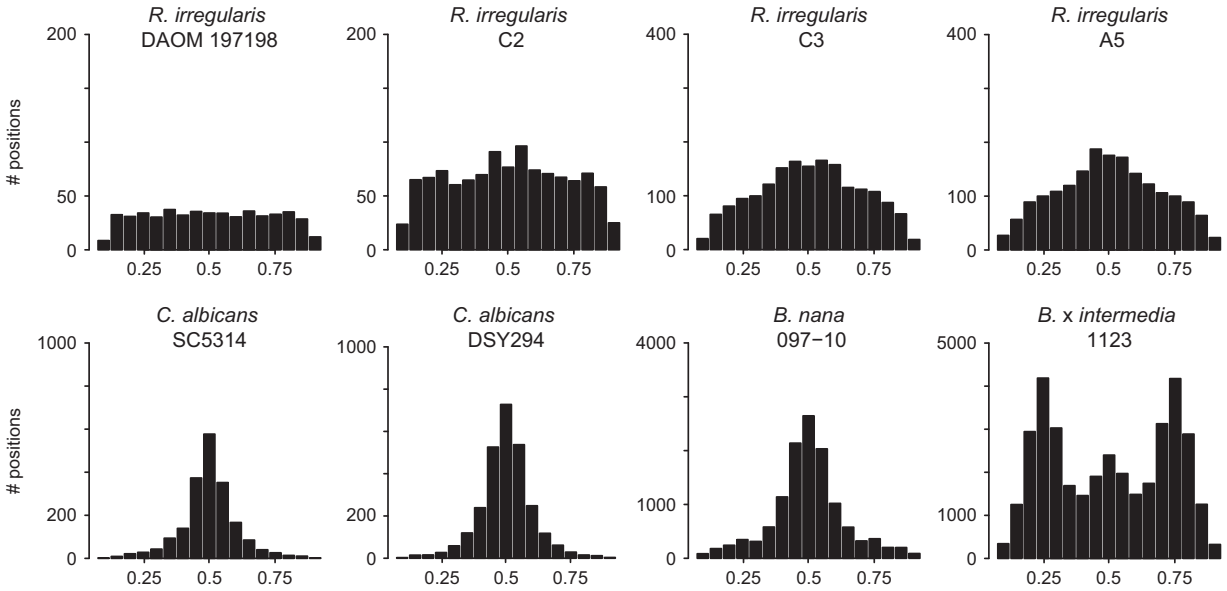

Figure 5 Distribution of allele frequencies in poly-allelic sites (SNPs) in non-repeated regions. Allele frequencies at sites that are identical among replicates and located in non-repeated regions are shown for four *R. irregularis* isolates, two diploids (*Candida albicans* and *Betula nana*) and one tetraploid *(Betula* x *intermedia*).

Discussion

Our study revealed a wide range of within-isolate polymorphism in a population of the AM fungus *R. irregularis*. We detected low levels of polymorphism within the reference isolate DAOM 197198, consistent with previous reports (Tisserant *et al.*, 2013; Lin *et al.*, 2014). We found that all Swiss isolates showed higher levels of within-isolate genetic polymorphism than DAOM 197198 when aligned against the DAOM 197198 genome assembly. However, the levels of within-isolate polymorphism we observed do not support the high numbers of alleles at individual loci found by Boon *et al.* (2015). We found that variation in coding regions was as high as in non-coding regions, suggesting that polymorphism within and among isolates of *R. irregularis* could be functionally important.

Robustness of the analyses

The challenge in detecting within-isolate polymorphism using next-generation sequencing techniques is to be able to identify which alleles are true variants. The data we generated are reliable. First, Illumina sequencing is not prone to consistent errors that would give rise to the same false SNP in different samples (Minoche *et al.*, 2011; Laehnemann *et al.*, 2015). Second, we sequenced at least three replicates of every isolate. A very high proportion of the SNPs we recorded were found in all three replicates. Therefore, the numbers of alleles that were only seen a small number of times was a very small fraction of the total alleles we found at poly-allelic sites (Supplementary Figure 2). The absence of poly-allelic SNPs in *S. cerevisiae* and *S. pombe* indicated that the ddRAD-seq protocol and Illumina sequencing that we used did not generate false poly-allelic sites. Fourth, in the case of experimental mixes of DNA from two *R. irregularis* isolates, allele frequencies reflected the expected proportions (Supplementary Results). Fifth, genetic relationships among isolates, based on the locations of poly-allelic sites alone (Supplementary Figure 4), were the same as when all loci were included in the analysis (Figure 2 and Supplementary Figure 3). The replicates would not have clustered together in the phylogenies if there were many false polymorphic sites in the data set. Finally, we are aware that the N6 genome assembly has a high number of contigs. Therefore, we used a greatly improved assembly of DAOM 197198 with around 1000 contigs (unpublished, Francis Martin, personal communication) and the assembly from another isolate (C2, unpublished, Nicolas Corradi, personal communication). The use of these two genomes did not change our conclusions.

A study questioning the results of Boon *et al.* (2015) suggests that the identification of true within-isolate polymorphism depends on the level of filtering (Ropars and Corradi, 2015). Ropars and Corradi (2015) advocate the use of Sanger sequencing to identify true variants from high-throughput sequencing results. However, direct Sanger sequencing of PCR products, without cloning, containing a mixture of different alleles in different frequencies requires intense optimization and does not allow detection of rare alleles, as illustrated by Kim *et al.* (2015) and references therein. The use of independent replication combined with reliable sequencing techniques such as Illumina allows the detection of low frequency alleles with much more certainty.

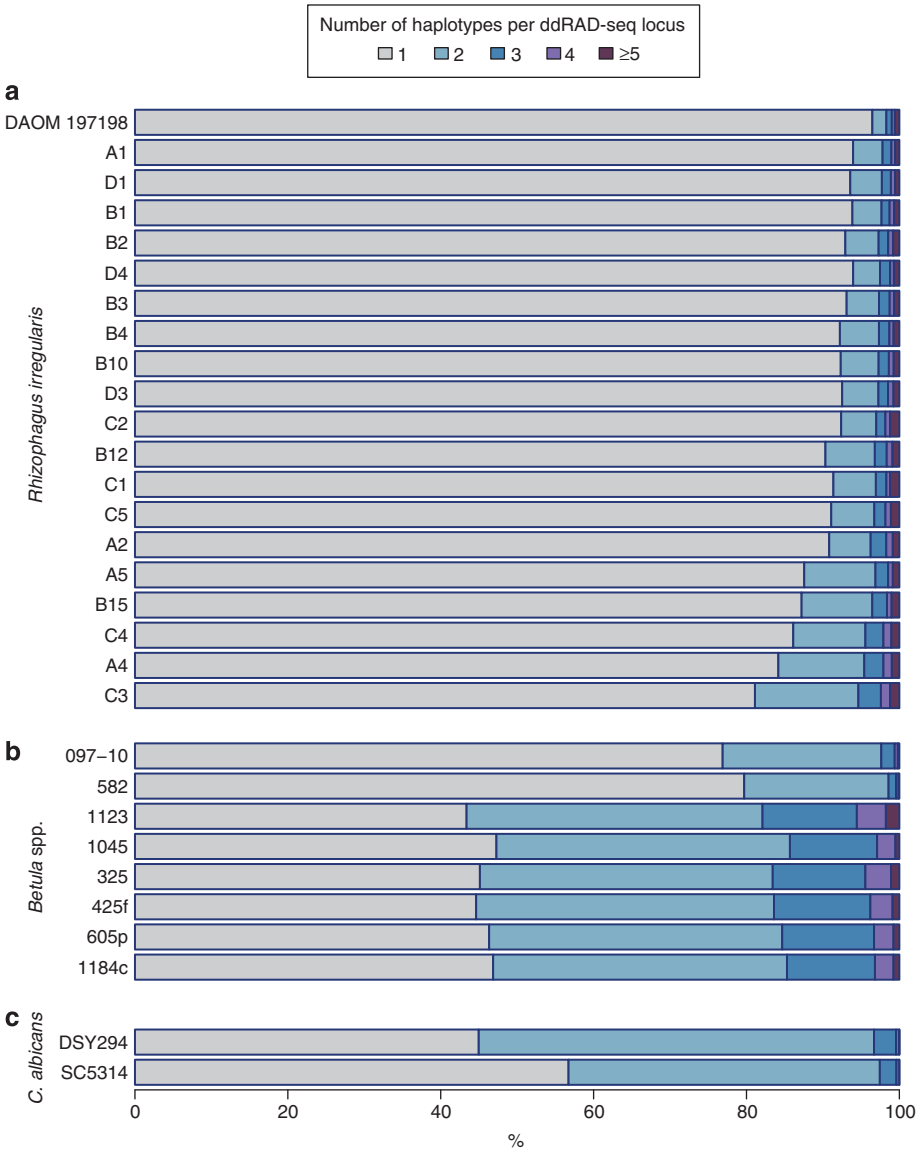

Figure 6 Percentage of ddRAD-seq loci with 1 to ⩾5 haplotypes per site. Sites within non-repeated-coding regions are shown (other categories are shown in Supplementary Figure 9). The number of haplotypes/ddRAD-seq locus per isolate is an average of 3–5 biological replicates (s.e. are not shown for ease of viewing). Numbers of haplotypes/ddRAD-seq locus should only be compared among individuals within a genus, but not among genera. (**a**) *Rhizophagus irregularis* isolates. (**b**) Diploids *Betula nana* (097-10, 582); tetraploids *Betula* x *intermedia* (1123, 1045, 325), *Betula pubescens* (425f, 605p) and *Betula pendula* (1148c). (**c**) Diploids *Candida albicans* (DSY294, SC5314).

Arrangement of within-fungus genetic variation in Rhizophagus irregularis

All of the *R. irregularis* isolates exhibited higher SNP densities than DAOM 197198, when mapped to the DAOM 197198 genome. Coupled with the distribution of allele frequencies (Figure 5, Supplementary Figure 7), our data indicate that some isolates may have a more complex genome than a strict homokaryotic haploid, or that the genomes of the isolates diverge greatly.

The polymorphism and allele frequency distributions observed might be caused by several phenomena that are not necessarily mutually exclusive. One possibility is that these isolates are heterokaryotic. This is supported by the density of SNPs and the number of haplotypes that occur in non-repeated regions of the genome, and the distributions of allele frequencies. A second possibility is that the observed polymorphism is not distributed among nuclei but among multi-copy regions within each nucleus. We aligned sequence reads of all isolates to the DAOM 197198 genome. If a considerable amount of copy number variation (CNV) existed among *R. irregularis* isolates, then sites that are single copy in DAOM 197198 may be multi-copy in the other isolates. Variation in the number and sequence of these copies could inflate the SNP density in other isolates relative to DAOM 197198. CNV in

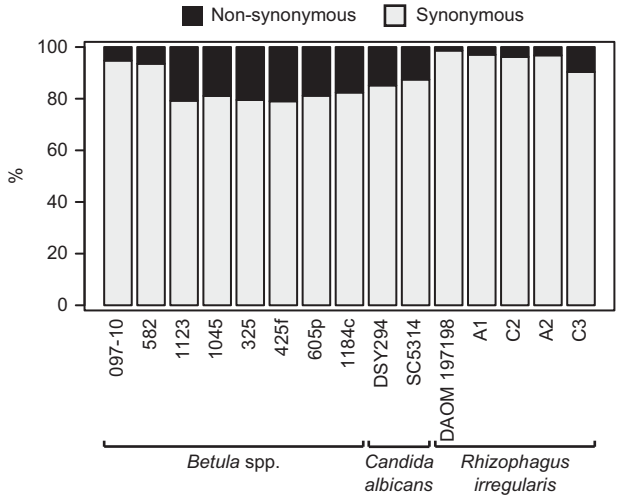

Figure 7 Proportions of synonymous and non-synonymous codon pairs within partial transcripts. Codon pairs were compared at coding ddRAD-seq loci showing two haplotypes in *Rhizophagus irregularis* (five isolates) and reference species with different levels of ploidy: diploids *Betula nana* (097-10, 582); tri- or tetraploids *Betula* x *intermedia* (1123, 1045, 325), *Betula pubescens* (425f, 605p) and *Betula pendula* (1148c), diploids *Candida albicans* (DSY294, SC5314). Synonymous codon pairs, gray bars; non-synonymous codon pairs, black bars.

R. irregularis has previously been observed among some of these isolates (Corradi *et al*., 2007). Isolates could vary, both in regions considered as single copy, or as repeated in DAOM 197198. A repeated region in DAOM 197198 could be single copy in other isolates, and vice versa. Finally, if one or more chromosomes are duplicated in some isolates but not in DAOM 197198, aneuploidy could explain the genetic diversity we observed.

Some isolates showed peaks of allele frequencies at 0.5 (for example, A4, A5, B15, C4 and C3), which would be expected if the isolates were diploid or heterokaryotic with two different nucleotypes. However, the allele frequency distributions of these isolates never showed the clear 0.5 allele frequency distribution of confirmed diploids or 50:50 mix of genomic DNA. Indeed, a high number of alleles existed in a frequency that deviated from 0.5. Three alternative scenarios to explain this would be: (1) A mixture of several genetically different nuclei with a predominance of two nuclear genotypes; (2) A heterokaryote with two different nucleotypes, that also contains significant CNVs in sites that are single copy in the reference genome; (3) Aneuploidy, with some chromosomes that are duplicated, combined with CNVs. Interestingly, the *B. nana* genome was found to contain the highest level of repeated regions (Supplementary Table 3) and yet allele frequencies still showed a clear 0.5 frequency peak (Figure 5, sample 097-10). Aligning the *B. pubescens* ddRAD-seq data to the *B. nana* genome also revealed a clear tetraploid pattern (Supplementary Figure 7, samples 425f and 605p). Thus, our strategy appears to sufficiently filter repeated elements that can interfere with the determination of ploidy.

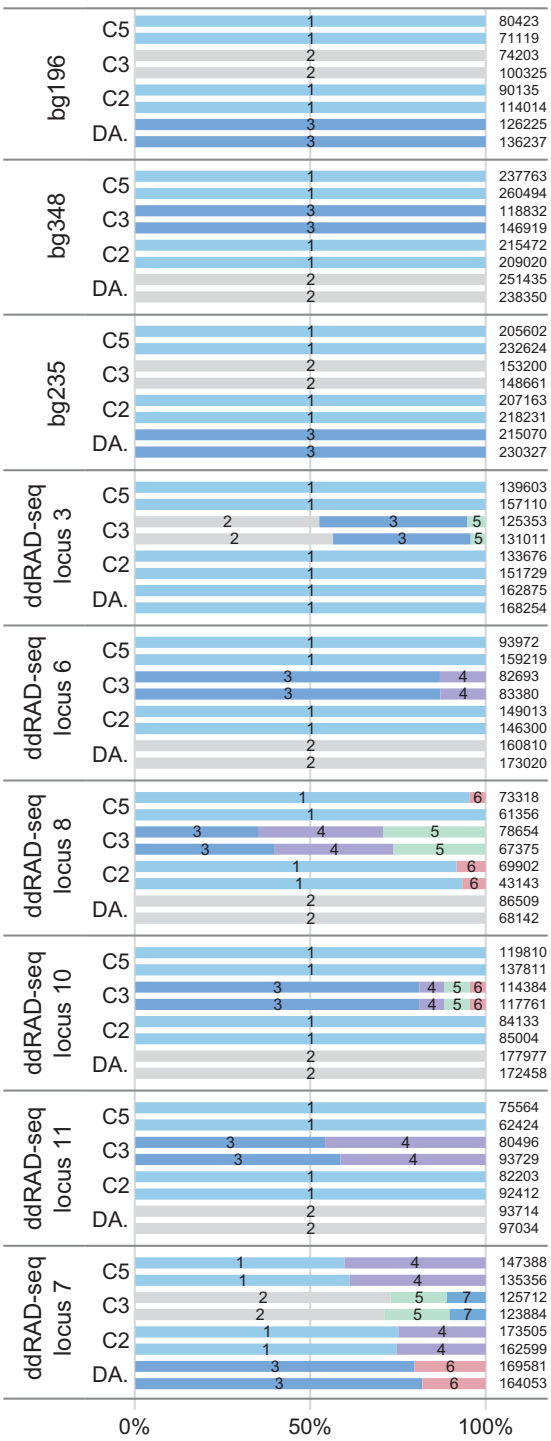

Figure 8 Proportion of different haplotypes found in nine loci within four isolates of *Rhizophagus irregularis*. Isolate DAOM 197198 (DA.) originated in Canada, while C2, C3 and C5 originated in Switzerland. The identity and proportion of haplotypes at each locus were determined on two independent DNA samples per isolate (represented by the two colored tracks per isolate). Each different color and number combination corresponds to a different haplotype sequence. Loci *Bg348*, *Bg196* and *Bg235* are microsatellite loci, and ddRAD-seq loci were selected from the ddRAD-seq data presented in this study. Numbers on the right side correspond to the total number of sequences analyzed.

Whether caused by differences in CNVs, heterokaryosis or aneuploidy, genetic polymorphism within and among *Rhizophagus irregularis* isolates other than DAOM 197198 is biologically interesting. Genetic variation in one fungal isolate gave rise to clonal progeny with differing allele frequencies, different phenotypes and different effects on plant growth (Angelard *et al.*, 2010; Ehinger *et al.*, 2012). Whether allele frequency changes were the result of spatial arrangement on different nuclei, of aneuploidy or of CNVs arising during the formation of new progeny, such variation is biologically interesting. Thus, documenting this variation within a population is important, irrespective of its location among or within nuclei. The range of within-isolate SNP densities that we reported may appear small. However, pairs of isolates such as C2 and B1 that differentially alter plant growth (Koch *et al.*, 2006) only differ by 3.70 mono-allelic SNPs/kb in non-repeated coding regions; a level of within-isolate genetic polymorphism similar in isolate C3.

Maintenance of genetic polymorphism in R. irregularis *populations*

Our study indicates that the genetic polymorphism in *R. irregularis* is maintained in populations. The high degree of congruence between the phylogenies based on mono-allelic and poly-allelic sites showed that the patterns of within-fungus polymorphism did not strongly diverge, and that random drift, remixing of heterokaryotic nuclei or dynamic generation of CNVs among isolates did not occur. Therefore, we conclude that the genetic polymorphism within each isolate has been maintained in the population over time.

Conclusions

Our analyses indicated that Swiss isolates of this important fungal species harbor more within-isolate polymorphism than that expected from the two existing whole-genome sequence studies on isolate DAOM 197198, but much less than that suggested by Boon *et al.* (2015). Although genetic polymorphism among and within isolates may appear small compared with other species (for example, *Rhizoctonia solani* AG8 (Hane *et al.*, 2014), *Puccinia striiformis* f. sp *tritici* (Cantu *et al.*, 2013)), such levels of variation are still of biological interest. The 1000 human genome project revealed that only 1 SNP/kb occurs among 1000 individuals coming from 14 ethnically different populations (1000 Genomes Project Consortium, 2012), and yet the phenotypic differences are considerable. In *R. irregularis*, sibling isolates with potentially small differences in within-isolate SNP densities had different phenotypes and strongly altered gene expression and growth in rice, the most globally important food plant (Angelard *et al.*, 2010; Angelard and Sanders, 2011; Colard *et al.*, 2011). For this reason, genetic polymorphism in *R. irregularis* isolates should not be ignored, as it could be ecologically a very important determinant of plant growth. Generation of genome assemblies from other isolates is now essential to determine the exact cause of genetic diversity in *Rhizophagus irregularis*. Characterization and use of genetic polymorphism opens the way for more efficient AM fungi particularly suited to globally important crops (Angelard *et al.*, 2010; Ceballos *et al.*, 2013).

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

We are grateful to Jeremy Bonvin and Ivan Dario Mateus Gonzalez who helped us with culturing and DNA extraction; Yves Poirier, Sophie Martin and Dominique Sanglard for providing cultures and DNA of non-AM fungal species; Alan Brelsford and Nicolas Perrin for providing us with Illumina P1 adapters and an earlier version of the ddRAD-Seq protocol. We thank Keith Harshman and Johann Weber (Lausanne University Genomic Technologies Facility), Nicolas Salamin and Thomas Junier for advice. We are grateful to Francis Martin and Nicolas Corradi for giving us privileged access to unpublished genome assemblies of isolates DAOM 197198 and C2. All bioinformatics computations were performed at the Vital-IT (www.vital-it.ch) Center for High Performance Computing of the Swiss Institute of Bioinformatics. This research was funded by a Swiss National Science Foundation grant to IRS (Bonus of Excellence 310030B_144079). ddRAD-seq reads were deposited in the NCBI SRA database (BioProject Accession Number: PRJNA268659). Barcode sequences are available in Supplementary Table 11.

References

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

Angelard C, Sanders IR. (2011). Effect of segregation and genetic exchange on arbuscular mycorrhizal fungi in colonization of roots. *New Phytol* **189**: 652–657.

Angelard C, Tanner CJ, Fontanillas P, Niculita-Hirzel H, Masclaux F, Sanders IR. (2014). Rapid genotypic change and plasticity in arbuscular mycorrhizal fungi is caused by a host shift and enhanced by segregation. *ISME J* **8**: 284–294.

Arnaud MB, Inglis DO, Skrzypek MS, Binkley J, Shah P, Wymore F *et al.* (2013). Candida Genome Database. Available from http://www.candidagenome.org/.

Baird NA, Etter PD, Atwood TS, Currey MC, Shiver AL, Lewis ZA *et al.* (2008). Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**: e3376.

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

Boon E, Zimmerman E, St-Arnaud M, Hijri M. (2013). Allelic differences within and among sister spores of the arbuscular mycorrhizal fungus *Glomus etunicatum* suggest segregation at sporulation. *PLoS One* **8**: e83301.

Börstler B, Raab PA, Thiéry O, Morton JB, Redecker D. (2008). Genetic diversity of the arbuscular mycorrhizal fungus *Glomus intraradices* as determined by mitochondrial large subunit rRNA gene sequences is considerably higher than previously expected. *New Phytol* **180**: 452–465.

Börstler B, Thiéry O, Sýkorová Z, Berner A, Redecker D. (2010). Diversity of mitochondrial large subunit rDNA haplotypes of *Glomus intraradices* in two agricultural field experiments and two semi-natural grasslands. *Mol Ecol* **19**: 1497–1511.

Cantu D, Segovia V, MacLean D, Bayles R, Chen X, Kamoun S *et al.* (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genom* **14**: 1–18.

Catchen J, Hohenlohe PA, Bassham S, Amores A, Cresko WA. (2013). Stacks: an analysis tool set for population genomics. *Mol Ecol* **22**: 3124–3140.

Ceballos I, Ruiz M, Fernández C, Peña R, Rodríguez A, Sanders IR. (2013). The in vitro mass-produced model mycorrhizal fungus, *Rhizophagus irregularis*, significantly increases yields of the globally important food security crop cassava. *PLoS One* **8**: e70633.

Colard A, Angelard C, Sanders IR. (2011). Genetic exchange in an arbuscular mycorrhizal fungus results in increased rice growth and altered mycorrhiza-specific gene transcription. *Appl Environ Microbiol* **77**: 6510–6515.

Corradi N, Croll D, Colard A, Kuhn G, Ehinger M, Sanders IR. (2007). Gene copy number polymorphisms in an arbuscular mycorrhizal fungal population. *Appl Environ Microbiol* **73**: 366–369.

Croll D, Wille L, Gamper HA, Mathimaran N, Lammers PJ, Corradi N *et al.* (2008). Genetic diversity and host plant preferences revealed by simple sequence repeat and mitochondrial markers in a population of the arbuscular mycorrhizal fungus *Glomus intraradices*. *New Phytol* **178**: 672–687.

Ehinger M, Koch AM, Sanders IR. (2009). Changes in arbuscular mycorrhizal fungal phenotypes and genotypes in response to plant species identity and phosphorus concentration. *New Phytol* **184**: 412–423.

Ehinger MO, Croll D, Koch AM, Sanders IR. (2012). Significant genetic and phenotypic changes arising from clonal growth of a single spore of an arbuscular mycorrhizal fungus over multiple generations. *New Phytol* **196**: 853–861.

Emerson KJ, Merz CR, Catchen JM, Hohenlohe PA, Cresko WA, Bradshaw WE *et al.* (2010). Resolving postglacial phylogeography using high-throughput sequencing. *Proc Natl Acad Sci USA* **107**: 16196–16200.

Garrison E, Marth G. (2012). Haplotype-based variant detection from short-read sequencing. arXiv preprint arXiv:1207.3907 [q-bio.GN].

Hane JK, Anderson JP, Williams AH, Sperschneider J, Singh KB. (2014). Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. *PLoS Genet* **10**: e1004281.

Hijri M, Sanders IR. (2005). Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. *Nature* **433**: 160–163.

Hoffman JI, Simpson F, David P, Rijks JM, Kuiken T, Thorne MAS *et al.* (2014). High-throughput sequencing reveals inbreeding depression in a natural population. *Proc Natl Acad Sci USA* **111**: 3775–3780.

Hohenlohe PA, Bassham S, Etter PD, Stiffler N, Johnson EA, Cresko WA. (2010). Population genomics of parallel adaptation in threespine stickleback using sequenced RAD tags. *PLoS Genet* **6**: e1000862.

Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB *et al.* (2004). The diploid genome sequence of *Candida albicans*. *Proc Natl Acad Sci USA* **101**: 7329–7334.

Katoh K, Standley DM. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol Biol Evol* **30**: 772–780.

Kim H, Erlich HA, Calloway CD. (2015). Analysis of mixtures using next generation sequencing of mitochondrial DNA hypervariable regions. *Croat Med J* **56**: 208–217.

Koch AM, Croll D, Sanders IR. (2006). Genetic variability in a population of arbuscular mycorrhizal fungi causes variation in plant growth. *Ecol Lett* **9**: 103–110.

Koch AM, Kuhn G, Fontanillas P, Fumagalli L, Goudet J, Sanders IR. (2004). High genetic variability and low local diversity in a population of arbuscular mycorrhizal fungi. *Proc Natl Acad Sci USA* **101**: 2369–2374.

Kuhn G, Hijri M, Sanders IR. (2001). Evidence for the evolution of multiple genomes in arbuscular mycorrhizal fungi. *Nature* **414**: 745–748.

Laehnemann D, Borkhardt A, McHardy AC. (2015). Denoising DNA deep sequencing data-high-throughput sequencing errors and their correction. *Brief Bioinform* **17**: 154–179.

Lange V, Böhme I, Hofmann J, Lang K, Sauter J, Schöne B *et al.* (2014). Cost-efficient high-throughput HLA typing by MiSeq amplicon sequencing. *BMC Genomics* **15**: 63.

Larsen BB, Chen L, Maust BS, Kim M, Zhao H, Deng W *et al.* (2013). Improved detection of rare HIV-1 variants using 454 pyrosequencing. *PLoS One* **8**: e76502.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N *et al.* (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**: 2078–2079.

Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D *et al.* (2014). Single nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genet* **10**: e1004078.

Magoč T, Salzberg SL. (2011). FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**: 2957–2963.

Minoche AE, Dohm JC, Himmelbauer H. (2011). Evaluation of genomic high-throughput sequencing data generated on Illumina HiSeq and Genome Analyzer systems. *Genome Biol* **12**: R112.

Munkvold L, Kjøller R, Vestberg M, Rosendahl S, Jakobsen I. (2004). High functional diversity within species of arbuscular mycorrhizal fungi. *New Phytol* **164**: 357–364.

Nielsen R, Paul JS, Albrechtsen A, Song YS. (2011). Genotype and SNP calling from next-generation sequencing data. *Nat Rev Genet* **12**: 443–451.

Novocraft-Technologies. (2014). Novoalign. Available from http://www.novocraft.com.

Öpik M, Zobel M, Cantero JJ, Davison J, Facelli JM, Hiiesalu I *et al.* (2013). Global sampling of plant roots expands the described molecular diversity of arbuscular mycorrhizal fungi. *Mycorrhiza* **23**: 411–430.

Paradis E, Claude J, Strimmer K. (2004). APE: Analyses of Phylogenetics and Evolution in R language. *Bioinformatics* **20**: 289–290.

Parchman TL, Gompert Z, Mudge J, Schilkey FD, Benkman CW, Buerkle CA. (2012). Genome-wide association genetics of an adaptive trait in lodgepole pine. *Mol Ecol* **21**: 2991–3005.

Pearson WR, Lipman DJ. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci USA* **85**: 2444–2448.

Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE. (2012). Double digest RADseq: an inexpensive method for de novo SNP discovery and genotyping in model and non-model species. *PLoS One* **7**: e37135.

1000 Genomes Project Consortium. (2012). An integrated map of genetic variation from 1092 human genomes. *Nature* **491**: 56–65.

Rambaut A. (2012). FigTree v1.4. Available from http://tree.bio.ed.ac.uk/software/figtree/.

Reitzel AM, Herrera S, Layden MJ, Martindale MQ, Shank TM. (2013). Going where traditional markers have not gone before: utility of and promise for RAD sequencing in marine invertebrate phylogeography and population genomics. *Mol Ecol* **22**: 2953–2970.

Ronquist F, Huelsenbeck JP. (2003). MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19**: 1572–1574.

Ropars J, Corradi N. (2015). Homokaryotic vs heterokaryotic mycelium in arbuscular mycorrhizal fungi: different techniques, different results? *New Phytol* **208**: 638–641.

Sanders IR, Croll D. (2010). Arbuscular mycorrhiza: the challenge to understand the genetics of the fungal partner. *Annu Rev Genet* **44**: 271–292.

Sanglard D, Ischer F, Calabrese D, Micheli Mde, Bille J. (1998). Multiple resistance mechanisms to azole antifungals in yeast clinical isolates. *Drug Resist Updat* **1**: 255–265.

Scaglione D, Acquadro A, Portis E, Tirone M, Knapp SJ, Lanteri S. (2012). RAD tag sequencing as a source of SNP markers in *Cynara cardunculus* L. *BMC Genomics* **13**: 3.

Sedzielewska KA, Fuchs J, Temsch EM, Baronian K, Watzke R, Kunze G. (2011). Estimation of the *Glomus intraradices* nuclear DNA content. *New Phytol* **192**: 794–797.

Smit AFA, Hubley R. (2008). RepeatModeler Open-1.0. Available fom http://www.repeatmasker.org.

Smit AFA, Hubley R, Green P. (1996). RepeatMasker Open-3.0. Available from http://www.repeatmasker.org.

Smith SE, Read DJ. (2008). *Mycorrhizal Symbiosis* 3rd edn. Elsevier Ltd.: London, UK.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* **18**: 1979–1990.

Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R *et al.* (2013). Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc Natl Acad Sci USA* **110**: 20117–20122.

van der Heijden MGA, Bardgett RD, van Straalen NM. (2008). The unseen majority: soil microbes as drivers of plant diversity and productivity in terrestrial ecosystems. *Ecol Lett* **11**: 296–310.

van der Heijden MGA, Klironomos JN, Ursic M, Moutoglis P, Streitwolf-engel R, Boller T *et al.* (1998). Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**: 69–72.

Wang N, Thomson M, Bodles WJA, Crawford RMM, Hunt HV, Featherstone AW *et al.* (2013). Genome sequence of dwarf birch (*Betula nana*) and cross-species RAD markers. *Mol Ecol* **22**: 3098–3111.

Wibberg D, Jelonek L, Rupp O, Hennig M, Eikmeyer F, Goesmann *et al.* (2013). Establishment and interpretation of the genome sequence of the phytopathogenic fungus *Rhizoctonia solani* AG1-IB isolate 7/3/14. *J Biotechnol* **167**: 142–155.

Supplementary Information accompanies this paper on The ISME Journal website (http://www.nature.com/ismej)