1 Investigating unexplained genetic variation and its expression in the arbuscular
2 mycorrhizal fungus *Rhizophagus irregularis*

3

4 Running header

5 Genome and transcriptome variation in AMF

6

7 Authors

8 Frédéric G. Masclaux*^{1,2}, Tania Wyss*¹, Marco Pagni², Pawel Rosikiewicz¹, Ian R.
 9 Sanders¹

10

11 Affiliations

12 ¹Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne,
 13 Switzerland

14 ²Vital-IT Group, Swiss Institute of Bioinformatics, 1015 Lausanne, Switzerland

15 * These authors contributed equally to this work.

16

17 Corresponding author

18 I. R. Sanders, Department of Ecology and Evolution, University of Lausanne, Biophore
 19 building, 1015 Lausanne, Switzerland

20 E-mail: ian.sanders@unil.ch

21

22

23 Summary

24 Arbuscular mycorrhizal fungi (AMF) are important symbionts of plants. Recently,
 25 studies of the AMF *Rhizophagus irregularis* recorded within-isolate genetic variation
 26 that does not completely match the proposed homokaryon or heterokaryon state (where
 27 heterokaryons comprise a population of two distinct nucleus genotypes). We re-
 28 analysed published data showing that bi-allelic sites (and their frequencies), detected in
 29 proposed homo- and heterokaryote *R. irregularis* isolates, were similar across
 30 independent studies using different techniques. This indicated that observed within-
 31 fungus genetic variation was not an artefact of sequencing and that such within- fungus
 32 genetic variation possibly exists. We looked to see if bi-allelic transcripts from three *R.*
 33 *irregularis* isolates matched those observed in the genome as this would give a strong
 34 indication of whether bi-allelic sites recorded in the genome were reliable variants. In
 35 putative homokaryon isolates, very few bi-allelic transcripts matched those in the
 36 genome. In a putative heterokaryon, a large number of bi-allelic transcripts matched
 37 those in the genome. Bi-allelic transcripts also occurred in the same frequency in the
 38 putative heterokaryon as predicted from allele frequency in the genome. Our results
 39 indicate that while within-fungus genome variation in putative homokaryon and
 40 heterokaryon AMF was highly similar in 2 independent studies, there was little support
 41 that this variation is transcribed in homokaryons. In contrast, within-fungus variation
 42 thought to be segregated among two nucleus genotypes in a heterokaryon isolate was
 43 indeed transcribed in a way that is proportional to that seen in the genome.

44 **Keywords:** homokaryon, heterokaryon, mycorrhizal symbiosis, arbuscular
 45 mycorrhizal fungi, genome variation, transcriptome variation

46

47 Introduction

48 Arbuscular mycorrhizal fungi (AMF) colonize the roots of the majority of land plants,
 49 improving plant nutrient and water uptake in exchange for plant-assimilated
 50 carbohydrates and lipids [1, 2]. Arbuscular mycorrhizal fungi form hyphae and spores
 51 that contain a continuous cytoplasm with many co-existing nuclei. There is no known
 52 stage during the AMF life cycle in which only one nucleus initiates a new generation,
 53 suggesting that a multinucleate organization is important [3].

54 Over the last two decades there has been some debate about whether AMF are
 55 homokaryons (possessing genetically identical nuclei) or heterokaryons (possessing
 56 two or more genetically different nuclei). The ensuing, and lengthy, debate about
 57 whether or not AMF are heterokaryons has largely confounded two separate questions.

58 1. Are AMF heterokaryons or homokaryons, without considering the quantity of
 59 genetic differences among nuclei? 2. How much genetic variation exists within an AMF
 60 isolate that could potentially be distributed among nuclei? Resolving this issue of
 61 genetic variation in this fungus is biologically highly relevant because clonally
 62 produced single spore offspring of *R. irregularis* show high variation in their phenotype
 63 [4] and extremely high variation in their effects on plant growth [5].

64 Studies over the last two decades that have tried to address the two questions outlined
 65 above are highly inconsistent in their conclusions regarding the amount of genetic
 66 variation existing within an AMF isolate and the prediction of the number of genetically
 67 different nuclei (summarized in Figure 1). However, there has been a lack of studies
 68 focused on the same species or isolate. Some species studied probably diverged from
 69 each other hundreds of millions of years ago, making comparisons between some of

these studies as largely meaningless [6]. Furthermore, different approaches have been used as technologies have developed over the years.

Figure 1. Range of reported levels of intra-isolate genetic variation in different AMF species and different studies. The level of intra-isolate genetic variation is expressed as the number of alleles observed per position in the genome or the predicted number of genetically different nuclei per AMF isolate on the grey horizontal bar. Levels of intra-isolate genetic variation were determined by sequencing methods. Green represents *Glomus etunicatum* although the studies did not use the same isolates. Blue represents *Rhizophagus irregularis* isolate DAOM 197198. Purple represents studies including isolates of *R. irregularis* other than DAOM 197198. It is important to note that within some studies a greater number of alleles was observed compared to the number of different nuclei predicted, owing to assumptions about ploidy or gene copy number.

More recently, studies using different techniques and sequencing platforms have concentrated on describing the genome of one haploid species of AMF, *R. irregularis*, and its within and among isolate variation [7-11]. Thus, there are now directly comparable data originating from independent studies. The whole genome of the *R. irregularis* isolate, DAOM197198, was sequenced showing that within fungus polymorphism existed but was low. The authors concluded that this isolate was a homokaryon [7, 12] and did not try and explain the variation. Double-digest restriction site-associated sequencing (ddRAD-seq) was performed on DAOM197198 and 19 other *R. irregularis* isolates [8], showing different levels of within-fungus polymorphism among the isolates that suggested isolates with a predominantly

94 homokaryon or heterokaryon state; where heterokaryons comprised predominantly
95 only two different genotypes of nuclei, but with the possibility of additional genetically
96 different nuclei at low frequency. Ropars *et al.* (2016) sequenced the whole genome of
97 five of the 19 isolates and concluded that three of these isolates; A1, B3 and C2 were
98 homokaryons and that two isolates; A4 and A5 were heterokaryons with two genetically
99 different nucleus genotypes [9]. Notably, none of these studies have recorded high
100 levels of polymorphism as those observed by Boon *et al.* [13, 14] (Figure 1). Despite
101 this, Bruns *et al.* (2018) [15] suggested a strong dichotomy between, on the one hand,
102 the studies of Boon *et al.* and Wyss *et al.* [8, 13, 14] and on the other hand, the studies
103 by Ropars *et al.*, Lin *et al.* and Tisserant *et al.* [7, 9, 12] (see discussion by Sanders
104 2018).

105 All studies revealed a level of polymorphism unexplainable by a strict homokaryon
106 state. In each case, bi-allelic sites existed in isolates that had been predicted to be
107 homokaryons. However, this variation was largely discounted as being due to
108 sequencing errors. Plots of allele frequency at bi-allelic sites in proposed homokaryons
109 and heterokaryon isolates revealed a peak at 0.5 in several isolates but with additional
110 variation at other frequencies[8]. A clear 0.5 peak of frequency at bi-allelic sites is
111 expected for a heterokaryon with equal proportions of two nucleus genotypes or a
112 diploid. Ropars *et al.* (2016) found a peak at 0.5 frequency in A4 and A5, but a U-
113 shaped distribution in the other isolates, which is difficult to explain and not consistent
114 with [8, 9].

115 Wyss *et al.* (2016) observed within-fungus genetic variation among biological
116 replicates of the same isolates, thereby, reducing the probability that the variation was
117 the result of sequencing error [8]. Ropars *et al.* (2016) also constructed replicate

libraries for whole genome sequencing, although they pooled them for further analysis, and did not use them as replicates[9]. We hypothesized that bi-allelic sites observed in the genome of a given isolate in a replicated study, and between the two independent studies, will be consistent if they did not arise from sequencing error. Thus, observed variation represents true variation and is not due to sequencing error. Using the replicates of Wyss et al. and the data from Ropars et al., treated as replicates, allows to test whether the unexplained variation in these studies is indeed real biological variation or an artefact of sequencing. This was the first goal of this study.

So far there has been almost no attention paid to the functional significance of any within-fungus genome variation observed in AMF. However, single spore offspring of clonally produced *R. irregularis* shows a high degree of variation in their phenotype and their effects on plant growth [4, 5]. Transcriptome profiling of *R. irregularis* would reveal whether the genome variation observed at bi-allelic sites is transcribed or whether unequal or single-allele expression occurs. However, as yet, no studies exist in AMF relating variation at the transcriptome level back to variation observed at the genome level. The detection of bi-allelic transcripts that match observed variation in the genome would confirm the existence of bi-alleles in the genome and would indicate that the variation could potentially have consequences on the phenotype. Expression of two alleles of a given gene, unequal expression of the two alleles, or expression of only one of them, could all have consequences on the phenotype of the fungus or its effect on plant growth. If alleles were truly located on two genetically different nuclei, it would mean that it would also represent a measure of the contribution by the two different AMF nucleus genotypes to overall gene expression in the fungus. In plants, the contribution of two different genomes (for example in hybrids or polyploids) is often highly unequal and the contribution from each genome can be tissue specific [16].

Thus, the existence of two different genotypes of nuclei does not necessarily mean that gene expression will be in equal proportions from the two alleles at a given locus. Thus, the second goal of our study was, therefore, to investigate whether variation observed in the *R. irregularis* transcriptome matched variation observed in the genome as an additional test of whether the variation really exists. Additionally, we wanted to know the contribution of each nucleus genotype to overall gene expression in this fungus.

We, therefore, used the available genome data on *R. irregularis* to compare within-isolate genome polymorphism and test whether the unexplained variation in these studies is indeed real biological variation or an artefact of sequencing. We also conducted transcriptome profiling on the isolate C3 (that has been predicted to be a heterokaryon with two nucleus genotypes) and the isolate C2 (that has been predicted to be a homokaryon) to see if variation found at the genome level is mirrored at the transcriptome level, i.e. whether this variation is co-expressed, and in what proportion the alleles are expressed. Finally, we looked at the extent of structural variation occurring in the genomes of these isolates because structural variation can also strongly affect the phenotype of an organism. Although heterokaryon can refer to the co-existence of two or more genetically distinct nuclei, here we refer to heterokaryon isolates as those possessing primarily two genetically different nuclei.

162 **Materials and Methods**

163 *Fungal material, RNA extraction and library preparation*

164 Three *R. irregularis* isolates (C2, C3 and DAOM197198) were cultured *in vitro* with
 165 *Ri* T-DNA transformed carrot roots [17] and a split plate system [18]. Fungal material
 166 was extracted from the medium in citrate buffer (450 mL ddH₂O, 8.5 mL 0.1 N citric
 167 acid, 41.5 mL 0.1 N Na citrate) for 20 min and washed with sterile double-deionized
 168 water (ddH₂O). Material from three Petri plates was pooled per sample. There were
 169 three samples per isolate. RNA was extracted with the RNeasy Plant Mini kit (Qiagen).
 170 Two micrograms of RNA from each sample was used to prepare libraries for RNA
 171 sequencing (RNA-seq) using the TruSeq Stranded Total RNA Library Prep Kit
 172 (Illumina) with a PCR enrichment step of 15 cycles. The libraries were pooled and
 173 sequenced using Illumina HiSeq 2500 (100 bp paired-end reads). Sequences were
 174 deposited in the NCBI SRA database (BioProject Accession Number: [PRJNA509102](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA509102)).

175 *All data used in this study*

176 The origin and characteristics of whole genome (WG), ddRAD-seq and RNA-seq data
 177 from *R. irregularis* and other species that was used in this study is summarized in
 178 Tables S1-S3. The name of the isolate “DAOM197198” is shortened to “DAOM” in
 179 some of the figures.

180 Additionally, in order to find examples to test the RNA-seq KisSplice pipeline
 181 (described below), we screened the NCBI sequence archive to select RNA-seq data of
 182 species that have homokaryon and heterokaryon states or have homozygote and
 183 heterozygote states (Table S3). The idea was to find species where both states exist like
 184 *Arabidopsis* homozygote (*Col-0* accession) and *Arabidopsis* heterozygote (Hybrid line).

Some other fungal species with known homokaryon and heterokaryon states were also included (Table S3).

Processing sequence data

Raw reads were processed with the script Tagcleaner.pl to trim Illumina adapters [19]. Reads were processed with PrinSeq-lite.pl version 0.20.4 [20] to remove reads containing ‘Ns’, to trim low quality 3’-ends and retain reads longer than 50 bp. A summary of the read size after filtering is in Table S4.

In silico prediction of ddRAD-seq fragments

To compare the same sites in ddRAD-seq and WG data, analysis was restricted to ddRAD-seq fragments existing in WG data. *In silico* digestion with *EcoRI* and *MseI* was performed on each genome assembly to define predicted ddRAD-seq fragments longer than 49 bp and delimited by a *EcoRI* site and a *MseI* site. Predicted fragments were aligned to their respective genome with Novoalign v3.02.00 (Novocraft-Technologies, 2014). Fragments that could not be re-aligned were removed. Re-aligned predicted fragments define regions that were considered in subsequent analyses. We refer to ‘predicted ddRAD-seq fragment’ as one of these fragments and to ‘predicted ddRAD-seq region’ to designate the location where the fragment mapped (Figure S1).

Identification of interspersed repeats and repeated elements

We use the term ‘repeats’ to include the interspersed repeats (mobile elements and transposon-like regions) and repeated elements (multicopy genes and duplicated regions). The interspersed repeat families were predicted *de novo* with the program RepeatModeler Open-1.0.8 (<http://www.repeatmasker.org>) and interspersed repeats were annotated with RepeatMasker Open-4.0.6 (<http://www.repeatmasker.org>).

Every method of repeat detection has its own stringency, which has direct consequences on the number of bi-allelic positions that can be discovered. We wanted to test that identification of poly-allelic sites was not strongly biased by the method of repeat detection. Thus, we used three methods to identify repeated elements to test the influence of underestimating or overestimating the repeated elements on within-isolate polymorphism.

Method M03 was used previously [8]. 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 [21]. This was done to identify globally similar predicted fragments among the ddRAD-seq fragments. Fragments having a match with another fragment were labelled ‘repeated elements’. The M03 repeats corresponded to interspersed repeats and M03 repeated elements. Method M03 is the default method used in this study (Figure S1).

Method M07 had a larger spectrum than M03. *In silico* predicted ddRAD-seq fragments were aligned to the whole reference genome using glsearch36 (package fasta-36.3.5e) to identify fragments matching more than one time in the reference genome. Fragments with more than one match were labelled ‘repeated elements’. The M07 repeats corresponded to a combination of interspersed repeats and M07 repeated elements. It is the most stringent of the three methods.

With a third method, M12, *in silico* predicted ddRAD-seq fragments were mapped to the genome assembly with Novoalign V3.02.00 using parameters -r All 100 -R 400 -s0. Fragments with more than one match in the assembly were labelled ‘repeated elements’. The M12 repeats were a combination of interspersed repeats and M12 repeated elements. This method is of medium stringency.

232 *Gene prediction in genome assemblies*

233 Coding regions were defined with the *ab initio* gene predictor Augustus version 3.1
234 using the gene model previously defined for the N6 assembly, constructed for the
235 DAOM197198 isolate [8].

236 *Mapping of WG and ddRAD-seq sequence data on reference genomes*

237 Wyss et al. (2016) mapped reads to the DAOM197198 genome assembly because other
238 genome assemblies of *R. irregularis* isolates were not available at the time [8]. For this
239 study, we used genome assemblies of isolates A1, A4, A5, B3 and C2 for mapping. We
240 aligned reads against the genome assemblies with Novoalign v3.02.12 (Novocraft-
241 Technologies, 2014). The correctly mapped paired-end reads were selected for
242 subsequent analyses using Samtools version 1.3 [22]. Single nucleotide polymorphisms
243 (SNPs), insertions and deletions (indels) and multiple nucleotide polymorphisms
244 (MNPs) were called for each sample using Freebayes version 1.0.2 [23] at positions
245 with a minimum 10× depth of coverage. Only alleles with frequencies greater or equal
246 to 0.1 were recorded. Variant files were filtered to keep the positions with a phred-
247 scaled quality score greater or equal to 30. For every sample, and at each site, we
248 reported whether there was a single allele, two alleles or more than two alleles. A
249 missing value (NA) was assigned if the depth of coverage was below 10×. Because the
250 objective of the study was to study all the within-isolate polymorphism, no filter for the
251 allelic ratio was applied to a position to call it bi-allelic or poly-allelic position.

252 We compared variants in ddRADseq data [8] and WG data [9]. If the same variants are
253 found among replicates and between two independent studies, then it is highly unlikely
254 that the observed variants are artefacts caused by error generated in the sequencing.
255 Wyss et al. (2016) included 3 replicates per isolate. Ropars et al. (2016) assembled

genomes of *R. irregularis* that were pooled samples of different libraries that can also be treated as replicates. The numbers of replicates in WG sequencing was between 2 and 3, depending on the isolate.

Bi-allelic site density and allele frequency

In order to only consider variants in single copy genes, the variant file of each sample was filtered to only retain positions with two alleles in non-repeated and coding regions. The number of bi-allelic positions was divided by the number of positions from predicted ddRAD-seq regions that were sequenced with a coverage greater or equal to 10.

All positions from the variant file were included except those in repeated regions defined by RepeatModeler/RepeatMasker and with less than 10× coverage. The threshold was altered to test the effect of depth of coverage on allele frequency distribution. The list of all allele frequencies was plotted using the density function from the R package ggplot2.

Depth of coverage and predicted ddRAD-seq regions

Low coverage in samples would limit the ability to reliably detect polymorphisms. Samtools version 1.3 [22] was used to calculate depth of coverage in ddRAD-seq and WG data. Samtools and custom perl script were used to calculate the mean depth of coverage per predicted ddRAD-seq region. A mean coverage greater, or equal to, 10 was necessary to consider a predicted ddRAD-seq region as covered. A similar method was used to calculate the depth of coverage per position required to calculate bi-allelic position density.

Identification of common bi-allelic positions

The first goal of this study was to compare whether the same bi-allelic sites were found among replicates and studies. The R package UpSetR [24] was applied on lists of bi-allelic positions in non-repeated, coding regions to find common bi-allelic positions in different replicates and generate UpSet plots.

Discovery of bi-allelic SNPs in RNA-seq data

KisSplice performs a local assembly of RNA-seq reads and detects variants directly from a De Bruijn graph, independently of a reference genome [25]. The methodology to analyse RNA-seq data is summarized in Figure S2. Because KisSplice reports a very local context around SNPs, we generated reference full-length transcriptome assemblies from RNA-seq data with Trinity [26]. Trinity v2.3.2 was run with standard parameters (except --SS_lib_type RF). Predicted ORFs were identified in the Trinity assembled transcripts with TransDecoder (v5.0.2). Next, we identified full-length, or nearly full-length, transcripts (<https://github.com/trinityrnaseq/trinityrnaseq/wiki/Counting-Full-Length-Trinity-Transcripts>, accessed on July 2018). The assembled transcripts were compared to known proteins in Swissprot UniProtKB proteins (www.uniprot.org, accessed on June 2017). The comparison was performed with BLASTX (2.6.0+) with the following parameters: -evalue 1e-20 -max_target_seqs 1 -outfmt 6. The length of the top hit and percentage of the hit length, included in the alignment to the Trinity transcript, were added to the BLASTX result file with the provided script analyze_blastPlus_topHit_coverage.pl. Next, a custom Perl script was applied on the resulting file to keep only the best isoform of each gene and to remove genes with multiple groups (following Trinity's specifications for 'group' and 'isoform'). KisSplice version 2.4.0 was run on RNA-seq data of each organism with the

parameters: --experimental -s 2 -k 41. We focused on the ‘Type0a bubbles’ (KisSplice nomenclature), that are two identical short sequences differing by a single SNP at a given position. Bubble sequences were positioned on the Trinity-reference transcriptome using BLAT with the option -minIdentity=80. To predict amino acid changes of a SNP, the program KisSplice2RefTranscriptome (K2RT) was run taking as input the predicted ORFs, the Type0a bubbles produced by KisSplice, and the mapping results of BLAT. A custom Perl script removed bubbles that matched two sequences in the reference transcriptome and those with mapping ambiguity. The script also removed bubbles with a read coverage less than six and bubbles with one of the two alleles having less than two observations. Bubbles matching a transcript without a BLASTX hit were removed. These filtering steps ensured that sets of SNPs had the highest possible quality. The KisSplice/K2RT pipeline reports counts of both alleles at bi-allelic positions. Using these values, the distribution of allele frequencies at bi-allelic positions was plotted for each organism.

We tested the pipeline on RNA-seq data generated from different species with known ploidy or heterokaryon/homokaryon status to demonstrate the pipelines efficiency and robustness. The pipeline was applied on *R. irregularis* RNA-seq data to evaluate within-isolate polymorphism in RNA transcripts.

Analysis of bubbles and bi-allelic transcripts

Transcript and bubble sequences were mapped to the reference genome with BLAT using the parameter -minIdentity=90. The alignments were parsed to quantify the proportion of uniquely mapped sequences.

To investigate whether bi-allelic positions discovered in RNA-seq data matched to bi-allelic sites in the genome, bubble sequences containing 2 alleles were mapped to the

reference genome to provide the genomic coordinates of the bi-allelic positions. Next, allelic composition of each bi-allelic position in bubbles was compared to the allele composition of the position in the WG data. Any difference in allele composition was recorded as “inconsistent”. An “inconsistent” position is a bi-allelic position detected in RNA-seq that was not bi-allelic in WG data or that had a different allele composition to the WG data.

Structural variation

WG samples (A1-WG4, A4-WG0, A4-WG9, A5-WG7, B3-WG6, C2-WG0, C2-WG3, DAOM-DNA1 and DAOM-DNA2), with a mean coverage greater than 20 were analysed to detect structural variations (SV). The calling of SV was performed using Lumpy version v2.16.2 [27]. This integrates multiple SV signals: read-depth variation events, paired-end events and split-read events. A new read mapping was performed with SpeedSeq version 0.1.2 [28]. All read sets were mapped to every genome to generate all possible combinations. The resulting .bam files were processed with Lumpy to call variants. Results were filtered to keep scaffolds with a size greater or equal to 10000, SV events that were supported by at least nine total events, three paired-end events and three split-read events. Only deletions detected with Lumpy were considered in this study.

SV calling was also performed using CNVnator [29]. The .bam files generated with novoalign to call the variants were reused to call the SV with CNVnator version 0.3.3. Results were filtered using the fthresholds: scaffold size greater or equal to 10000, p-value 1 ($p\text{-val1}$) < 0.05, q0 value < 0.5. Only deletions detected with CNVnator were considered in this study.

Genome and transcriptome variation in AMF

350 Genes overlapped by a deletion were found using BEDtools intersect. The script
351 sma3_v2.pl was run on proteins deduced from genes to perform functional annotation
352 and gene ontology enrichment analysis.

353

Results

Level of within-isolate polymorphism depends on the identity of the reference genome

The identity of the reference genome indeed had a strong impact on the density of poly-allelic positions (Figure S3), with the lowest density when ddRAD-seq data were mapped onto the respective genome assembly.

Variants in ddRAD-seq and WG occurred at the same positions

The mean depth of coverage across the samples ranged from 22 to 33 and from 13 to 47 for ddRAD-seq and WG data, respectively (Figure S4). A greater number of predicted ddRAD-seq regions in the five isolates were covered by reads from WG data than by ddRAD-seq data (Figure S5). Poly-allelic positions were almost all bi-allelic. Tri-allelic positions represented approximately 0.5% of the poly-allelic sites. Only one tetra-allelic position was found (Table S5). Bi-allelic positions were found in all isolates and both types of data (Figure 2a). The density of bi-allelic positions ranged from 0.66 to 1.79 bi-allelic positions/kb (Figure 2b). In general, the density of bi-allelic positions was higher when the depth of coverage was higher (Figure 2b). The majority of bi-allelic sites that were found in the ddRAD-seq datasets also occurred in the WG datasets (Figure 3). In most cases, these occurred in all replicates of the two datasets or in several replicates of the ddRAD-seq data and WG data, confirming that it is very unlikely that they occurred because of random error (Figure 3). Hypergeometric tests showed that the overlap of which fragments from ddRAD-seq and WG mapped to the in silico predicted RAD fragments was highly unlikely to be due to chance (Cumulative hypergeometric test (function *phyper*), $P < 0.00001$, in all isolates). Alternative approaches to determining non-repeated regions yielded very similar results (Figures S6 and S7) meaning that the detection of within fungus SNPs was not strongly sensitive

to the method of repeat detection. There were many more bi-allelic positions in the WG data that were not detected in the ddRAD-seq data simply because so many more predicted ddRAD-seq regions were covered by reads in WG data than in RD data (see Figure S5).

Figure 2. Detection of poly-allelic positions in whole genome (WG) and ddRAD-seq data. (a) Number of bi-allelic positions among replicate samples of the different isolates. (b) Mean density of poly-allelic positions (composed of SNPs) versus mean depth coverage (excluding ddRAD-seq loci with coverage lower than 10×) in predicted ddRAD-seq regions which are coding and non-repeated (using method 03). "WGx" corresponds to the whole genome sequencing data, where "x" is the replicate identifier. "RSx" corresponds to the ddRAD-seq data, where "x" is the replicate identifier. Replicates of each isolate are shown as separate dots (RS) or triangle (WG).

Figure 3. Comparison of bi-allelic sites found in replicates of WG and ddRAD-seq data in five *R. irregularis* isolates. Number of replicate samples from WG and ddRAD-seq data that contain the same bi-allelic sites in each isolate. The numbers above the bars in the histograms represent the number of bi-allelic sites shared among a given set of replicates from WG and ddRAD-seq. Bullets connected by vertical lines below the histograms show which replicates contained the same bi-allelic positions. The Up-set plots show numbers of common positions among samples ranked from highest on the left of the graph for the highest 15 combinations of samples. A very small number of bi-allelic sites were found among other samples that are not shown but these numbers are so low as not to alter the conclusions that can be drawn from this analysis.

402 *Allele frequency similarities*

403 Assuming nuclei are haploid, allele frequency distribution at bi-allelic loci should
 404 reveal the number and proportion of different nuclei. Very similar allele frequency
 405 distributions were observed among the replicates of each isolate originating from WG
 406 and ddRAD-seq data (Figure 4; Figure S8). A peak at 50:50 was detected in A4 and A5
 407 (Figure 2b), as expected for a 50:50 heterokaryon and as shown previously [9]. There
 408 was a slight peak at 50:50, as well as slight peaks at approximately 20:80 and 80:20, in
 409 A1, B3 and C2 (Figure 4). It is also noticeable that the small peaks at 20:80 and 80:20
 410 also existed in the isolates A4 and A5 despite a clear peak at 50:50. The isolate
 411 DAOM197198 was also subjected to the same analysis. These results are presented in
 412 the Figure S9.

413 **Figure 4. Comparison of distribution of allele frequencies in five *R.***
 414 ***irregularis* isolates observed in WG and ddRAD-seq data.** Distribution of
 415 allele frequencies at poly-allelic positions (composed of SNPs) in non-repeated
 416 and coding regions (using method 03) of the isolates A1, A4, A5, B3 and C2.
 417 Dark coloured lines represent WG data and clear lines ddRAD-seq data.

418 *Effects of coverage and other parameters*

419 We investigated how different parameters can affect the discovery of bi-allelic positions.
 420 All highest-covered samples had a depth of coverage lower than 50× (Figure S4). In
 421 Figure 2b, the density of bi-allelic positions was only reported in non-repeated, coding
 422 regions of predicted ddRAD-seq fragments and was higher when the depth of coverage
 423 was higher. The number of bi-allelic positions was measured in the whole genome and
 424 not only in the predicted ddRAD-seq regions (Figure S10). The greatest numbers of
 425 poly-allelic positions were found in samples having the highest coverage depth.

Increasing the depth of coverage by sequencing more deeply lead to higher number of bi-allelic positions and no saturation was detected (Figure S10).

We applied different thresholds on the depth of coverage in order to evaluate the effects of coverage on the discovery of bi-allelic positions and on patterns of allele frequencies (Figure S11). Increasing the coverage thresholds had a strong impact on allele frequency distribution. At higher thresholds, the signal around 50:50 disappeared, including the strong 50:50 peaks found for A5 and A4. High thresholds led to a U-shape distribution similar to what was observed by Ropars et al. (2016) [9].

Bi-allelic positions within RNA transcripts

We used RNA-seq data to explore whether within-isolate polymorphism was detectable in transcripts. We first tested the pipeline on data generated from different species for which the ploidy level is known or for which the states homokaryon, heterokaryon, homozygote and heterozygote is already known. The pipeline clearly demonstrated its efficiency to distinguish heterozygotes from homozygotes, and heretokaryons from homokaryons (Figure S12 and Supplementary information).

The number of transcripts assembled independently from the reference genome was greater in the isolate C3 than in the isolates C2 and DAOM197198 when no filtration was applied (Figure 5a). However, the number of genes deduced from Trinity-assembled transcripts was very similar in the three isolates (Figure 5a). The number of genes recovered in the transcripts represented 71%, 77%, and 71% of the 22 812 genes that are thought to be present in the *R. irregularis* genome in C2, C3 and DAOM 197198, respectively. This meant that the RNA-seq datasets from these isolates should have allowed the recovery of a significant proportion of the bi-allelic transcripts, if they exist. Transcripts with bi-allelic positions were found in all isolates (Figure 5b), with a greater

proportion in C3 than in C2 and DAOM197198. Detailed results can be found at <https://github.com/FredMasc/PAAT2019>. The percentage of bi-allelic sites with synonymous codons was 54%, 41% and 37% in C3, C2 and DAOM, respectively.

Figure 5. Detection of bi-allelic positions in RNA-seq data. (a) Number of transcripts (black) and genes (grey) obtained with the de-novo transcript assembly approach. (b) Percentage of transcripts containing bi-allelic positions in isolates C2, C3 and DAOM197198. (c) Allele frequency graphs of RNA-seq data at bi-allelic sites. (d) Mapping percentages of transcripts (grey) and bubbles (black) on reference assemblies. Sets of bubbles and transcripts are mapped on the appropriate assembly for each isolate. (e) Number of bi-allelic positions detected in RNA-seq data that have the same allele composition (black) or divergent allele composition (light grey) compared to WG data. Error bars represent ± 1 S.E.

Allele frequencies in C3 displayed a peak centred at 50:50, although the peak was not as clear as in some of the diploid or heterokaryons shown in Figure S12. There were two additional peaks at 10 and 90. C2 and DAOM197198 allele frequency plots did not exhibit a 50:50 peak (Figure 5c). U-shaped distributions were observed in C2 and DAOM197198. In addition, moderate peaks were observed at 10:90 and 90:10 in the three isolates. Such patterns are very similar to the patterns observed in the ddRAD-seq and WG data (Figure 4 and S10).

Testing for the existence of matching bi-alleles in the R. irregularis transcriptome and genome

We tested whether bi-alleles found in the transcriptome of the three isolates were also detected independently in the genome. Moderate peaks at 0.1 and 0.9 in the frequency

plots (Figure 5c) could be due to wrongly called SNPs because pairs of bubbles do not belong to single transcripts. We tested if all the bubble sequences produced by KisSplice mapped unambiguously to their corresponding reference genome. Both bubble and whole transcript sequences showed similar, and high, proportions of transcripts mapping a single time to the reference genome (Figure 5d). Second, these peaks could be due to sequencing errors. We tested if the positions with two alleles found in RNA-seq data also displayed the same two alleles in WG data. If the bi-alleles are located on two genetically-different nuclei then the same alleles should be found in RNA-seq data and WG data. A total of 326 identical bi-allelic positions occurred in both RNA-seq and WG data in the isolate C3 (Figure 5e). A very small number of bi-allelic positions found in RNA-seq data were also found in WG data in C2 and DAOM197198 (17 in C2 and 5 in DAOM197198). Most of the bi-allelic positions detected in RNA-seq data in C2 and DAOM197198 were not found in WG data (depicted by grey bars and labelled as “divergent allele composition”; Figure 5e). An example of inconsistent positions in C2 is provided in Figure S13. As a comparison, an example of consistent positions in C3 is shown in Figure S14. A notably larger number of bi-allelic positions occurred in RNA-seq data that were not found in WG data in isolate C3 compared to C2 and DAOM197198 (Figure 5e).

Impact of unsolved regions on the within-isolate polymorphism

Each position was visually inspected in the WG read alignment to the reference genome. Bi-allelic positions in C2 and DAOM197198 were frequently found around problematic regions in the genome assembly. Problematic regions were either unsolved regions (gaps) in the genome assembly or regions where the depth of coverage shows an unexpected increase compared to the mean depth of coverage. Both types of

problematic regions are frequently found in close locations. Unsolved regions are regions in which Ns are reported in the middle or at the extremities of the scaffold sequences. They correspond to genomic regions that are difficult to assemble. The unsolved regions, which are likely to occur in repeated regions, led to the reads to map wrongly and to accumulate at the wrong locations. This transfer of reads generated bi-allelic positions containing frequently two alleles with equal proportions. Examples of unsolved regions and their effect on the detected within-isolate polymorphism are provided in Figure S15.

Structural variation in the genome of the three isolates

Within isolate variation could also be potentially caused by structural variation existing among nuclei. Structural variation was found in all isolates. Both CNVnator and Lumpy enabled us to detect deletions in all isolates (Figure 6a and 6b). More deletions were detected in C3 than in C2 and DAOM197198 with both methods. The number of deletions detected was greater with CNVnator than with Lumpy, but the pattern was very similar. Most deletions were smaller than 5 kbp (Figure 6c). The largest sized deletions (ranging between 15 and 20 Kbp) were found in the genome of A4, but their number was low and could not be linked to the enrichment of a particular group of genes of similar function. Genes that occurred where a deletion had also occurred were identified and their number is shown in Figure 6d. Detailed results can be found at <https://github.com/FredMasc/PAAT2019>. Functional annotation and GO enrichment analysis of these genes did not reveal large differences among the isolates (Table S6). Notably, the GO category “DNA binding” was prevalent and many genes matched with transposases. Visual inspection of the alignments was conducted to characterize the deletions. Many deletions appeared to be associated with unsolved regions. The

assembly quality is an important parameter to accurately detect deletions using Illumina small read mapping. We looked for genes with functions unrelated to transposition. Regions where structural variants were found in all isolates with a strong drop in coverage were discarded because they likely reflect problems in genome assembly. We provide two examples of genes in A4/C3 that overlap deletions that did not occur in the other isolates (Figure 6e and 6f). The region around each gene was affected by a decrease of approximately 50% in coverage. In addition, SV at gene *g11881* (Figure 6e) was supported by 41 read pairs that had an abnormal insert size (too large) and SV at gene *g22991* (Figure 4f) was supported by 19 read pairs that had excessively large insert size. The gene *g11881* encodes a hypothetical protein (148 amino acids) with no obviously identifiable domain. The gene *g22991* encodes a serine/threonine protein kinase (487 amino acids).

Figure 6. Structural variation (SV) in WG data. (a) Number of deletions found with Lumpy. (b) Number of deletions found with CNVnator. (c) Deletion sizes among the isolates. (d) Number of genes overlapped by deletions. (e) and (f) Examples of structural variation found in A4 at the gene *g11881* (e) and at the gene *g22991* (f). The first track represents the gene model. The other tracks correspond to the depth of coverage measured in different isolates. The genomic region affected by an SV is labelled with a red box and red arrows.

542 Discussion

543 Here we demonstrated that studies of within and among-isolate SNP variation in *R.*
 544 *irregularis*, conducted independently, and using different techniques produced
 545 remarkably similar results. We have also made the first analysis linking within-AMF
 546 isolate variation at the genomic level to within-fungus transcription at bi-allelic sites
 547 indicating that in a heterokaryon *R. irregularis*, transcription of two alleles at bi-allelic
 548 sites is consistent with the proportion of the two genetically different nuclei but that in
 549 homokaryons little evidence exists for any expression of possible within fungus genetic
 550 variation. Finally, we have demonstrated the level of structural variation in the genome
 551 in 3 isolates that was due to deletions. The implications of these findings are discussed
 552 below.

553 *Within and among-isolate SNP variation in R. irregularis*

554 Our analyses indicated that SNP variation within isolates of *R. irregularis* were
 555 remarkably similar in two studies where the same fungal isolates (or clones of the
 556 isolates) were cultured and then sequenced independently in two different laboratories
 557 and using two different sequencing approaches. Not only were many of the same SNPs
 558 identified in the two studies but they were repeatedly observed in replicates within and
 559 between the studies (Figure 3). The same bi-allelic SNPs observed both among
 560 replicates and between independent studies means that the probability that they
 561 occurred in the sequence datasets by chance is exceptionally unlikely. Furthermore,
 562 allele frequency plots among samples from the two studies are also remarkably similar
 563 (Figure 4). This analysis shows that for detection of SNPs in *R. irregularis*, both
 564 approaches are equally valuable for SNP discovery in this fungal species, with the only
 565 differences being that more SNPs can be discovered with whole genome sequencing

but more isolates can be screened for the same cost using ddRAD-seq. In both cases, sufficient sequencing depth is required and the fact that deeper sequencing reveals more SNPs means that deep sequencing might lead to discovery of a greater number of rare alleles. However, this should also be coupled with replication to ensure rare alleles truly exist.

The discrepancies between the studies are that Ropars et al. (2016) reported lower within-fungus polymorphism than Wyss et al. (2016), although this differed greatly among isolates with values for some isolates being very close and others more distant [8, 9]. As shown in this study, one reason for this is that Wyss et al. (2016) mapped reads to the only available reference genome at that time; namely, an assembly of the isolate DAOM 197198. This is because of the very large amount of among-isolate genetic variation. Mapping to a genome assembly of the isolate rather than a reference genome greatly reduced the number of bi-allelic SNPs identified (Figure S3). This is because of the very large genetic differences among isolates in this species. This is an important result as looking at SNPs among individuals in population genetics studies is usually conducted by comparing SNPs found in the genome or transcriptome of a given individual to a reference genome. However, artefacts arising from mapping to one reference genome still does not account for the discovery of consistent among replicate or between-study within-fungus variation observed in each isolate. In the case of isolates considered heterokaryons, these likely represent differences between the two nucleus genotypes or some artefact of mapping where problematic parts of the genome assembly occur. In isolates considered homokaryons, bi-allelic SNPs could either represent artefacts due to problematic parts of the genome assembly or rare alleles.

Genome and transcriptome variation in AMF

Allele frequency plots of some isolates reported in Wyss et al. (2016) did not exhibit the U-shaped curves [8]. However, Ropars et al. (2016) observed U-shaped curves in homokaryons but did not explain what accounted for such a distribution [9]. In this study, we show that the shape of the allele frequency distribution is highly sensitive to the read threshold number used as a cut-off (Figure S11). At a low threshold of 10, almost all isolates appeared to have multiple bi-alleles with many different proportions; even those claimed to be homokaryons. It is difficult to set a standard threshold for all isolates at which homokaryons appear as homokaryons and heterokaryons reveal a 0.5 peak of allele distribution. A higher threshold sets a bias for where only high coverage variants are retained and these usually represent areas of poor assembly or repeated regions. Thus, this represents an artefact where even heterokaryons will exhibit a U-shaped curve of allele distribution making it difficult to detect heterokaryons. Given that setting a read threshold number is arbitrary, this makes distinction between homokaryons and heterokaryons difficult.

Bi-allelic variation in the R. irregularis genome and its transcription

While there has been considerable effort in documenting within and among isolate genome variation in AMF, little attention has been paid whether this variation is transcribed. The number of bi-allelic transcripts that were observed in the isolates expected to be homokaryons were indeed low and only a very small number matched those observed in the genome SNP data (Figure 5e). Of those, most mapped to problematic regions of the genome assembly indicating that these are probably artefacts. A notably large number of alleles in the heterokaryon C3 were observed as 2 alleles in the transcript data and in the genome. We assumed that bi-allelic transcripts that did not have corresponding alleles in the genome data were artefacts (light grey bars in Figure

Genome and transcriptome variation in AMF

5e). However, these were notably higher in C3 than in the other two isolates which would not be expected if they occurred because of sequencing errors. Some bi-allelic positions were not discovered in the genome of this isolate because the coverage of genome data is insufficient. The results show that indeed, genes of the two different nucleus genotypes present in the heterokaryon C3 are transcribed. About 50% of these bi-alleles in C3 contained non-synonymous substitutions meaning that this variation among nuclei could potentially affect the fungal phenotype. The allele frequency graphs of bi-allelic SNPs in the genome and those observed independently in the transcriptome both revealed an allele frequency of 0.5. This suggests that in isolate C3, the two nucleus genotypes are present in roughly equal proportions; a measurement that was also corroborated by a separate study [30] using the allele frequencies of a single marker and replicate measurements on the same isolate. The 0.5 allele frequency in the transcripts suggests that allele expression from a pair of given alleles located on the two nuclei are expressed in equal amounts from two differing parts of the two nucleus genomes. In several studies from our group, sibling cultures of AMF, initiated from single sibling spores for the parent C3 have been shown to exhibit quantitative traits such as hyphal growth and spore production, as well as their effects on plant growth, that were significantly different from the parent C3 [4, 5]. This is highly unusual for clonal offspring. Recently, the ratios of the two nucleus genotypes in some of these sibling offspring lines were shown to deviate significantly from the 50:50 ratio observed in the parent C3 [30]. If gene transcription from the two nucleus genotypes is proportional to the frequency of those genotypes (as shown in the present study) then we could expect that sibling clonal offspring of C3 could have unequal transcription of bi-alleles and this could account for the differences in quantitative traits observed among single spore clonally produced siblings of C3.

Structural variation in homokaryon and heterokaryon R. irregularis

The analysis of structural variation in the isolates revealed a higher level of structural variation in the dikaryon C3 than the other two isolates. If C2 and DAOM197198 are truly homokaryons then this structural variation would not be expected. Thus, we assume that this variation is also due to parts of the assembly where data is of poor quality. In C3, the higher level of structural variation should represent true structural differences occurring between the two different nucleus genotypes and such structural variation among the two nucleus genotypes could also have a regulatory effect influencing the fungal phenotype. These structural variations could either have arisen because of the fusion of two homokaryons carrying genetically different nuclei or by rearrangement and recombination among nuclei within C3, as reported previously [31].

Conclusions

Investigating whether within-AMF isolate genetic variation is extremely difficult. Independent studies using different techniques have identified the same variants and we show that this has obviously not occurred by chance. Transcription profiling is a good way to verify some of these bi-allelic positions in the genome. Transcription profiling did not reveal significant transcription of bi-allelic sites in a homokaryon isolate. However, transcription profiling reveals that indeed, within fungus genetic variation existing in heterokaryons is transcribed from both of the co-existing genomes and in equal proportions from both nucleus genotypes. Further studies are needed to determine the nature of the genetic differences between these two genotypes and how their expression influences the fungal phenotype and its effects on plants.

661 **Acknowledgements**

662 We thank Vincent Lacroix for his advice on KisSplice and the Genomic Technologies
663 Facility, University of Lausanne for sequencing. Bioinformatics computations were
664 performed at the Vital-IT (www.vital-it.ch) Centre for High Performance Computing,
665 Swiss Institute of Bioinformatics. This research was funded Swiss National Science
666 Foundation grant (number: 31003A_162549).

667

668 **Author Contribution**

669 Design of the research: FGM, TW, MP, IRS; material and sample preparation: TW, PR;
670 data analysis and interpretation: FGM, TW, MP, IRS; writing the manuscript: FGM,
671 IRS, TW.

672

673

674 **References**

- 675 1. van der Heijden MGA, Martin FM, Selosse M-A, Sanders IR. Mycorrhizal
676 ecology and evolution: the past, the present, and the future. *New Phytol.* 2015;205:
677 1406-1423. pmid: 25639293

- 678 2. Keymer A, Pimprakar P, Wewer V, Huber C, Brands M, Bucerius SL, et al.
679 Lipid transfer from plants to arbuscular mycorrhiza fungi. *eLife.* 2017;6: 1-33. pmid:
680 28726631

- 681 3. Sanders IR, Croll D. Arbuscular mycorrhiza: the challenge to understand the
682 genetics of the fungal partner. *Ann Rev Gen.* 2010;44: 271-292. pmid: 20822441

- 683 4. Ehinger MO, Croll D, Koch AM, Sanders IR. Significant genetic and
684 phenotypic changes arising from clonal growth of a single spore of an arbuscular
685 mycorrhizal fungus over multiple generations. *New Phytol.* 2012;196: 853-861. pmid:
686 22931497

- 687 5. Angelard C, Colard A, Niculita-Hirzel H, Croll D, Sanders IR. Segregation in a
688 mycorrhizal fungus alters rice growth and symbiosis-specific gene transcription. *Curr*
689 *Biol.* 2010;20: 1216-1221. pmid: 20541408

- 690 6. Sanders IR. Sex, plasticity, and biologically significant variation in one
691 Glomeromycotina species. *New Phytol.* 2018;220: 968-970. pmid: 29480929

- 692 7. Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, et al.
693 Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant
694 symbiosis. *Proc Natl Acad Sci USA.* 2013;110: 20117-20122. pmid: 24277808

- 695 8. Wyss T, Masclaux FG, Rosikiewicz P, Pagni M, Sanders IR. Population
696 genomics reveals that within-fungus polymorphism is common and maintained in
697 populations of the mycorrhizal fungus *Rhizophagus irregularis*. *ISME J.* 2016: 1-13.
698 pmid: 26953600

- 699 9. Ropars J, Toro KS, Noel J, Pelin A, Charron P, Farinelli L, et al. Evidence for
700 the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi. *Nature Microbiol.*
701 2016;1: 16033. pmid: 27572831

- 702 10. Savary R, Masclaux FG, Wyss T, Droh G, Cruz Corella J, Machado AP, et al.
703 A population genomics approach shows widespread geographical distribution of
704 cryptic genomic forms of the symbiotic fungus *Rhizophagus irregularis*. *ISME J.*
705 2018;12: 17-30. pmid: 29027999

- 706 11. Maeda T, Kobayashi Y, Kameoka H, Okuma N, Takeda N, Yamaguchi K, et al.
707 Evidence of non-tandemly repeated rDNAs and their intragenomic heterogeneity in
708 *Rhizophagus irregularis*. *Commun Biol.* 2018;1: 87. pmid: 30271968

- 709 12. Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, et al. Single
710 nucleus genome sequencing reveals high similarity among nuclei of an
711 endomycorrhizal fungus. PLoS Gen. 2014;10: e1004078. pmid: 24415955
- 712 13. Boon E, Zimmerman E, Lang BF, Hijri M. Intra-isolate genome variation in
713 arbuscular mycorrhizal fungi persists in the transcriptome. J Evol Biol. 2010;23: 1519-
714 1527. pmid: 20492090
- 715 14. Boon E, Halary S, Baptiste E, Hijri M. Studying genome heterogeneity within
716 the arbuscular mycorrhizal fungal cytoplasm. Genome Biol Evol. 2015;7: 505-521.
717 pmid: 25573960
- 718 15. Bruns TD, Corradi N, Redecker D, Taylor JW, Öpik M. Glomeromycotina:
719 what is a species and why should we care? New Phytol. 2018;220: 963-967.
- 720 16. Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, et al.
721 Evolutionary genetics of genome merger and doubling in plants. Ann Rev Gen.
722 2008;42: 443-461. pmid: 18983261
- 723 17. Bécard G, Fortin JA. Early events of vesicular-arbuscular mycorrhiza formation
724 on Ri T-DNA transformed roots. New Phytol. 1988;108: 211-218.
- 725 18. Rosikiewicz P, Bonvin J, Sanders IR. Cost-efficient production of *in vitro*
726 *Rhizophagus irregularis*. Mycorrhiza. 2017;27: 477-486. pmid: 28210812
- 727 19. Schmieder R, Lim YW, Rohwer F, Edwards R. TagCleaner: Identification and
728 removal of tag sequences from genomic and metagenomic datasets. BMC
729 Bioinformatics. 2010;11: 341. pmid: 20573248
- 730 20. Schmieder R, Edwards R. Quality control and preprocessing of metagenomic
731 datasets. Bioinformatics. 2011;27: 863-864. pmid: 21278185
- 732 21. Pearson WR, Lipman DJ. Improved tools for biological sequence comparison.
733 Proc Natl Acad Sci USA. 1988;85: 2444-2448. pmid: 3162770
- 734 22. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The
735 sequence alignment/map format and SAMtools. Bioinformatics. 2009;25: 2078-2079.
736 pmid: 19505943
- 737 23. Garrison E, Marth G. Haplotype-based variant detection from short-read
738 sequencing. 2012: 1-9. Preprint. Available from: arXiv:1207.3907 [q-bio.GN]. Cited 1
739 July 2019
- 740 24. Conway JR, Lex A, Gehlenborg N. UpSetR: an R package for the visualization
741 of intersecting sets and their properties. Bioinformatics. 2017;33: 2938-2940. pmid:
742 28645171

- 743 25. Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot MF, et al.
744 KisSplice: *De-novo* calling alternative splicing events from RNA-seq data. BMC
745 Bioinformatics. 2012;13: S5. pmid: 22537044
- 746 26. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, et al.
747 Full-length transcriptome assembly from RNA-Seq data without a reference genome.
748 Nature Biotechnology. 2011;29: 644-652. pmid: 21572440
- 749 27. Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A probabilistic
750 framework for structural variant discovery. Genome Biol. 2014;15: 1-19. pmid:
751 24970577
- 752 28. Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, et al.
753 SpeedSeq: Ultra-fast personal genome analysis and interpretation. Nature Methods.
754 2015;12: 966-968. pmid: 26258291
- 755 29. Abyzov A, Urban AE, Snyder M, Gerstein M. CNVnator: An approach to
756 discover, genotype, and characterize typical and atypical CNVs from family and
757 population genome sequencing. Genome Res. 2011;21: 974-984. pmid: 21324876
- 758 30. Masclaux FG, Wyss T, Mateus-Gonzalez ID, Aletti C, Sanders IR. Variation in
759 allele frequencies at the bg112 locus reveals unequal inheritance of nuclei in a
760 dikaryotic isolate of the fungus *Rhizophagus irregularis*. Mycorrhiza. 2018;28: 369-
761 377. pmid: 29675619
- 762 31. Chen EC, Mathieu S, Hoffrichter A, Sedzielewska-Toro K, Peart M, Pelin A, et
763 al. Single nucleus sequencing reveals evidence of inter-nucleus recombination in
764 arbuscular mycorrhizal fungi. eLife. 2018;7: 1-17. pmid: 30516133
765

Figure 1

Figure 2

a

b

Figure 3

Figure 4

Figure 5

Figure 6**a** **b** **c** **d** **e** **f**