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

Running header

Genome and transcriptome variation in AMF

Authors

Frédéric G. Masclaux*[1,2], Tania Wyss*[1], Marco Pagni[2], Pawel Rosikiewicz[1], Ian R. Sanders[1]

Affiliations

[1]Department of Ecology and Evolution, University of Lausanne, 1015 Lausanne, Switzerland

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

* These authors contributed equally to this work.

Corresponding author

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

E-mail: ian.sanders@unil.ch

Summary

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

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

Introduction

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

Over the last two decades there has been some debate about whether AMF are homokaryons (possessing genetically identical nuclei) or heterokaryons (possessing two or more genetically different nuclei). The ensuing, and lengthy, debate about whether or not AMF are heterokaryons has largely confounded two separate questions. 1. Are AMF heterokaryons or homokaryons, without considering the quantity of genetic differences among nuclei? 2. How much genetic variation exists within an AMF isolate that could potentially be distributed among nuclei? Resolving this issue of genetic variation in this fungus is biologically highly relevant because clonally produced single spore offspring of *R. irregularis* show high variation in their phenotype [4] and extremely high variation in their effects on plant growth [5].

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

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

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

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

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

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

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

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

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

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

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

161

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 ddH_2O, 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 (ddH_2O). 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).

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).

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

187 *Processing sequence data*

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

192 *In silico prediction of ddRAD-seq fragments*

193 To compare the same sites in ddRAD-seq and WG data, analysis was restricted to
194 ddRAD-seq fragments existing in WG data. I*n silico* digestion with *Eco*RI and *Mse*I
195 was performed on each genome assembly to define predicted ddRAD-seq fragments
196 longer than 49 bp and delimited by a *Eco*RI site and a *Mse*I site. Predicted fragments
197 were aligned to their respective genome with Novoalign v3.02.00 (Novocraft-
198 Technologies, 2014). Fragments that could not be re-aligned were removed. Re-aligned
199 predicted fragments define regions that were considered in subsequent analyses. We
200 refer to 'predicted ddRAD-seq fragment' as one of these fragments and to 'predicted
201 ddRAD-seq region' to designate the location where the fragment mapped (Figure S1).

202 *Identification of interspersed repeats and repeated elements*

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

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

214 Method M03 was used previously [8]. The *in silico* predicted ddRAD-seq fragments
215 were submitted to pairwise comparisons using ggsearch36; a global pairwise alignment
216 algorithm from the package fasta-36.3.5e [21]. This was done to identify globally
217 similar predicted fragments among the ddRAD-seq fragments. Fragments having a
218 match with another fragment were labelled 'repeated elements'. The M03 repeats
219 corresponded to interspersed repeats and M03 repeated elements. Method M03 is the
220 default method used in this study (Figure S1).

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

227 With a third method, M12, *in silico* predicted ddRAD-seq fragments were mapped to
228 the genome assembly with Novoalign V3.02.00 using parameters -r All 100 -R 400 -s0.
229 Fragments with more than one match in the assembly were labelled 'repeated elements'.
230 The M12 repeats were a combination of interspersed repeats and M12 repeated
231 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

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

259 *Bi-allelic site density and allele frequency*

260 In order to only consider variants in single copy genes, the variant file of each sample
261 was filtered to only retain positions with two alleles in non-repeated and coding regions.
262 The number of bi-allelic positions was divided by the number of positions from
263 predicted ddRAD-seq regions that were sequenced with a coverage greater or equal to
264 10.

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

270 *Depth of coverage and predicted ddRAD-seq regions*

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

278 *Identification of common bi-allelic positions*

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

283 *Discovery of bi-allelic SNPs in RNA-seq data*

284 KisSplice performs a local assembly of RNA-seq reads and detects variants directly
285 from a De Bruijn graph, independently of a reference genome [25]. The methodology
286 to analyse RNA-seq data is summarized in Figure S2. Because KisSplice reports a very
287 local context around SNPs, we generated reference full-length transcriptome
288 assemblies from RNA-seq data with Trinity [26]. Tritnity v2.3.2 was run with standard
289 parameters (except --SS_lib_type RF). Predicted ORFs were identified in the Trinity
290 assembled transcripts with TransDecoder (v5.0.2). Next, we identified full-length, or
291 nearly full-length, transcripts
292 (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Counting-Full-Length-Trinity-
293 Transcripts, accessed on July 2018). The assembled transcripts were compared to
294 known proteins in Swissprot UniProtKB proteins (www.uniprot.org, accessed on June
295 2017). The comparison was performed with BLASTX (2.6.0+) with the following
296 parameters: -evalue 1e-20 -max_target_seqs 1 -outfmt 6. The length of the top hit and
297 percentage of the hit length, included in the alignment to the Trinity transcript, were
298 added to the BLASTX result file with the provided script
299 analyze_blastPlus_topHit_coverage.pl. Next, a custom Perl script was applied on the
300 resulting file to keep only the best isoform of each gene and to remove genes with
301 multiple groups (following Trinity's specifications for 'group' and 'isoform').
302 KisSplice version 2.4.0 was run on RNA-seq data of each organism with the

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

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

321 *Analysis of bubbles and bi-allelic transcripts*

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

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

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

333 *Structural variation*

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

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

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

354 **Results**

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

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

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

360 The mean depth of coverage across the samples ranged from 22 to 33 and from 13 to
361 47 for ddRAD-seq and WG data, respectively (Figure S4). A greater number of
362 predicted ddRAD-seq regions in the five isolates were covered by reads from WG data
363 than by ddRAD-seq data (Figure S5). Poly-allelic positions were almost all bi-allelic.
364 Tri-allelic positions represented approximately 0.5% of the poly-allelic sites. Only one
365 tetra-allelic position was found (Table S5). Bi-allelic positions were found in all isolates
366 and both types of data (Figure 2a). The density of bi-allelic positions ranged from 0.66
367 to 1.79 bi-allelic positions/kb (Figure 2b). In general, the density of bi-allelic positions
368 was higher when the depth of coverage was higher (Figure 2b). The majority of bi-
369 allelic sites that were found in the ddRAD-seq datasets also occurred in the WG datasets
370 (Figure 3). In most cases, these occurred in all replicates of the two datasets or in several
371 replicates of the ddRAD-seq data and WG data, confirming that it is very unlikely that
372 they occurred because of random error (Figure 3). Hypergeometric tests showed that
373 the overlap of which fragments from ddRAD-seq and WG mapped to the in silico
374 predicted RAD fragments was highly unlikely to be due to chance (Cumulative
375 hypergeometric test (function *phyper*), $P < 0.00001$, in all isolates). Alternative
376 approaches to determining non-repeated regions yielded very similar results (Figures
377 S6 and S7) meaning that the detection of within fungus SNPs was not strongly sensitive

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

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

391 **Figure 3. Comparison of bi-allelic sites found in replicates of WG and**
392 **ddRAD-seq data in five *R. irregularis* isolates.** Number of replicate samples
393 from WG and ddRAD-seq data that contain the same bi-allelic sites in each
394 isolate. The numbers above the bars in the histograms represent the number of
395 bi-allelic sites shared among a given set of replicates from WG and ddRAD-seq.
396 Bullets connected by vertical lines below the histograms show which replicates
397 contained the same bi−allelic positions. The Up-set plots show numbers of
398 common positions among samples ranked from highest on the left of the graph
399 for the highest 15 combinations of samples. A very small number of bi-allelic
400 sites were found among other samples that are not shown but these numbers are
401 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.

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

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

434 *Bi-allelic positions within RNA transcripts*

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

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

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

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

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

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

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

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

492 *Impact of unsolved regions on the within-isolate polymorphism*

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

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

506 *Structural variation in the genome of the three isolates*

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

522 assembly quality is an important parameter to accurately detect deletions using Illumina
523 small read mapping. We looked for genes with functions unrelated to transposition.
524 Regions where structural variants were found in all isolates with a strong drop in
525 coverage were discarded because they likely reflect problems in genome assembly. We
526 provide two examples of genes in A4/C3 that overlap deletions that did not occur in the
527 other isolates (Figure 6e and 6f). The region around each gene was affected by a
528 decrease of approximately 50% in coverage. In addition, SV at gene *g11881* (Figure
529 6e) was supported by 41 read pairs that had an abnormal insert size (too large) and SV
530 at gene *g22991* (Figure 4f) was supported by 19 read pairs that had excessively large
531 insert size. The gene *g11881* encodes a hypothetical protein (148 amino acids) with no
532 obviously identifiable domain. The gene *g22991* encodes a serine/threonine protein
533 kinase (487 amino acids).

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

541

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

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

571 The discrepancies between the studies are that Ropars et al. (2016) reported lower
572 within-fungus polymorphism than Wyss et al. (2016), although this differed greatly
573 among isolates with values for some isolates being very close and others more distant
574 [8, 9]. As shown in this study, one reason for this is that Wyss et al. (2016) mapped
575 reads to the only available reference genome at that time; namely, an assembly of the
576 isolate DAOM 197198. This is because of the very large amount of among-isolate
577 genetic variation. Mapping to a genome assembly of the isolate rather than a reference
578 genome greatly reduced the number of bi-allelic SNPs identified (Figure S3). This is
579 because of the very large genetic differences among isolates in this species. This is an
580 important result as looking at SNPs among individuals in population genetics studies is
581 usually conducted by comparing SNPs found in the genome or transcriptome of a given
582 individual to a reference genome. However, artefacts arising from mapping to one
583 reference genome still does not account for the discovery of consistent among replicate
584 or between-study within-fungus variation observed in each isolate. In the case of
585 isolates considered heterokaryons, these likely represent differences between the two
586 nucleus genotypes or some artefact of mapping where problematic parts of the genome
587 assembly occur. In isolates considered homokaryons, bi-allelic SNPs could either
588 represent artefacts due to problematic parts of the genome assembly or rare alleles.

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

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

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

613 5e). However, these were notably higher in C3 than in the other two isolates which
614 would not be expected if they occurred because of sequencing errors. Some bi-allelic
615 positions were not discovered in the genome of this isolate because the coverage of
616 genome data is insufficient. The results show that indeed, genes of the two different
617 nucleus genotypes present in the heterokaryon C3 are transcribed. About 50% of these
618 bi-alleles in C3 contained non-synonymous substitutions meaning that this variation
619 among nuclei could potentially affect the fungal phenotype. The allele frequency graphs
620 of bi-allelic SNPs in the genome and those observed independently in the transcriptome
621 both revealed an allele frequency of 0.5. This suggests that in isolate C3, the two
622 nucleus genotypes are present in roughly equal proportions; a measurement that was
623 also corroborated by a separate study [30] using the allele frequencies of a single marker
624 and replicate measurements on the same isolate. The 0.5 allele frequency in the
625 transcripts suggests that allele expression from a pair of given alleles located on the two
626 nuclei are expressed in equal amounts from two differing parts of the two nucleus
627 genomes. In several studies from our group, sibling cultures of AMF, initiated from
628 single sibling spores for the parent C3 have been shown to exhibit quantitative traits
629 such as hyphal growth and spore production, as well as their effects on plant growth,
630 that were significantly different from the parent C3 [4, 5]. This is highly unusual for
631 clonal offspring. Recently, the ratios of the two nucleus genotypes in some of these
632 sibling offspring lines were shown to deviate significantly from the 50:50 ratio
633 observed in the parent C3 [30]. If gene transcription from the two nucleus genotypes is
634 proportional to the frequency of those genotypes (as shown in the present study) then
635 we could expect that sibling clonal offspring of C3 could have unequal transcription of
636 bi-alleles and this could account for the differences in quantitative traits observed
637 among single spore clonally produced siblings of C3.

638 *Structural variation in homokaryon and heterokaryon* R. irregularis

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

649 *Conclusions*

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

660

Acknowledgements

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

Author Contribution

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

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, Pimprikar 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, Bapteste 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 i*n 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

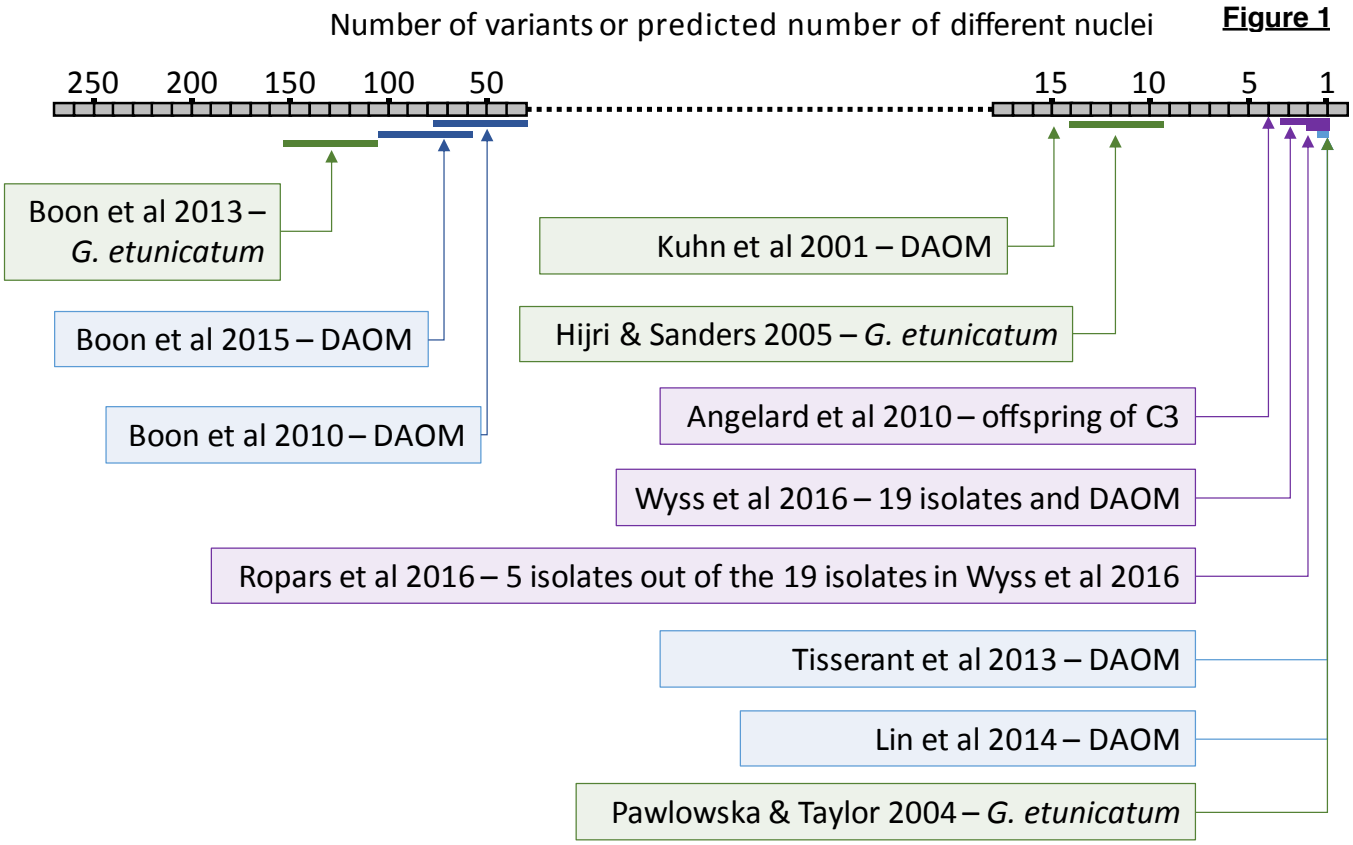
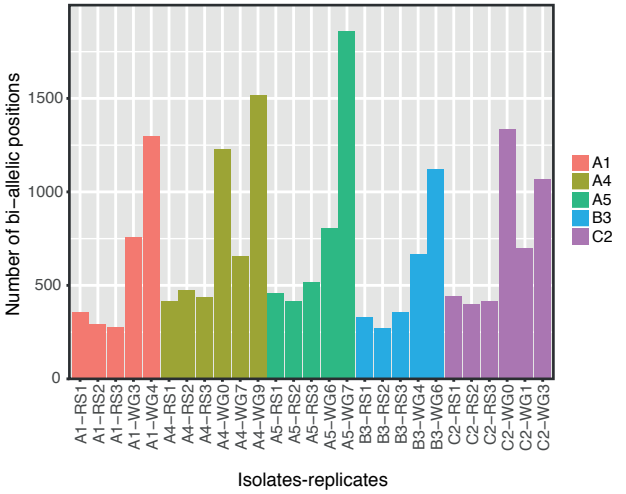

Number of variants or predicted number of different nuclei
250
200
150
100
50
15
10
5
1
Boon et al 2013 – G. etunicatum
Boon et al 2015 – DAOM
Boon et al 2010 – DAOM
Kuhn et al 2001 – DAOM
Hijri & Sanders 2005 – G. etunicatum
Angelard et al 2010 – offspring of C3
Wyss et al 2016 – 19 isolates and DAOM
Ropars et al 2016 – 5 isolates out of the 19 isolates in Wyss et al 2016
Tisserant et al 2013 – DAOM
Lin et al 2014 – DAOM
Pawlowska & Taylor 2004 – G. etunicatum

Figure 2

a


Number of bi−allelic positions
0
500
1000
1500
A1−RS1
A1−RS2
A1−RS3
A1−WG3
A1−WG4
A4−RS1
A4−RS2
A4−RS3
A4−WG0
A4−WG7
A4−WG9
A5−RS1
A5−RS2
A5−RS3
A5−WG6
A5−WG7
B3−RS1
B3−RS2
B3−RS3
B3−WG4
B3−WG6
C2−RS1
C2−RS2
C2−RS3
C2−WG0
C2−WG1
C2−WG3
Isolates-replicates
A1
A4
A5
B3
C2

b

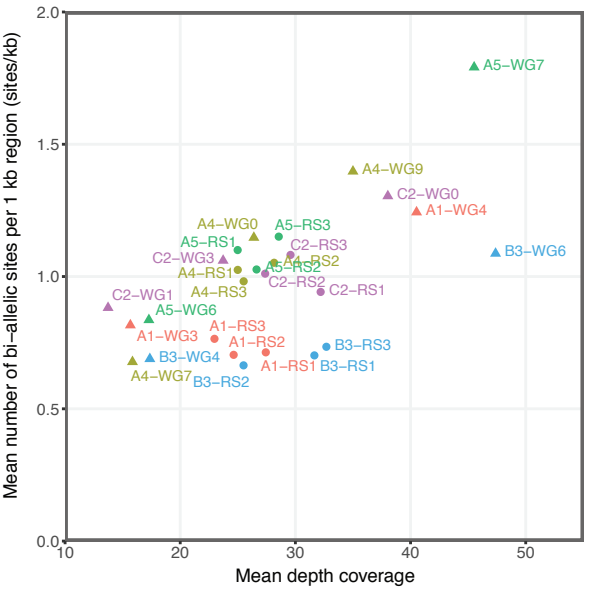
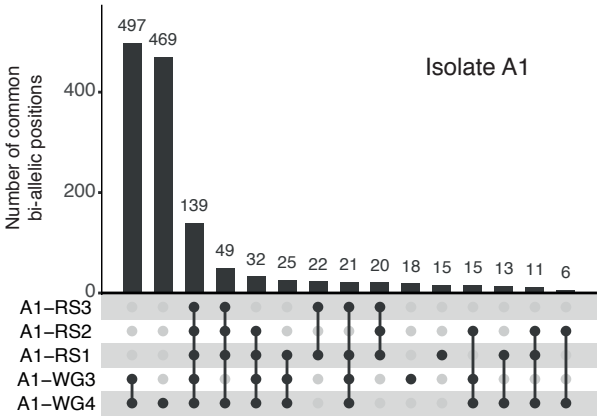
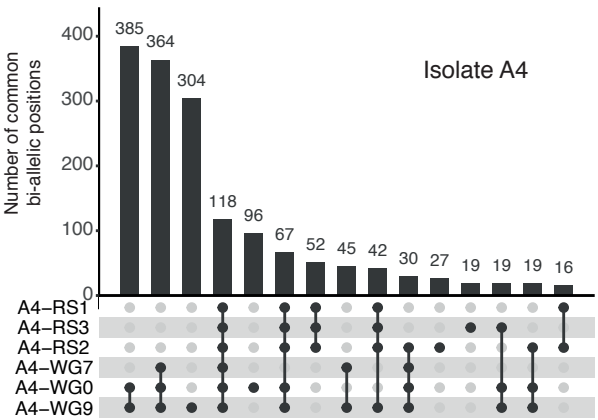
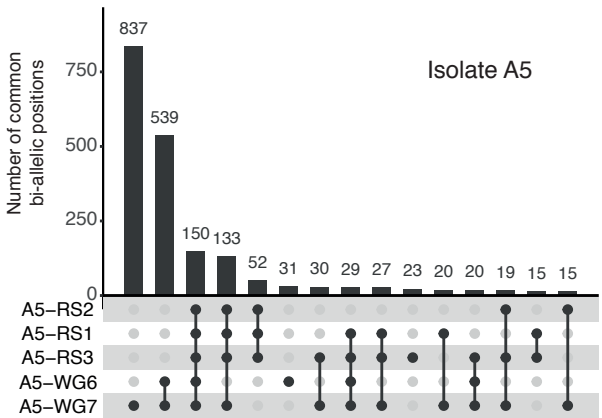
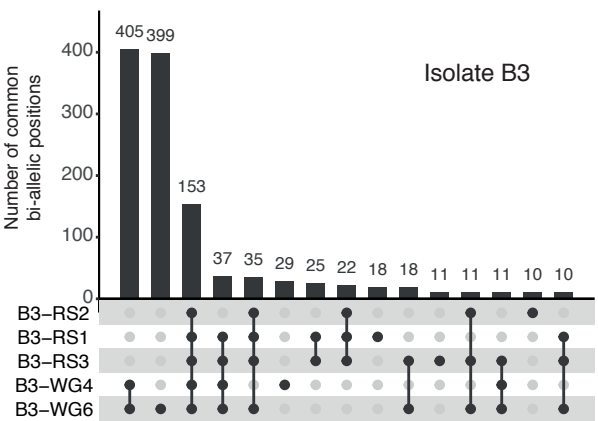

Mean number of bi−allelic sites per 1 kb region (sites/kb)
0.0
0.5
1.0
1.5
2.0
10
20
30
40
50
Mean depth coverage
A5−WG7
A4−WG9
C2−WG0
A1−WG4
B3−WG6
A4−WG0
A5−RS3
A5−RS1
C2−RS3
C2−WG3
A4−RS2
A4−RS1
A5−RS2
C2−RS2
C2−RS1
A4−RS3
C2−WG1
A5−WG6
A1−RS3
A1−WG3
A1−RS2
B3−RS3
B3−WG4
A1−RS1
B3−RS1
A4−WG7
B3−RS2

Figure 3


Isolate A1
Number of common bi-allelic positions
0
200
400
497
469
139
49
32
25
22
21
20
18
15
15
13
11
6
A1−RS3
A1−RS2
A1−RS1
A1−WG3
A1−WG4


Isolate A4
Number of common bi-allelic positions
0
100
200
300
400
385
364
304
118
96
67
52
45
42
30
27
19
19
19
16
A4−RS1
A4−RS3
A4−RS2
A4−WG7
A4−WG0
A4−WG9


Isolate A5
Number of common bi-allelic positions
0
250
500
750
837
539
150
133
52
31
30
29
27
23
20
20
19
15
15
A5−RS2
A5−RS1
A5−RS3
A5−WG6
A5−WG7


Isolate B3
Number of common bi-allelic positions
0
100
200
300
400
405
399
153
37
35
29
25
22
18
18
11
11
11
10
10
B3−RS2
B3−RS1
B3−RS3
B3−WG4
B3−WG6

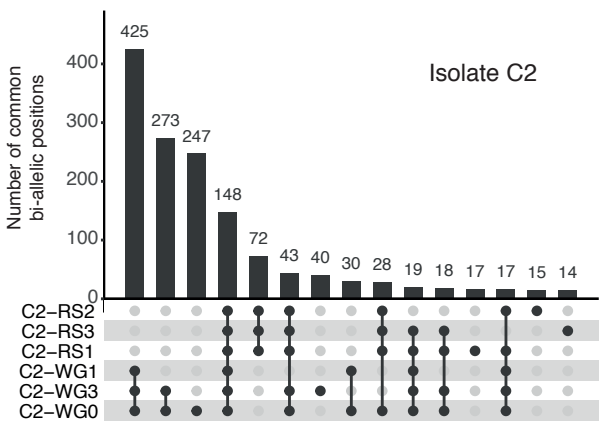

Isolate C2
Number of common bi-allelic positions
0
100
200
300
400
425
273
247
148
72
43
40
30
28
19
18
17
17
15
14
C2−RS2
C2−RS3
C2−RS1
C2−WG1
C2−WG3
C2−WG0

Figure 4

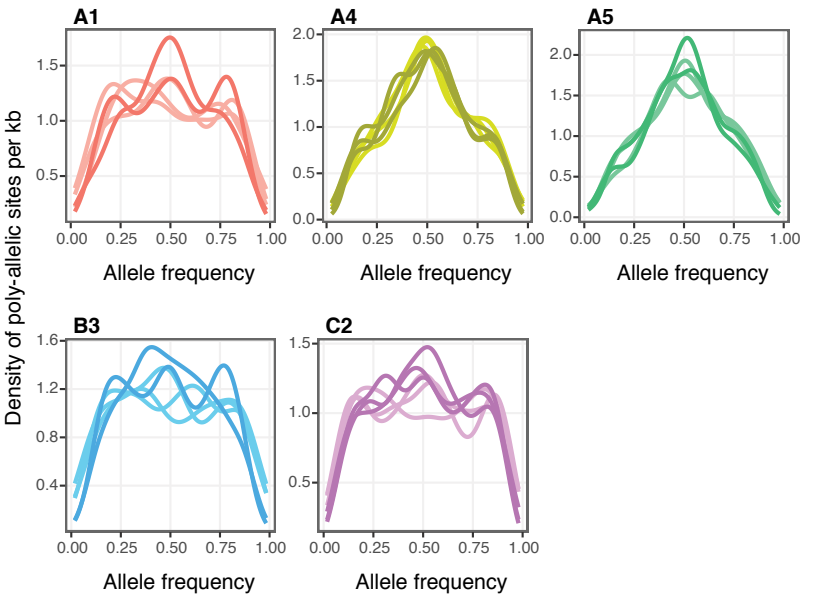

A1
A4
A5
B3
C2
Density of poly-allelic sites per kb
Allele frequency
0.00
0.25
0.50
0.75
1.00

Figure 5

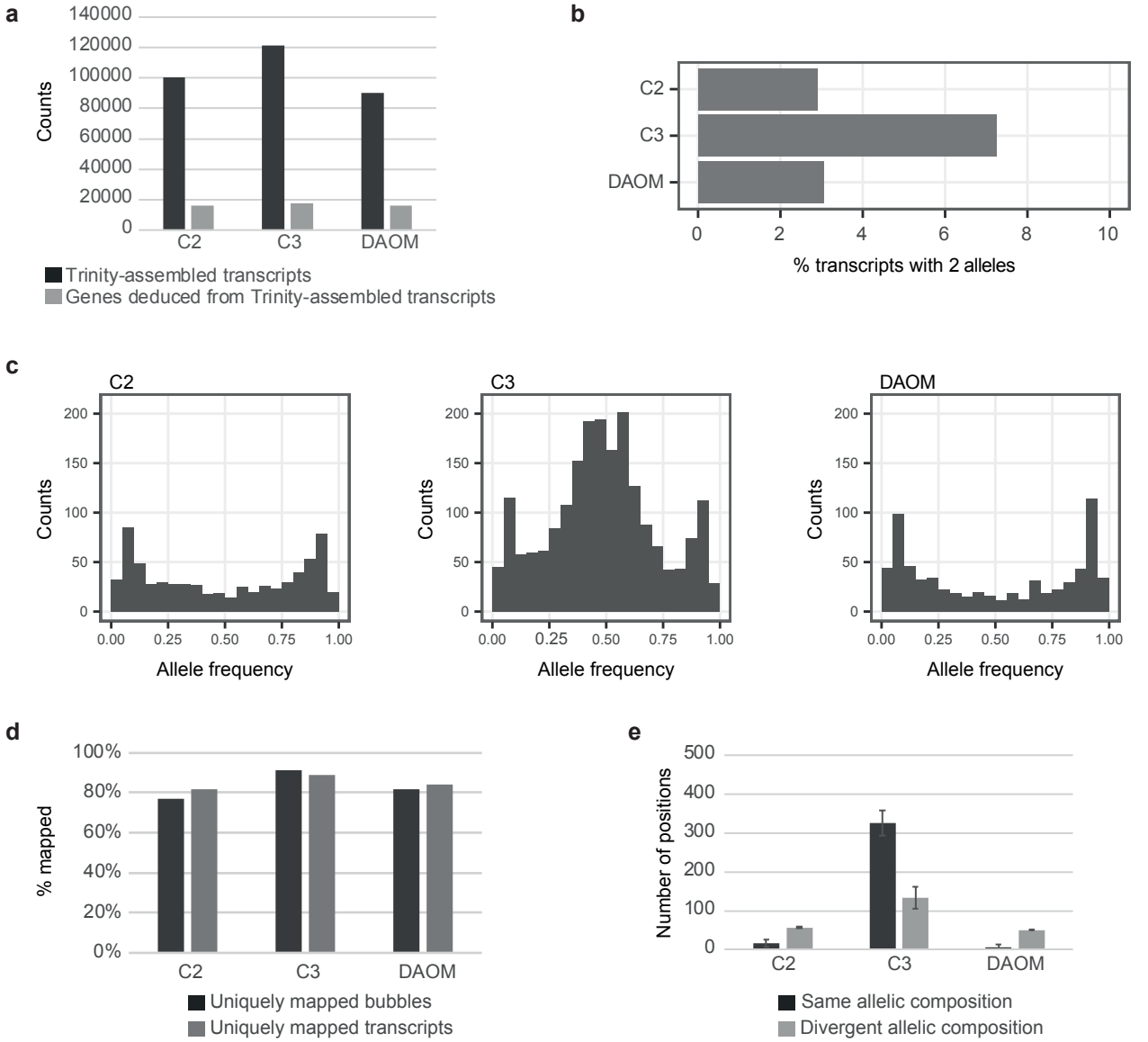
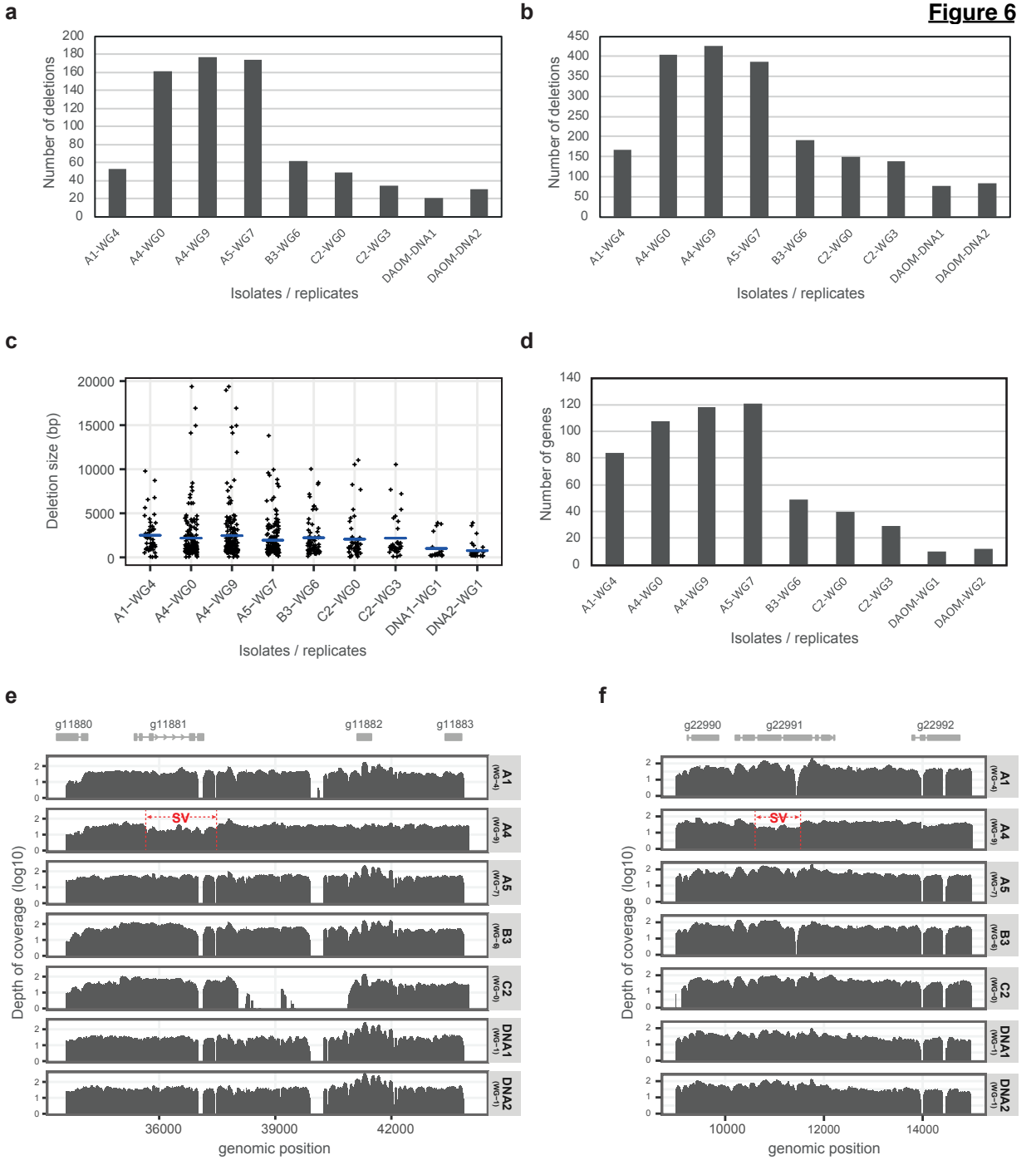

a
Counts
140000
120000
100000
80000
60000
40000
20000
0
C2
C3
DAOM
Trinity-assembled transcripts
Genes deduced from Trinity-assembled transcripts
b
C2
C3
DAOM
0
2
4
6
8
10
% transcripts with 2 alleles
c
C2
C3
DAOM
Counts
200
150
100
50
0
0.00
0.25
0.50
0.75
1.00
Allele frequency
d
% mapped
100%
80%
60%
40%
20%
0%
C2
C3
DAOM
Uniquely mapped bubbles
Uniquely mapped transcripts
e
Number of positions
500
400
300
200
100
0
C2
C3
DAOM
Same allelic composition
Divergent allelic composition

Figure 6


a
Number of deletions
Isolates / replicates
A1-WG4
A4-WG0
A4-WG9
A5-WG7
B3-WG6
C2-WG0
C2-WG3
DAOM-DNA1
DAOM-DNA2
b
Number of deletions
Isolates / replicates
c
Deletion size (bp)
Isolates / replicates
A1–WG4
A4–WG0
A4–WG9
A5–WG7
B3–WG6
C2–WG0
C2–WG3
DNA1–WG1
DNA2–WG1
d
Number of genes
Isolates / replicates
DAOM-WG1
DAOM-WG2
e
g11880
g11881
g11882
g11883
SV
Depth of coverage (log10)
genomic position
f
g22990
g22991
g22992
SV
Depth of coverage (log10)
genomic position