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 match the proposed monokaryon or dikaryon state.
- We re-analysed published data showing that bi-allelic sites (and their frequencies), detected in monosporic *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 exists.
- Bi-allelic transcripts from three *R. irregularis* isolates matched those observed in the genome. In putative monokaryon isolates, very few bi-allelic transcripts could be found in the genome. In a putative dikaryon, a large number of bi-allelic transcripts matched those in the genome. Bi-allelic transcripts also occurred in the same frequency in the putative dikaryon as predicted from allele frequency in the genome.
- Our results indicate that while within-fungus genome variation in putative monokaryon and dikaryon AMF was highly similar in 2 independent studies, there was little support that this variation is transcribed in monokaryons. In contrast, within-fungus variation thought to be segregated among nuclei in a dikaryon isolate is indeed transcribed in a way that is proportional to that seen in the genome.

Keywords: monokaryon, dikaryon, 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 (van der Heijden *et al.*, 2015; Keymer *et al.*, 2017). 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 (Sanders & Croll, 2010).

Over the last two decades there has been some debate about whether AMF are monokaryons (possessing genetically identical nuclei) or heterokaryons (possessing two or more genetically different nuclei). This debate has largely arisen because it is also unknown whether these fungi are sexual or not. Arbuscular mycorrhizal fungi reproduce by producing asexual or clonal spores on the termini of hyphae. Clonality is clearly a common mode of reproduction in these fungi as identical genotypes of the AMF *Rhizophagus irregularis* can be found in very distant geographic locations (Savary *et al.*, 2018). No direct evidence for sexual processes in AMF has been documented yet. Indirect evidence for possible sex exists (Croll & Sanders, 2009; Halary *et al.*, 2011; Riley & Corradi, 2013; Ropars *et al.*, 2016; Chen *et al.*, 2018). Kuhn et al. hypothesized that AMF nuclei accumulate mutations resulting in AMF containing divergent nuclei (Kuhn *et al.*, 2001). They used the null model of lack of sex in these fungi because of the absence of any evidence of sexual reproduction in AMF at the time of their publication. The fungi would, therefore, be heterokaryons. The ensuing, and lengthy, debate about whether or not AMF are heterokaryons has largely confounded two separate questions. 1. Are AMF heterokaryons or monokaryons,

71 without considering the quantity of genetic differences among nuclei? 2. How much
72 genetic variation exists within an AMF isolate that could potentially be distributed
73 among nuclei? Resolving this issue of genetic variation in this fungus is biologically
74 highly relevant because clonally produced single spore offspring of *R. irregularis* show
75 high variation in their phenotype (Ehinger *et al.*, 2012) and extremely high variation in
76 their effects on plant growth (Angelard *et al.*, 2010).

77 Studies over the last two decades that have tried to address the two questions outlined
78 above are highly inconsistent in their conclusions regarding the amount of genetic
79 variation existing within an AMF isolate and the prediction of the number of genetically
80 different nuclei (summarized in Figure 1). However, there has been a lack of studies
81 focused on the same species or isolate. Some species studied probably diverged from
82 each other hundreds of millions of years ago, making comparisons between some of
83 these studies as largely meaningless (Sanders, 2018). Furthermore, different approaches
84 have been used as technologies have developed over the years.

85 More recently, studies using different techniques and sequencing platforms have
86 concentrated on describing the genome of one haploid species of AMF, *R. irregularis,*
87 and its within and among isolate variation (Tisserant *et al.*, 2013; Wyss *et al.*, 2016;
88 Ropars *et al.*, 2016; Savary *et al.*, 2018; Maeda *et al.*, 2018). Thus, there are now
89 directly comparable data originating from independent studies. The whole genome of
90 the *R. irregularis* isolate, DAOM197198, was sequenced showing that within fungus
91 polymorphism existed but was low. The authors concluded that this isolate was a
92 monokaryon (Tisserant *et al.*, 2013; Lin *et al.*, 2014) and did not try and explain the
93 variation. Double-digest restriction site-associated sequencing (ddRAD-seq) was
94 performed on DAOM197198 and 19 other *R. irregularis* isolates (Wyss *et al.*, 2016),

showing different levels of within-fungus polymorphism among the isolates that suggested isolates with a predominantly monokaryon or dikaryon state but 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 monokaryons and that two isolates; A4 and A5 were dikaryons. Notably, none of these studies have recorded high levels of polymorphism as those observed by Boon et al. (Boon *et al.*, 2010, 2015) (Figure 1). Despite this, Bruns et al. (2018) suggest a strong dichotomy between, on the one hand, the studies of Boon et al. (2010 and 2015) and Wyss et al. (2016) and on the other hand, the studies by Ropars et al. (2016), Lin et al. (2014) and Tisserant et al. (2014) (see discussion by Sanders 2018).

All studies revealed a level of polymorphism unexplainable by a strict monokaryon or dikaryon state, even if they concluded the fungi were monokaryons or dikaryons. In each case, bi-allelic, and even tri- and tetra-allelic, sites existed in isolates that had been predicted to be monokaryons. However, this variation was ignored. Plots of allele frequency at bi-allelic sites revealed a peak at 0.5 in several isolates but with additional variation at other frequencies (Wyss *et al.*, 2016). A clear 0.5 peak of frequency at bi-allelic sites is expected for a dikaryon with equal proportions of nuclei 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 (Wyss *et al.*, 2016).

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. Ropars et al. (2016) also constructed replicate libraries

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

127 So far there has been almost no attention paid to the functional significance of any
128 within-fungus genome variation observed in AMF. However, single spore offspring of
129 clonally produced *R. irregularis* shows a high degree of variation in their phenotype
130 and their effects on plant growth (Angelard *et al.*, 2010; Ehinger *et al.*, 2012).
131 Transcriptome profiling of *R. irregularis* would reveal whether the genome variation
132 observed at bi-allelic sites is transcribed or whether unequal or single-allele expression
133 occurs. However, as yet, no studies exist in AMF relating variation at the transcriptome
134 level back to variation observed at the genome level. The detection of bi-allelic
135 transcripts that match observed variation in the genome would confirm the existence of
136 bi-alleles in the genome and would indicate that the variation could potentially have
137 consequences on the phenotype. Expression of two alleles of a given gene, unequal
138 expression of the two alleles, or expression of only one of them, could all have
139 consequences on the phenotype of the fungus or its effect on plant growth. If alleles
140 were truly located on two genetically different nuclei, it would mean that it would also
141 represent a measure of the contribution by the two different AMF nucleus genotypes to
142 overall gene expression in the fungus. In plants, the contribution of two different
143 genomes (for example in hybrids or polyploids) can be highly unequal and the

contribution from each genome can be tissue specific (Doyle *et al.*, 2008). 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 dikaryon) and the isolate C2 (that has been predicted to be a monokaryon) 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.

161 **Materials and Methods**

162 *Fungal material, RNA extraction and library preparation*

163 Three *R. irregularis* isolates (C2, C3 and DAOM197198) were cultured *in vitro* with
164 *Ri* T-DNA transformed carrot roots (Bécard & Fortin, 1988) and a split plate system
165 (Rosikiewicz *et al.*, 2017). Fungal material was extracted from the medium in citrate
166 buffer (450 mL ddH_2O, 8.5 mL 0.1 N citric acid, 41.5 mL 0.1 N Na citrate) for 20 min
167 and washed with sterile double-deionized water (ddH_2O). Material from three Petri
168 plates was pooled per sample. There were three samples per isolate. RNA was extracted
169 with the RNeasy Plant Mini kit (Qiagen). Two micrograms of RNA from each sample
170 was used to prepare libraries for RNA sequencing (RNA-seq) using the TruSeq
171 Stranded Total RNA Library Prep Kit (Illumina) with a PCR enrichment step of 15
172 cycles. The libraries were pooled and sequenced using Illumina HiSeq 2500 (100 bp
173 paired-end reads). Sequences were deposited in the NCBI SRA database (BioProject
174 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 monokaryon 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 monokaryon and dikaryon 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
189 (Schmieder *et al.*, 2010). Reads were processed with PrinSeq-lite.pl version 0.20.4
190 (Schmieder & Edwards, 2011) to remove reads containing 'Ns', to trim low quality 3'-
191 ends and retain reads longer than 50 bp. A summary of the read size after filtering is in
192 Table S4.

193 *In silico prediction of ddRAD-seq fragments*

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

203 *Identification of interspersed repeats and repeated elements*

204 We use the term 'repeats' to include the interspersed repeats (mobile elements and
205 transposon-like regions) and repeated elements (multicopy genes and duplicated
206 regions). The interspersed repeat families were predicted *de novo* with the program

207 RepeatModeler Open-1.0.8 (Smit & Hubley, 2008) and interspersed repeats were
208 annotated with RepeatMasker Open-4.0.6 (Smit *et al.*, 1996).

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

215 Method M03 was previously used (Wyss *et al.*, 2016). The *in silico* predicted ddRAD-
216 seq fragments were submitted to pairwise comparisons using ggsearch36; a global
217 pairwise alignment algorithm from the package fasta-36.3.5e (Pearson & Lipman,
218 1988). This was done to identify globally similar predicted fragments among the
219 ddRAD-seq fragments. Fragments having a match with another fragment were labelled
220 'repeated elements'. The M03 repeats corresponded to interspersed repeats and M03
221 repeated elements. Method M03 is the default method used in this study (Figure S1).

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

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

231 The M12 repeats were a combination of interspersed repeats and M12 repeated
232 elements. This method is of medium stringency.

233 *Gene prediction in genome assemblies*

234 Coding regions were defined with the *ab initio* gene predictor Augustus version 3.1
235 using the gene model previously defined for the N6 assembly, constructed for the
236 DAOM197198 isolate (Wyss *et al.*, 2016).

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

238 Wyss et al. (2016) mapped reads to the DAOM197198 genome assembly because other
239 genome assemblies of *R. irregularis* isolates were not available at the time. For this
240 study, we used genome assemblies of isolates A1, A4, A5, B3 and C2 for mapping. We
241 aligned reads against the genome assemblies with Novoalign v3.02.12 (Novocraft-
242 Technologies, 2014). The correctly mapped paired-end reads were selected for
243 subsequent analyses using Samtools version 1.3 (Li *et al.*, 2009). Single nucleotide
244 polymorphisms (SNPs), insertions and deletions (indels) and multiple nucleotide
245 polymorphisms (MNPs) were called for each sample using Freebayes version 1.0.2
246 (Garrison & Marth, 2012) at positions with a minimum 10× depth of coverage. Only
247 alleles with frequencies greater or equal to 0.1 were recorded. Variant files were filtered
248 to keep the positions with a phred-scaled quality score greater or equal to 30. For every
249 sample, and at each site, we reported whether there was a single allele, two alleles or
250 more than two alleles. A missing value (NA) was assigned if the depth of coverage was
251 below 10×. Because the objective of the study was to study all the within-isolate
252 polymorphism, no filter for the allelic ratio was applied to a position to call it bi-allelic
253 or poly-allelic position.

254 We compared variants in ddRADseq data (Wyss et al. 2016) and WG data (Ropars et
255 al. 2016). If the same variants are found among replicates and between two independent
256 studies, then it is highly unlikely that the observed variants are artefacts caused by error
257 generated in the sequencing. Wyss et al. (2016) included 3 replicates per isolate. Ropars
258 et al. (2016) assembled genomes of *R. irregularis* that were pooled samples of different
259 libraries that can also be treated as replicates. The numbers of replicates in WG
260 sequencing was between 2 and 3, depending on the isolate.

261 *Bi-allelic site density and allele frequency*

262 In order to only consider variants in single copy genes, the variant file of each sample
263 was filtered to only retain positions with two alleles in non-repeated and coding regions.
264 The number of bi-allelic positions was divided by the number of positions from
265 predicted ddRAD-seq regions that were sequenced with a coverage greater or equal to
266 10.

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

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

273 Low coverage in samples would limit the ability to reliably detect polymorphisms.
274 Samtools version 1.3 (Li *et al.*, 2009) was used to calculate depth of coverage in
275 ddRAD-seq and WG data. Samtools and custom perl script were used to calculate the
276 mean depth of coverage per predicted ddRAD-seq region. A mean coverage greater, or

277 equal to, 10 was necessary to consider a predicted ddRAD-seq region as covered. A
278 similar method was used to calculate the depth of coverage per position required to
279 calculate bi-allelic position density.

280 *Identification of common bi-allelic positions*

281 The first goal of this study was to compare whether the same bi-allelic sites were found
282 among replicates and studies. The R package UpSetR (Conway *et al.*, 2017) was
283 applied on lists of bi-allelic positions in non-repeated, coding regions to find common
284 bi-allelic positions in different replicates and generate UpSet plots.

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

286 KisSplice performs a local assembly of RNA-seq reads and detects variants directly
287 from a De Bruijn graph, independently of a reference genome (Sacomoto *et al.*, 2012).
288 The methodology to analyse RNA-seq data is summarized in Figure S2. Because
289 KisSplice reports a very local context around SNPs, we generated reference full-length
290 transcriptome assemblies from RNA-seq data with Trinity (Grabherr *et al.*, 2011).
291 Tritnity v2.3.2 was run with standard parameters (except --SS_lib_type RF). Predicted
292 ORFs were identified in the Trinity assembled transcripts with TransDecoder (v5.0.2).
293 Next, we identified full-length, or nearly full-length, transcripts
294 (https://github.com/trinityrnaseq/trinityrnaseq/wiki/Counting-Full-Length-Trinity-
295 Transcripts, accessed on July 2018). The assembled transcripts were compared to
296 known proteins in Swissprot UniProtKB proteins (www.uniprot.org, accessed on June
297 2017). The comparison was performed with BLASTX (2.6.0+) with the following
298 parameters: -evalue 1e-20 -max_target_seqs 1 -outfmt 6. The length of the top hit and
299 percentage of the hit length, included in the alignment to the Trinity transcript, were
300 added to the BLASTX result file with the provided script

301 analyze_blastPlus_topHit_coverage.pl. Next, a custom Perl script was applied on the
302 resulting file to keep only the best isoform of each gene and to remove genes with
303 multiple groups (following Trinity's specifications for 'group' and 'isoform').
304 KisSplice version 2.4.0 was run on RNA-seq data of each organism with the
305 parameters: --experimental -s 2 -k 41. We focused on the 'Type0a bubbles' (KisSplice
306 nomenclature), that are two identical short sequences differing by a single SNP at a
307 given position. Bubble sequences were positioned on the Trinity-reference
308 transcriptome using BLAT with the option -minIdentity=80. To predict amino acid
309 changes of a SNP, the program KisSplice2RefTranscriptome (K2RT) was run taking as
310 input the predicted ORFs, the Type0a bubbles produced by KisSplice, and the mapping
311 results of BLAT. A custom Perl script removed bubbles that matched two sequences in
312 the reference transcriptome and those with mapping ambiguity. The script also removed
313 bubbles with a read coverage less than six and bubbles with one of the two alleles
314 having less than two observations. Bubbles matching a transcript without a BLASTX
315 hit were removed. These filtering steps ensured that sets of SNPs had the highest
316 possible quality. The KissSplice/K2RT pipeline reports counts of both alleles at bi-
317 allelic positions. Using these values, the distribution of allele frequencies at bi-allelic
318 positions was plotted for each organism.

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

323 *Analysis of bubbles and bi-allelic transcripts*

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

327 To investigate whether bi-allelic positions discovered in RNA-seq data matched to bi-
328 allelic sites in the genome, bubble sequences containing 2 alleles were mapped to the
329 reference genome to provide the genomic coordinates of the bi-allelic positions. Next,
330 allelic composition of each bi-allelic position in bubbles was compared to the allele
331 composition of the position in the WG data. Any difference in allele composition was
332 recorded as "inconsistent". An "inconsistent" position is a bi-allelic position detected
333 in RNA-seq that was not bi-allelic in WG data or that had a different allele composition
334 to the WG data.

335 *Structural variation*

336 WG samples (A1-WG4, A4-WG0, A4-WG9, A5-WG7, B3-WG6, C2-WG0, C2-WG3,
337 DAOM-DNA1 and DAOM-DNA2), with a mean coverage greater than 20 were
338 analysed to detect structural variations (SV). The calling of SV was performed using
339 Lumpy version v2.16.2 (Layer *et al.*, 2014). This integrates multiple SV signals: read-
340 depth variation events, paired-end events and split-read events. A new read mapping
341 was performed with SpeedSeq version 0.1.2 (Chiang *et al.*, 2015). All read sets were
342 mapped to every genome to generate all possible combinations. The resulting .bam files
343 were processed with Lumpy to call variants. Results were filtered to keep scaffolds with
344 a size greater or equal to 10000, SV events that were supported by at least nine total
345 events, three paired-end events and three split-read events. Only deletions detected with
346 Lumpy were considered in this study.

347 SV calling was also performed using CNVnator (Abyzov *et al.*, 2011). The .bam files
348 generated with novoalign to call the variants were reused to call the SV with CNVnator
349 version 0.3.3. Results were filtered using the fthresholds: scaffold size greater or equal
350 to 10000, p-value 1 (p-val1) < 0.05, q0 value < 0.5. Only deletions detected with
351 CNVnator were considered in this study.

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

355

356 **Results**

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

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

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

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

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

384 *Allele frequency similarities*

385 Assuming nuclei are haploid, allele frequency distribution at bi-allelic loci should
386 reveal the number and proportion of different nuclei. Very similar allele frequency
387 distributions were observed among the replicates of each isolate originating from WG
388 and ddRAD-seq data (Figure 4; Figure S8). A peak at 50:50 was detected in A4 and A5
389 (Figure 2b), as expected for a 50:50 dikaryon and as shown previously (Ropars *et al.*,
390 2016). There was a slight peak at 50:50, as well as slight peaks at approximately 20:80
391 and 80:20, in A1, B3 and C2 (Figure 4). It is also noticeable that the small peaks at
392 20:80 and 80:20 also existed in the isolates A4 and A5 despite a clear peak at 50:50.
393 The isolate DAOM197198 was also subjected to the same analysis. These results are
394 presented in the Figure S9.

395 *Effects of coverage and other parameters*

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

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

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

411 *Bi-allelic positions within RNA transcripts*

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

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

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

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

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

439 We tested whether bi-alleles found in the transcriptome of the three isolates were also
440 detected independently in the genome. Moderate peaks at 0.1 and 0.9 in the frequency
441 plots (Figure 5c) could be due to wrongly called SNPs because pairs of bubbles do not
442 belong to single transcripts. We tested if all the bubble sequences produced by
443 KisSplice mapped unambiguously to their corresponding reference genome. Both
444 bubble and whole transcript sequences showed similar, and high, proportions of
445 transcripts mapping a single time to the reference genome (Figure 5d). Second, these
446 peaks could be due to sequencing errors. We tested if the positions with two alleles
447 found in RNA-seq data also displayed the same two alleles in WG data. If the bi-alleles
448 are located on two genetically-different nuclei then the same alleles should be found in
449 RNA-seq data and WG data. A total of 326 identical bi-allelic positions occurred in
450 both RNA-seq and WG data in the isolate C3 (Figure 5e). A very small number of bi-

451 allelic positions found in RNA-seq data were also found in WG data in C2 and
452 DAOM197198 (17 in C2 and 5 in DAOM197198). Most of the bi-allelic positions
453 detected in RNA-seq data in C2 and DAOM197198 were not found in WG data
454 (depicted by grey bars and labelled as "divergent allele composition"; Figure 5e). An
455 example of inconsistent positions in C2 is provided in Figure S13. As a comparison, an
456 example of consistent positions in C3 is shown in Figure S14. A notably larger number
457 of bi-allelic positions occurred in RNA-seq data that were not found in WG data in
458 isolate C3 compared to C2 and DAOM197198 (Figure 5e).

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

460 Each position was visually inspected in the WG read alignment to the reference genome.
461 Bi-allelic positions in C2 and DAOM197198 were frequently found around problematic
462 regions in the genome assembly. Problematic regions were either unsolved regions
463 (gaps) in the genome assembly or regions where the depth of coverage shows an
464 unexpected increase compared to the mean depth of coverage. Both types of
465 problematic regions are frequently found in close locations. Unsolved regions are
466 regions in which Ns are reported in the middle or at the extremities of the scaffold
467 sequences. They correspond to genomic regions that are difficult to assemble. The
468 unsolved regions, which are likely to occur in repeated regions, led to the reads to map
469 wrongly and to accumulate at the wrong locations. This transfer of reads generated bi-
470 allelic positions containing frequently two alleles with equal proportions. Examples of
471 unsolved regions and their effect on the detected within-isolate polymorphism are
472 provided in Figure S15.

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

474 Within isolate variation could also be potentially caused by structural variation existing
475 among nuclei. Structural variation was found in all isolates. Both CNVnator and Lumpy
476 enabled us to detect deletions in all isolates (Figure 6a and 6b). More deletions were
477 detected in C3 than in C2 and DAOM197198 with both methods. The number of
478 deletions detected was greater with CNVnator than with Lumpy, but the pattern was
479 very similar. Most deletions were smaller than 5 kbp (Figure 6c). The largest sized
480 deletions (ranging between 15 and 20 Kbp) were found in the genome of A4, but their
481 number was low and could not be linked to the enrichment of a particular group of
482 genes of similar function. Genes that occurred where a deletion had also occurred were
483 identified and their number is shown in Figure 6d. Detailed results can be found at
484 https://github.com/FredMasc/PAAT2019. Functional annotation and GO enrichment
485 analysis of these genes did not reveal large differences among the isolates (Table S6).
486 Notably, the GO category "DNA binding" was prevalent and many genes matched with
487 transposases. Visual inspection of the alignments was conducted to characterize the
488 deletions. Many deletions appeared to be associated with unsolved regions. The
489 assembly quality is an important parameter to accurately detect deletions using Illumina
490 small read mapping. We looked for genes with functions unrelated to transposition.
491 Regions where structural variants were found in all isolates with a strong drop in
492 coverage were discarded because they likely reflect problems in genome assembly. We
493 provide two examples of genes in A4/C3 that overlap deletions that did not occur in the
494 other isolates (Figure 6e and 6f). The region around each gene was affected by a
495 decrease of approximately 50% in coverage. In addition, SV at gene *g11881* (Figure
496 6e) was supported by 41 read pairs that had an abnormal insert size (too large) and SV
497 at gene *g22991* (Figure 4f) was supported by 19 read pairs that had excessively large
498 insert size. The gene *g11881* encodes a hypothetical protein (148 amino acids) with no

499 obviously identifiable domain. The gene *g22991* encodes a serine/threonine protein
500 kinase (487 amino acids).

501

502 **Discussion**

503 Here we demonstrated that studies of within and among-isolate SNP variation in *R.*
504 *irregularis*, conducted independently, and using different techniques produced
505 remarkably similar results. We have also made the first analysis linking within-AMF
506 isolate variation at the genomic level to within-fungus transcription at bi-allelic sites
507 indicating that in a dikaryon *R. irregularis*, transcription of two alleles at bi-allelic sites
508 is consistent with the proportion of the two genetically different nuclei but that in
509 monokaryons little evidence exists for any expression of possible within fungus genetic
510 variation. Finally, we have demonstrated the level of structural variation in the genome
511 in 3 isolates that was due to deletions. The implications of these findings are discussed
512 below.

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

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

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

531 The discrepancies between the studies are that Ropars et al. (2016) reported lower
532 within-fungus polymorphism than Wyss et al. (2016) (see Corradi & Brachmann (2017)
533 for comparison), although this differed greatly among isolates with values for some
534 isolates being very close and others more distant. As shown in this study, one reason
535 for this is that Wyss et al. (2016) mapped reads to the only available reference genome
536 at that time; namely, an assembly of the isolate DAOM 197198. This is because of the
537 very large amount of among-isolate genetic variation. Mapping to a genome assembly
538 of the isolate rather than a reference genome greatly reduced the number of bi-allelic
539 SNPs identified (Figure S3). This is because of the very large genetic differences among
540 isolates in this species. This is an important result as looking at SNPs among individuals
541 in population genetics studies is usually conducted by comparing SNPs found in the
542 genome or transcriptome of a given individual to a reference genome. However,
543 artefacts arising from mapping to one reference genome still does not account for the
544 discovery of consistent among replicate or between-study within-fungus variation
545 observed in each isolate. In the case of isolates considered dikaryons, these likely
546 represent differences between the two nucleus genotypes or some artefact of mapping
547 where problematic parts of the genome assembly occur. In isolates considered
548 monokaryotes, bi-allelic SNPs could either represent artefacts due to problematic parts
549 of the genome assembly or rare alleles.

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

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

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

574 5e). However, these were notably higher in C3 than in the other two isolates which
575 would not be expected if they occurred because of sequencing errors. Some bi-allelic
576 positions were not discovered in the genome of this isolate because the coverage of
577 genome data is insufficient. The results show that indeed, genes of the two different
578 nucleus genotypes present in the dikaryon C3 are transcribed. About 50% of these bi-
579 alleles in C3 contained non-synonymous substitutions meaning that this variation
580 among nuclei could potentially affect the fungal phenotype. The allele frequency graphs
581 of bi-allelic SNPs in the genome and those observed independently in the transcriptome
582 both revealed an allele frequency of 0.5. This suggests that in isolate C3, the two
583 nucleus genotypes are present in roughly equal proportions; a measurement that was
584 also corroborated by a separate study (Masclaux *et al.*, 2018) using the allele
585 frequencies of a single marker and replicate measurements on the same isolate. The 0.5
586 allele frequency in the transcripts suggests that allele expression from a pair of given
587 alleles located on the two nuclei are expressed in equal amounts from two differing
588 parts of the two nucleus genomes. In several studies from our group, sibling cultures of
589 AMF, initiated from single sibling spores for the parent C3 have been shown to exhibit
590 quantitative traits such as hyphal growth and spore production, as well as their effects
591 on plant growth, that were significantly different from the parent C3 (Angelard *et al.*,
592 2010; Ehinger *et al.*, 2012). This is highly unusual for clonal offspring. Recently,
593 Masclaux et al. (2018) showed that the ratios of the two nucleus genotypes in some of
594 these sibling offspring lines deviated significantly from the 50:50 ratio observed in the
595 parent C3. If gene transcription from the two nucleus genotypes is proportional to the
596 frequency of those genotypes (as shown in the present study) then we could expect that
597 sibling clonal offspring of C3 could have unequal transcription of bi-alleles and this

598 could account for the differences in quantitative traits observed among single spore
599 clonally produced siblings of C3.

600 *Structural variation in monokaryon and dikaryon* R. irregularis

601 The analysis of structural variation in the isolates revealed a higher level of structural
602 variation in the dikaryon C3 than the other two isolates. If C2 and DAOM197198 are
603 truly monokaryons then this structural variation would not be expected. Thus, we
604 assume that this variation is also due to parts of the assembly where data is of poor
605 quality. In C3, the higher level of structural variation should represent true structural
606 differences occurring between the two different nucleus genotypes and such structural
607 variation among the two nucleus genotypes could also have a regulatory effect
608 influencing the fungal phenotype. These structural variations could either have arisen
609 because of the fusion of two monokarons carrying genetically different nuclei or by
610 rearrangement and recombination among nuclei within C3, as reported by Chen et al.
611 (2018).

612 *Conclusions*

613 Investigating whether within-AMF isolate genetic variation is extremely difficult.
614 Independent studies using different techniques have identified the same variants and
615 we show that this has obviously not occurred by chance. Transcription profiling is a
616 good way to verify some of these bi-allelic positions in the genome. Transcription
617 profiling did not reveal significant transcription of bi-allelic sites in a monokaryon
618 isolate. However, transcription profiling reveals that indeed, within fungus genetic
619 variation existing in dikaryons is transcribed from both of the co-existing genomes and
620 in equal proportions from both nucleus genotypes. Further studies are needed to

621 determine the nature of the genetic differences between these two genotypes and how
622 their expression influences the fungal phenotype and its effects on plants.

623

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.

636 **References**

637 **Abyzov A, Urban AE, Snyder M, Gerstein M**. **2011**. CNVnator: An approach to
638 discover, genotype, and characterize typical and atypical CNVs from family and
639 population genome sequencing. *Genome Research* **21**: 974–984.

640 **Angelard C, Colard A, Niculita-Hirzel H, Croll D, Sanders IR**. **2010**. Segregation
641 in a mycorrhizal fungus alters rice growth and symbiosis-specific gene transcription.
642 *Current Biology* **20**: 1216–1221.

643 **Bécard G, Fortin JA**. **1988**. Early events of vesicular-arbuscular mycorrhiza formation
644 on *Ri* T-DNA transformed roots. *New Phytologist* **108**: 211–218.

645 **Boon E, Halary S, Bapteste E, Hijri M**. **2015**. Studying genome heterogeneity within
646 the arbuscular mycorrhizal fungal cytoplasm. *Genome Biology and Evolution* **7**: 505–
647 521.

648 **Boon E, Zimmerman E, Lang BF, Hijri M**. **2010**. Intra-isolate genome variation in
649 arbuscular mycorrhizal fungi persists in the transcriptome. *Journal of evolutionary*
650 *biology* **23**: 1519–27.

651 **Bruns TD, Corradi N, Redecker D, Taylor JW, Öpik M**. **2018**. Glomeromycotina:
652 what is a species and why should we care? *New Phytologist* **220**: 963–967.

653 **Chen EC, Mathieu S, Hoffrichter A, Sedzielewska-Toro K, Peart M, Pelin A,**
654 **Ndikumana S, Ropars J, Dreissig S, Fuchs J, *et al.* 2018**. Single nucleus sequencing
655 reveals evidence of inter-nucleus recombination in arbuscular mycorrhizal fungi. *eLife*
656 **7**: 1–17.

657 **Chiang C, Layer RM, Faust GG, Lindberg MR, Rose DB, Garrison EP, Marth**
658 **GT, Quinlan AR, Hall IM**. **2015**. SpeedSeq: Ultra-fast personal genome analysis and
659 interpretation. *Nature Methods* **12**: 966–968.

660 **Conway JR, Lex A, Gehlenborg N**. **2017**. UpSetR: an R package for the visualization

661 of intersecting sets and their properties. *Bioinformatics (Oxford, England)* **33**: 2938–
662 2940.
663 **Corradi N, Brachmann A**. **2017**. Fungal mating in the most widespread plant
664 symbionts? *Trends in Plant Science* **22**: 175–183.
665 **Croll D, Sanders IR**. **2009**. Recombination in *Glomus intraradices*, a supposed ancient
666 asexual arbuscular mycorrhizal fungus. *BMC evolutionary biology* **9**: 13.
667 **Doyle JJ, Flagel LE, Paterson AH, Rapp RA, Soltis DE, Soltis PS, Wendel JF**. **2008**.
668 Evolutionary genetics of genome merger and doubling in plants. *Annual review of*
669 *genetics* **42**: 443–61.
670 **Ehinger MO, Croll D, Koch AM, Sanders IR**. **2012**. Significant genetic and
671 phenotypic changes arising from clonal growth of a single spore of an arbuscular
672 mycorrhizal fungus over multiple generations. *The New phytologist* **196**: 853–61.
673 **Garrison E, Marth G**. **2012**. Haplotype-based variant detection from short-read
674 sequencing. *arXiv preprint arXiv:1207.3907*: 9.
675 **Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis**
676 **X, Fan L, Raychowdhury R, Zeng Q, *et al.* 2011**. Full-length transcriptome assembly
677 from RNA-Seq data without a reference genome. *Nature biotechnology* **29**: 644–52.
678 **Halary S, Malik S-B, Lildhar L, Slamovits CH, Hijri M, Corradi N**. **2011**.
679 Conserved meiotic machinery in *Glomus spp.*, a putatively ancient asexual fungal
680 lineage. *Genome biology and evolution* **3**: 950–8.
681 **van der Heijden MGA, Martin FM, Selosse M-A, Sanders IR, Heijden MGA Van**
682 **Der, Martin FM, Sanders IR**. **2015**. Mycorrhizal ecology and evolution: the past, the
683 present, and the future. *New Phytologist* **205**: n/a-n/a.
684 **Keymer A, Pimprikar P, Wewer V, Huber C, Brands M, Bucerius SL, Delaux PM,**
685 **Klingl V, von Röpenack-Lahaye E, Wang TL, *et al.* 2017**. Lipid transfer from plants

686 to arbuscular mycorrhiza fungi. *eLife* **6**: 1–33.

687 **Kuhn G, Hijri M, Sanders IR**. **2001**. Evidence for the evolution of multiple genomes
688 in arbuscular mycorrhizal fungi. *Nature* **414**: 745–8.

689 **Layer RM, Chiang C, Quinlan AR, Hall IM**. **2014**. LUMPY: a probabilistic
690 framework for structural variant discovery. *Genome Biology* **15**: R84.

691 **Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis**
692 **G, Durbin R**. **2009**. The Sequence Alignment/Map format and SAMtools.
693 *Bioinformatics (Oxford, England)* **25**: 2078–9.

694 **Lin K, Limpens E, Zhang Z, Ivanov S, Saunders DGO, Mu D, Pang E, Cao H, Cha**
695 **H, Lin T, *et al.* 2014**. Single nucleus genome sequencing reveals high similarity among
696 nuclei of an endomycorrhizal fungus. *PLoS genetics* **10**: e1004078.

697 **Maeda T, Kobayashi Y, Kameoka H, Okuma N, Takeda N, Yamaguchi K, Bino T,**
698 **Shigenobu S, Kawaguchi M**. **2018**. Evidence of non-tandemly repeated rDNAs and
699 their intragenomic heterogeneity in *Rhizophagus irregularis*. *Communications Biology*
700 **1**: 87.

701 **Masclaux FG, Wyss T, Mateus-Gonzalez ID, Aletti C, Sanders IR**. **2018**. Variation
702 in allele frequencies at the *bg112* locus reveals unequal inheritance of nuclei in a
703 dikaryotic isolate of the fungus *Rhizophagus irregularis*. *Mycorrhiza*.

704 **Novocraft-Technologies**. **2014**. Novoalign.

705 **Pearson WR, Lipman DJ**. **1988**. Improved tools for biological sequence comparison.
706 *Proceedings of the National Academy of Sciences of the United States of America* **85**:
707 2444–8.

708 **Riley R, Corradi N**. **2013**. Searching for clues of sexual reproduction in the genomes
709 of arbuscular mycorrhizal fungi. *Fungal Ecology* **6**: 44–49.

710 **Ropars J, Toro KS, Noel J, Pelin A, Charron P, Farinelli L, Marton T, Krüger M,**

711 **Fuchs J, Brachmann A, *et al.* 2016**. Evidence for the sexual origin of heterokaryosis
712 in arbuscular mycorrhizal fungi. *Nature Microbiology* **1**: 16033.
713 **Rosikiewicz P, Bonvin J, Sanders IR**. **2017**. Cost-efficient production of in vitro
714 *Rhizophagus irregularis*. *Mycorrhiza* **27**: 477–486.
715 **Sacomoto GAT, Kielbassa J, Chikhi R, Uricaru R, Antoniou P, Sagot M-F,**
716 **Peterlongo P, Lacroix V**. **2012**. KISSPLICE: de-novo calling alternative splicing
717 events from RNA-seq data. *BMC bioinformatics* **13**: S5.
718 **Sanders IR**. **2018**. Sex, plasticity, and biologically significant variation in one
719 Glomeromycotina species: A response to Bruns et al. (2017) 'Glomeromycotina: What
720 is a species and why should we care?' *New Phytologist*: 1–3.
721 **Sanders IR, Croll D**. **2010**. Arbuscular mycorrhiza: the challenge to understand the
722 genetics of the fungal partner. *Annual review of genetics* **44**: 271–92.
723 **Savary R, Masclaux FG, Wyss T, Droh G, Cruz Corella J, Machado AP, Morton**
724 **JB, Sanders IR**. **2018**. A population genomics approach shows widespread
725 geographical distribution of cryptic genomic forms of the symbiotic fungus
726 *Rhizophagus irregularis*. *The ISME journal* **12**: 17–30.
727 **Schmieder R, Edwards R**. **2011**. Quality control and preprocessing of metagenomic
728 datasets. *Bioinformatics (Oxford, England)* **27**: 863–4.
729 **Schmieder R, Lim YW, Rohwer F, Edwards R**. **2010**. TagCleaner: Identification and
730 removal of tag sequences from genomic and metagenomic datasets. *BMC*
731 *bioinformatics* **11**: 341.
732 **Smit AFA, Hubley R**. **2008**. RepeatModeler Open-1.0.
733 **Smit AFA, Hubley R, Green P**. **1996**. RepeatMasker Open-3.0.
734 **Tisserant E, Malbreil M, Kuo A, Kohler A, Symeonidi A, Balestrini R, Charron P,**
735 **Duensing N, Frei dit Frey N, Gianinazzi-Pearson V, *et al.* 2013**. Genome of an

arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proceedings of the National Academy of Sciences of the United States of America* **110**: 20117–22.

Wyss T, Masclaux FG, Rosikiewicz P, Pagni M, Sanders IR. **2016**. Population genomics reveals that within-fungus polymorphism is common and maintained in populations of the mycorrhizal fungus *Rhizophagus irregularis*. *The ISME journal*: 1–13.

746 **Figure captions**

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

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

766 **Figure 3. Comparison of bi-allelic sites found in replicates of WG and ddRAD-seq**
767 **data in five *R. irregularis* isolates.** Number of replicate samples from WG and
768 ddRAD-seq data that contain the same bi-allelic sites in each isolate. The numbers
769 above the bars in the histograms represent the number of bi-allelic sites shared among

770 a given set of replicates from WG and ddRAD-seq. Bullets connected by vertical lines
771 below the histograms show which replicates contained the same bi−allelic positions.

772

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

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

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

Figure 1

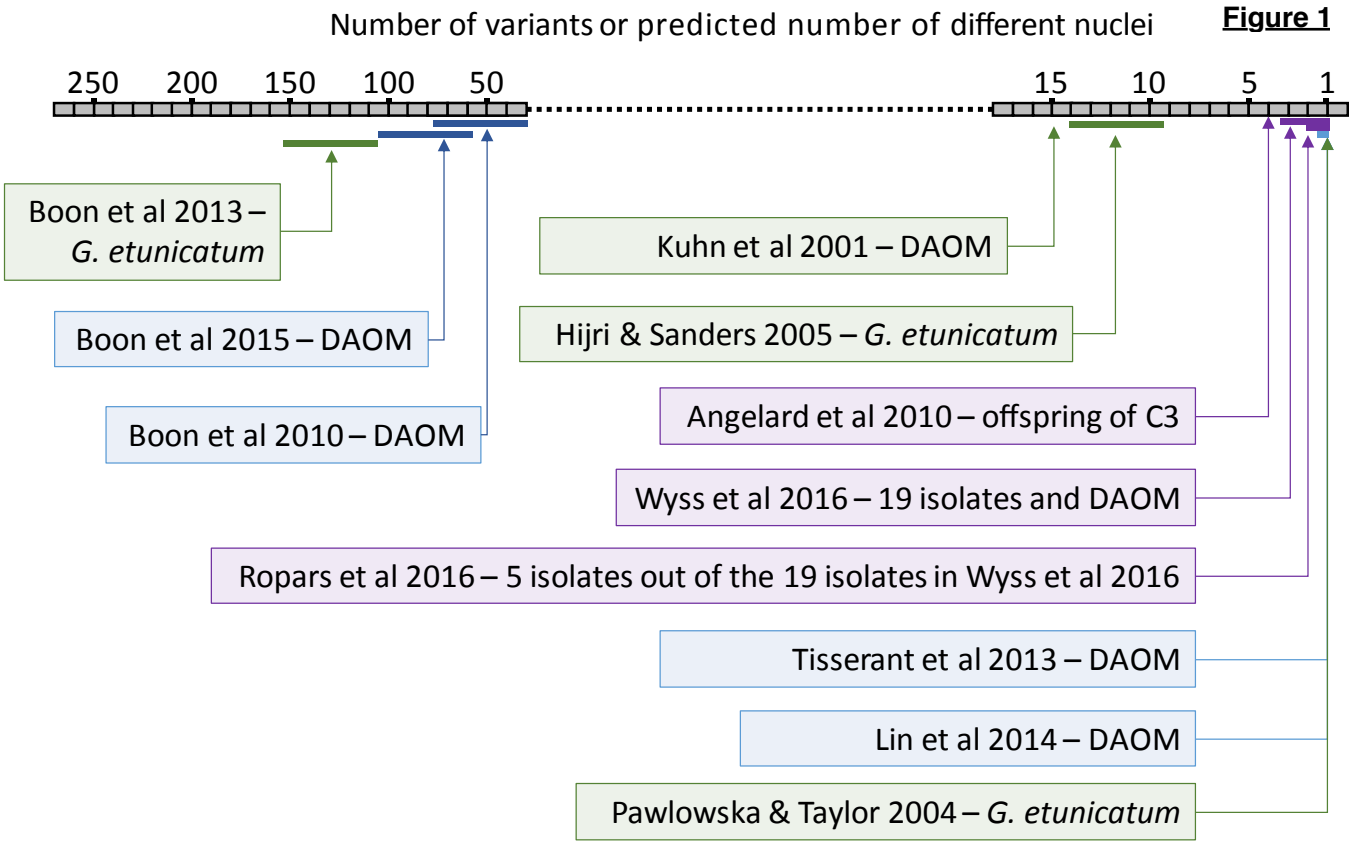
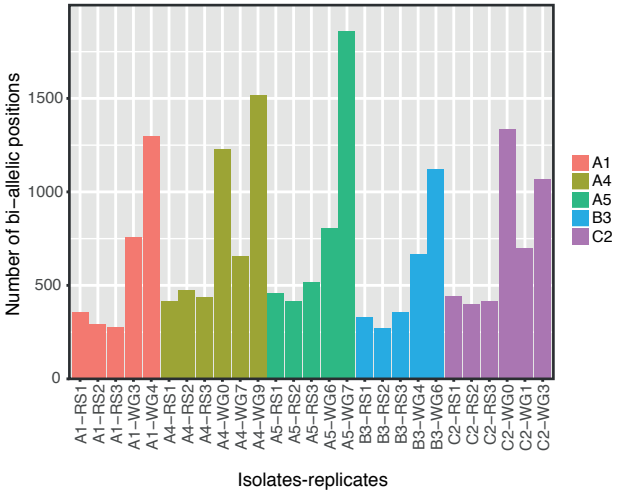

Number of variants or predicted number of different nuclei
250
200
150
100
50
15
10
5
1
Boon et al 2013 – G. etunicatum
Boon et al 2015 – DAOM
Boon et al 2010 – DAOM
Kuhn et al 2001 – DAOM
Hijri & Sanders 2005 – G. etunicatum
Angelard et al 2010 – offspring of C3
Wyss et al 2016 – 19 isolates and DAOM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

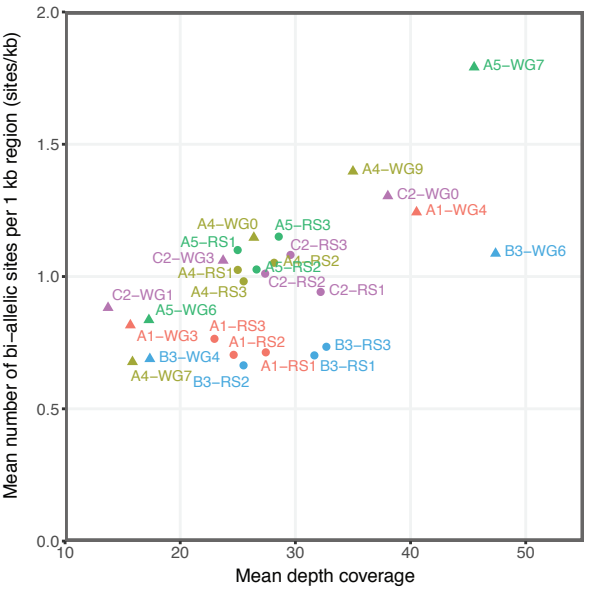
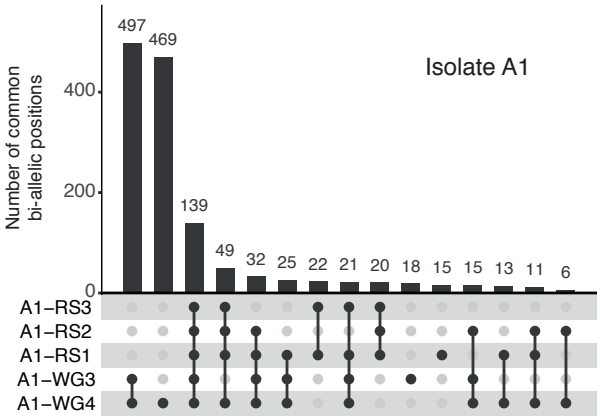
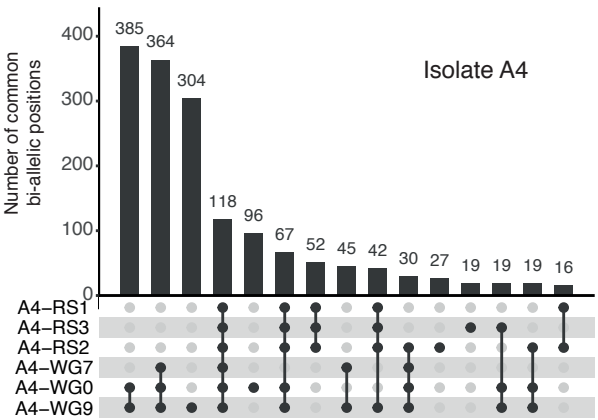
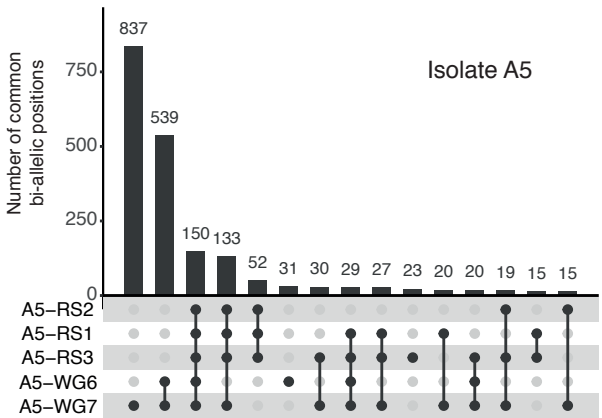
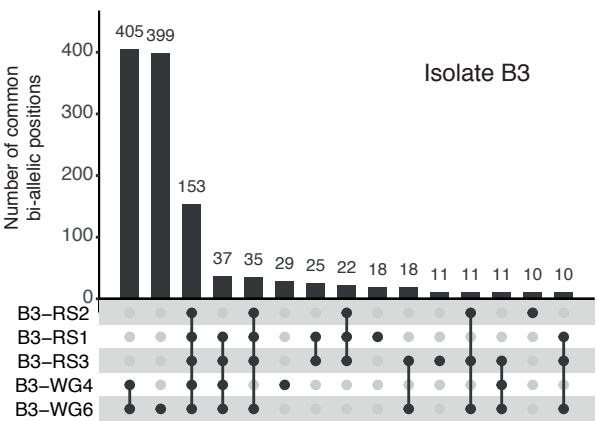

Mean number of bi−allelic sites per 1 kb region (sites/kb)
0.0
0.5
1.0
1.5
2.0
10
20
30
40
50
Mean depth coverage
A5−WG7
A4−WG9
C2−WG0
A1−WG4
B3−WG6
A4−WG0
A5−RS3
A5−RS1
C2−RS3
C2−WG3
A4−RS2
A4−RS1
A5−RS2
C2−RS2
C2−RS1
A4−RS3
C2−WG1
A5−WG6
A1−RS3
A1−WG3
A1−RS2
B3−RS3
B3−WG4
A1−RS1
B3−RS1
A4−WG7
B3−RS2

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

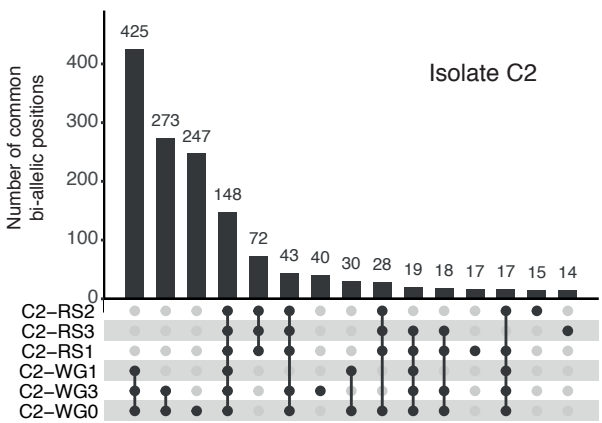

Isolate C2
Number of common bi-allelic positions
0
100
200
300
400
425
273
247
148
72
43
40
30
28
19
18
17
17
15
14
C2−RS2
C2−RS3
C2−RS1
C2−WG1
C2−WG3
C2−WG0

Figure 4

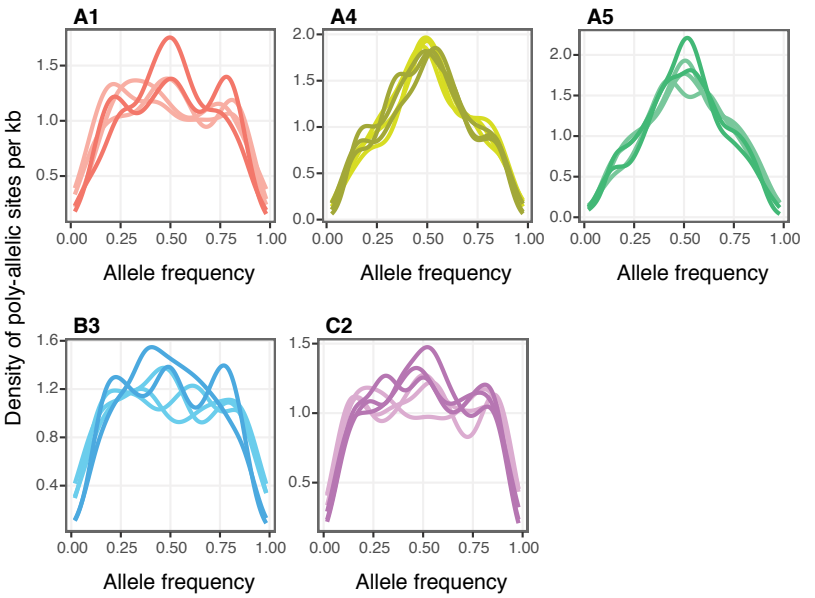

A1
A4
A5
B3
C2
Density of poly-allelic sites per kb
Allele frequency
0.00
0.25
0.50
0.75
1.00

Figure 5

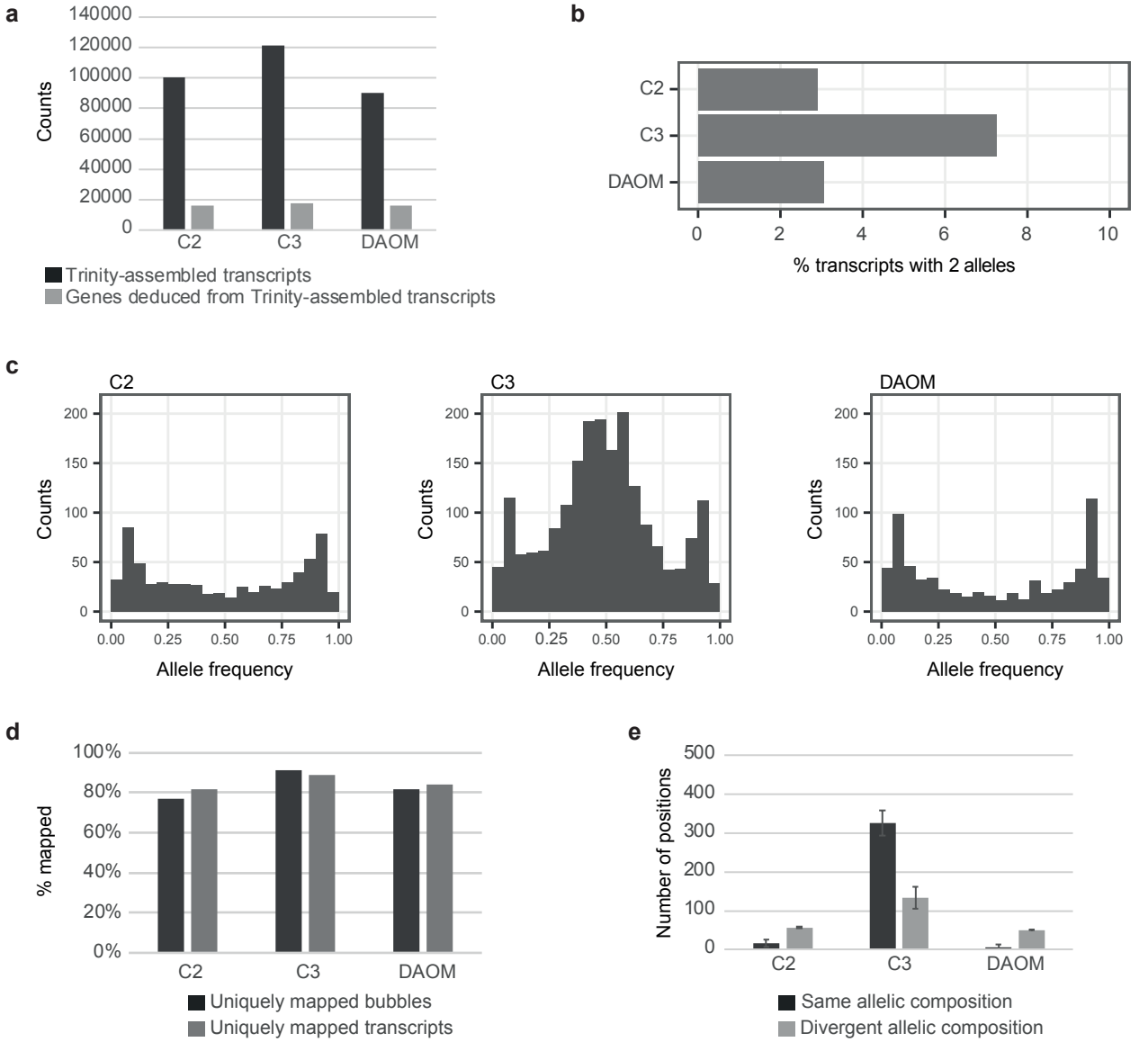
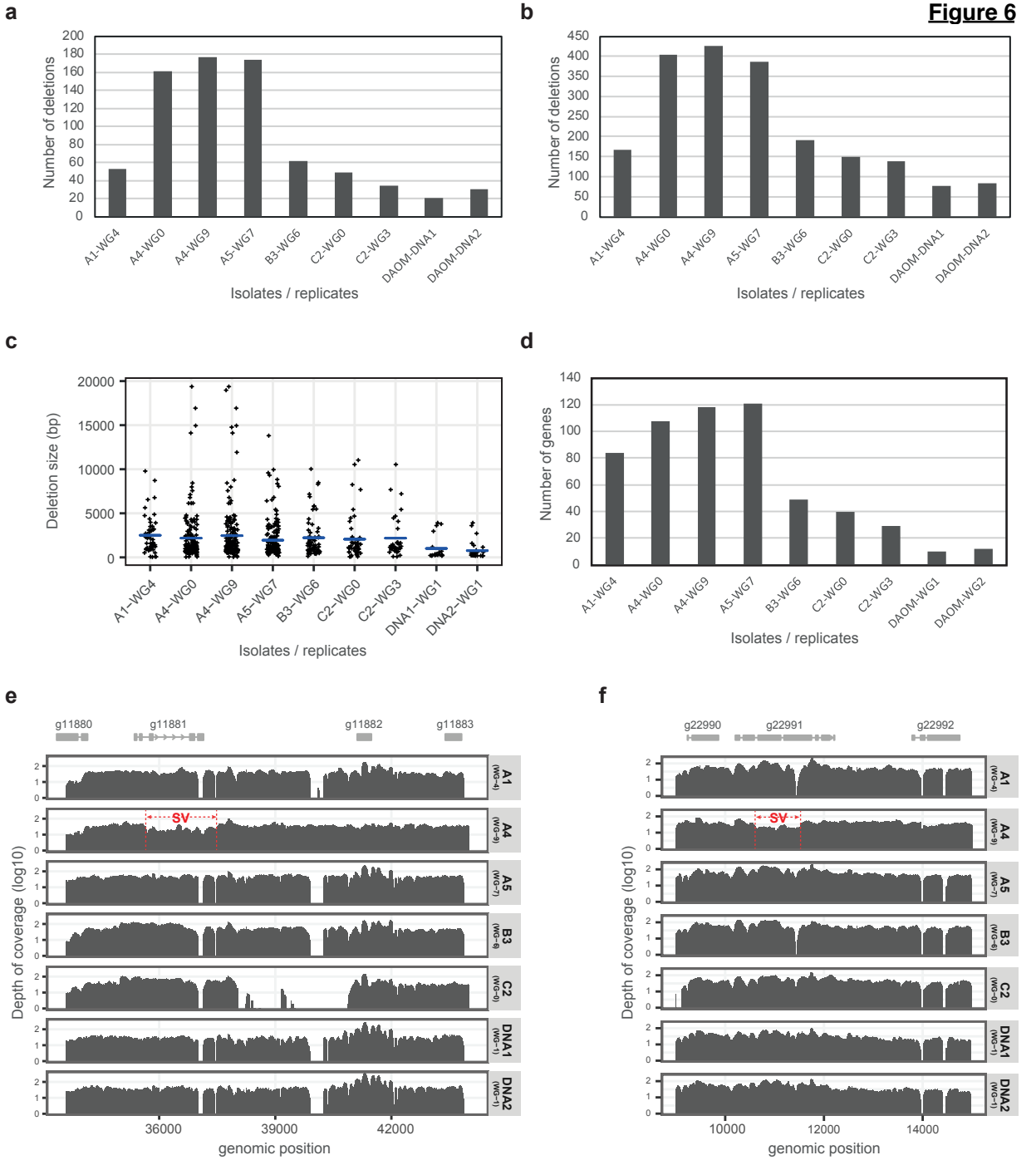

a
Counts
140000
120000
100000
80000
60000
40000
20000
0
C2
C3
DAOM
Trinity-assembled transcripts
Genes deduced from Trinity-assembled transcripts
b
C2
C3
DAOM
0
2
4
6
8
10
% transcripts with 2 alleles
c
C2
C3
DAOM
Counts
200
150
100
50
0
0.00
0.25
0.50
0.75
1.00
Allele frequency
d
% mapped
100%
80%
60%
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