Université de Montréal

The evolution of inter-genomic variation in arbuscular mycorrhizal fungi

par

Eva Boon

Institut de Recherche en Biologie Végétale

Faculté des Arts et des Sciences

Thèse présentée à la Faculté des Arts et des Sciences
en vue de l'obtention du grade de thèse de doctorat
en Sciences Biologiques

March 2012

© Eva Boon, 2012

Université de Montréal

Faculté des études supérieures et postdoctorales

Cette thèse intitulée :

The evolution of inter-genomic variation in arbuscular mycorrhizal fungi

Présentée par :

Eva Boon

a été évaluée par un jury composé des personnes suivantes :

Bernard Angers, président-rapporteur

Mohamed Hijri, directeur de recherche

Francois-Joseph Lapointe, membre du jury

David Walsh, examinateur externe

Frédéric Bouchard, répresentant du doyen

Résumé

Contexte

Les champignons mycorhiziens à arbuscules (AMF) établissent des relations symbiotiques avec la plupart des plantes grâce à leurs réseaux d'hyphes qui s'associent avec les racines de leurs hôtes. De précédentes études ont révélé des niveaux de variation génétique extrêmes pour des loci spécifiques permettant de supposer que les AMF peuvent contenir des milliers de noyaux génétiquement divergents dans un même cytoplasme. Si aucun processus de reproduction sexuée n'a jusqu'ici été observé chez ces mycorhizes, on constate cependant que des niveaux élevés de variation génétique peuvent être maintenus à la fois par l'échange de noyaux entre hyphes et par des processus fréquents de recombinaison entre noyaux. Les AMF se propagent par l'intermédiaire de spores qui contiennent chacune un échantillon d'une population initiale de noyaux hétérogènes, directement hérités du mycélium parent. À notre connaissance les AMF sont les seuls organismes qui ne passent jamais par un stade mononucléaire, ce qui permet aux noyaux de diverger génétiquement dans un même cytoplasme. Ces aspects singuliers de la biologie des AMF rendent l'estimation de leur diversité génétique problématique. Ceci constitue un défi majeur pour les écologistes sur le terrain mais également pour les biologistes moléculaires dans leur laboratoire. Au-delà même des problématiques de diversité spécifique, l'amplitude du polymorphisme entre noyaux mycorhiziens est mal connue. Le travail proposé dans ce manuscrit de thèse explore donc les différents aspects de l'architecture génomique singulière des AMF.

Résultats

L'ampleur du polymorphisme intra-isolat a été déjà observée pour la grande sous-unité d'ARN ribosomal de l'isolat *Glomus irregulare* DAOM-197198 (précédemment

identifié comme *G. intraradices*) et pour le gène de la *polymerase1-like* (*PLS*) de *Glomus etunicatum* isolat NPI. Dans un premier temps, nous avons pu confirmer ces résultats et nous avons également pu constater que ces variations étaient transcrites. Nous avons ensuite pu mettre en évidence la présence d'un goulot d'étranglement génétique au moment de la sporulation pour le locus *PLS* chez l'espèce *G. etunicatum* illustrant les importants effets d'échantillonnage qui se produisaient entre chaque génération de spore. Enfin, nous avons estimé la différentiation génétique des AMF en utilisant à la fois les réseaux de gènes appliqués aux données de séquençage haut-débit ainsi que cinq nouveaux marqueurs génomiques en copie unique. Ces analyses révèlent que la différenciation génomique est présente de manière systématique dans deux espèces (*G. irregulare* et *G. diaphanum*).

Conclusions

Les résultats de cette thèse fournissent des preuves supplémentaires en faveur du scénario d'une différenciation génomique entre noyaux au sein du même isolat mycorhizien. Ainsi, au moins trois membres du genre *Glomus*, *G. irregulare, G. diaphanum* and *G. etunicatum*, apparaissent comme des organismes dont l'organisation des génomes ne peut pas être décrit d'après un modèle Mendélien strict, ce qui corrobore l'hypothèse que les noyaux mycorhiziens génétiquement différenciés forment un pangenome.

Mots-clés *Glomus*, mycorhizes à arbuscules, hétérogénéité génétique intra-organismale, polymorphisme transcriptomique, polymorphisme des ARNs ribosomiques, polymorphisme nucléaire, goulot d'étranglement génétique, hétérokaryose, anastomose

Abstract

Background

Arbuscular mycorrhizal fungi (AMF) are root-inhabiting fungi whose hyphal networks form symbioses with plants. Previous studies have revealed extremely high levels of genetic variation for some loci, which has lead to the proposition that AMF contain thousands of genetically divergent nuclei that share the same cytoplasm, i.e. they are heterokaryotic coenocytes. No reproductive stage has as yet been observed in AMF, yet evidence is accumulating that the observed high levels of diversity could be maintained by the exchange of nuclei between hyphal systems and (meiotic) recombination. AMF spores contain varying fractions of this heterogeneous population of nuclei, which migrate directly from the parent mycelium. To our knowledge, AMF are the only organisms that never pass through a single nucleus stage in their life cycle, which allows nuclei to diverge into genetically distinct nuclei within the same cytoplasm. Thus, estimating genetic diversity in arbuscular mycorrhizal fungi (AMF) is a major challenge, not only for ecologists in the field but also for molecular biologists in the lab. It is unclear what the extent of polymorphism is in AMF genomes. The present thesis investigates different aspects of this peculiar genome organization.

Results

The second chapter in this thesis confirms the extensive intra-isolate polymorphism that was previously observed for large subunit rDNA (in *G. irregulare* DAOM-197198) and the polymerase1-like gene, *PLS* (in *G. etunicatum*), and shows that this polymorphism is transcribed. In the third chapter I report the presence of a bottleneck of genetic variation at sporulation for the *PLS* locus, in *G. etunicatum*. Analyses in the fourth chapter, based on a conservative network-based clustering approach and five novel single copy genomic markers, reveal extensive genome-

wide patterns of diversity in two different AMF species (*G. irregulare* and *G. diaphanum*).

Conclusions

The results from this thesis provide additional evidence in favor of genome differentiation between nuclei in the same isolate for AMF. Thus, at least three members of the *Glomus* genus, *G. irregulare, G. diaphanum* and *G. etunicatum* appear to be organisms whose genome organization cannot be described by a single genome sequence: genetically differentiated nuclei in AMF form a pangenome.

Keywords

Glomus, arbuscular mycorrhizal fungi, intra-organismal genetic heterogeneity, transcriptome variation, rDNA variation, nuclear variation, genetic bottleneck, heterokaryosis, anastomosis, polymerase1-like sequence

Acknowledgements

> *"No man is an Iland, intire of itselfe; every man is a peece of the Continent, a part of the maine; if a Clod bee washed away by the Sea, Europe is the lesse, as well as if a Promontorie were, as well as if a Manor of thy friends or of thine owne were; any mans death diminishes me, because I am involved in Mankinde; And therefore never send to know for whom the bell tolls; It tolls for thee."* John Donne (1572-1631), Meditation XVII, Devotions upon Emergent Occasions

I am grateful to acknowledge the people who much improved my work and helped me complete this thesis, even though of course all errors are mine. This text would not have existed without the enthusiasm, support, instruction and insight of Dr. Dr. Eric Bapteste, to whom I am much indebted. Eric helped me on my way from the first beginnings to the completion of this work, and gave me the confidence to confront other's implicit assumptions and mine. I could not have written chapter four without the help of Sébastien Halary, whose computing power has saved the day more than once and whose scientific rigor steered us through the quagmire that is *Glomus*. Mohamed, I cannot begin to describe your influence on my work. I know I have much benefited from your unbridled trust.

I am greatly obliged to Dr. Ford W. Doolittle, who welcomed me in his laboratory and helped me sharpen my ideas on fuzzy boundaries. Dr. Philippe Lopez kindly hosted me on several occasions in his research group in Paris, where I finally learned to face the command line. The Institut de Recherche en Biologie Végétale graciously financed a large part of my *stage* and conference travels through their travel grants, which were essential to many a trip budget.

Denis Beaudet helped me when I was unable to finish the last part of my labwork. Yves Terrat performed analyses that unfortunately did not make it into the final version of chapter four, kept up my spirits during the last months of writing, and corrected my French. Cristina Micali read the first draft of this thesis, and provided valuable comments –not only on the science aspect of our work.

None of this labor would have been finished without the patient support of Antonin Serpereau, who has pulled me through one of the most difficult periods in my life as my bodyparts gradually fell to pieces. My parents, brother and sister have always held up me with unconditional love and confidence, which is the most solid foundation anyone could hope for. Finally, I would like to thank my friends in Montréal and overseas for engaging and discussing my work with them, and to help me disconnect at the same time. One only fully understands one's work when it is explained to others.

Table of Contents

RÉSUMÉ V

ABSTRACT VII

ACKNOWLEDGEMENTS IX

TABLE OF CONTENTS XI

LIST OF TABLES XIII

LIST OF FIGURES XV

ABBREVIATIONS XIX

1. INTRODUCTION 1

2. INTRA-ISOLATE GENOME VARIATION IN ARBUSCULAR MYCORRHIZAL FUNGI PERSISTS IN THE TRANSCRIPTOME 25

3. ALLELIC DIFFERENCES AMONG SISTER SPORES SUGGEST GENETIC DRIFT AT SPORULATION IN AN ARBUSCULAR MYCORRHIZAL FUNGUS 53

4. ESTIMATING GENOME DIVERSITY IN ARBUSCULAR MYCORRHIZAL FUNGI: THE CASE FOR A PANGENOME IN AMF 85

5. DISCUSSION 123

REFERENCES 148

List of Tables

Table 2-1 Primers used for LSU rDNA and *PLS* amplification 33
Table 2-2 LSU rDNA polymorphism estimates 38
Table 2-3 Polymorphism estimates for *PLS* (*G. etunicatum*) 40
Table 2-4 Description of LSU regions (structure information is based on Schnare et al.(1996)) 47
Table 2-5 Additional *G. irregulare* (unless specified otherwise) genomic sequences from GenBank used for inter-isolate comparison in the LSU rDNA analysis .. 48
Table 3-1 Alpha-diversity estimates for the 1st (partial) and 2nd exonic sequence of *PLS.* Alpha diversity was calculated for individual spores, spores grouped per plant and for all spores combined. 66
Table 3-2 P-values of three different tests that evaluate whether *PLS* alleles show the same structure of genetic diversity between spores 68
Table 3-3 Statistical tests for the difference between 'plants' community parameters 69
Table 3-4 Changes on the amino acid level. Letters refer to groups in figure 3-2.. 70
Table 3-5 Environmental conditions under which tomato seedlings were grown ... 77
Table 4-1 Genomes used for pyrosequencing simulations 95
Table 4-2 Results from clustering analyses 100
Table 4-3 Amplification and genetic diversity of single copy markers 104
Table 4-4 Psi Blast results for single copy markers 104
Table 4-5 Parameters for *metasim* genome simulations 111
Table 4-6 Fungal species used for initial alignment of genomes..... 115
Table 4-7 Primer sequence, annealing temperatures and amplicon length for five single copy markers 116

List of figures

Figure 1-1 AMF are part of a complex system of interactions involving plants, fungi and bacteria, which ultimately determines many aspect of community and ecosystem functioning. AMF potentially exchange information among themselves (different isolates are indicated in light/dark grey and in white), with their host plants (upper panel), and with soil bacteria (lower panel). Furthermore, AMF connect individual plants via their hyphal systems and thus potentially play a role in information exchange *between* plants. 3

Figure 1-2: AMF are coenocytic and multinucleate **a)** typical AMF coenocytic hyphae (H) that lack septa with nuclei (N: green dots) and mitochondria (M: red dots) **b)** a live AMF multinucleate spore (S). These images were obtained by confocal microscopy, courtesy of Dr Hijri. 5

Figure 1-3: What kind of marker is needed? Standard molecular biology techniques do not allow us to distinguish situation **a)**, in which a gene occurs in single copy in genetically divergent nuclei, from situation **b)**, in which a gene occurs in multiple copies within genetically similar nuclei. A third possible scenario **c)**, in which the nuclei are polyploid, has been excluded by Hijri *et al.* (2004): *G. irregulare*, at least, is haploid. 16

Figure 2-1 Within-isolate nucleotide diversity (π) (Nei 1987) along the 5' LSU rDNA of *Glomus irregulare* for the entire 2053 bp LSU rDNA, with a comparison of compensatory base pair changes within *Glomus sp.* isolate DAOM-197198 (white bars) and between isolate DAOM-197198 and other *G. irregulare* isolates (black bars). The dotted rectangle indicates the relative location of the 220 bp subset alignment. 38

Figure 2-2 A Neighbor-Net network for the *PLS* locus, based on uncorrected P-distances. Only exons were taken into account. Reticulation of the branches indicates uncertainties due to homoplasy (recombination or multiple hits). Sequences that belong to the *PLS*1 sequence type are represented in red and

sequences that belong to the *PLS*2 sequence type are represented in blue. The shape of the symbol corresponds to the mode of sequencing and/or origin of the template; squares are sequences from the cDNA library, circles are gDNA sequences obtained by cloning and stars are gDNA sequences obtained by pyrosequencing. Numbers in the symbols correspond to the frequency of the allele in question. Bootstrap support values >70 are depicted above branches (1000 replicates). Scale bar indicates substitutions per site. The shading on the edges of the network reflects confidence values associated with each split 41

Figure 2-3 Absolute quantification of rDNA copy number. Results of real-time quantitative PCR showing linear regressions of the cycle threshold (CT values) and the Log concentration of **a)** plasmid DNA that was used as standard and **b)** LSU rDNA of *Glomus sp*. isolate DAOM-197198. Three replicate amplifications were performed for each sample. Data points shown in the graphs represent the CT values of the three replicates 51

Figure 3-1 Diagram of the experimental setup showing three parent spores taken from a single isolate of *G. etunicatum* grown in pot culture to produce eight progeny spores for pyrosequencing. 60

Figure 3-2 Phylogenetic analysis on genetic divergence between alleles (n=356). Evolutionary relations between alleles were inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa *et al.* 1985) with a discrete gamma distribution (5 categories (+G, parameter = 11.0677)). Only bootstrap values higher than 70 are depicted. Scale bar indicates number of substitutions per site. Positions with less than 95% site coverage were eliminated, in total 92 nucleotide sites were used. 72

Figure 3-3 Rarefaction analyses for **a)** all spores pooled together (all) **b-d)** spores pooled by plant (A-B-C) and **e-l)** individual spores (A1-A2-A4-B1-B2-B4-C1-C2). The number of recovered alleles (y axis, grey line, 95% confidence intervals indicated by vertical lines) is compared to the maximum Chao1 value

(Chao *et al.* 2005), which is the estimated minimum richness for each group (solid black line, 95% confidence intervals in dotted lines). 83

Figure 3-4 Nucleotide diversity along the fragment of the Polymerase-1 like gene under study. 84

Figure 4-1 Frequencies of similarity clusters compared to average PID per cluster for all *Glomus* and control genomes **a)** *G. irregulare* DAOM 234179 run 2 **b)** *G. diaphanum*. 102

Figure 4-2 Two variables employed to describe similarity clusters, **a)** the PID is the percentage of identical positions on the shortest sequence of an aligned reads pair and **b)** the clustering coefficient, which is the number of aligned reads pairs in the cluster, divided by the maximum number of possible pairs. 110

Figure 4-3 Frequencies of similarity clusters compared to average PID per cluster for *Glomus* and control genomes **a)** *Glomus irregulare* DAOM 197198 and **b)** *G. irregulare* DAOM 234179 run 1. 112

Figure 4-4 Scatterplot between PID and clustering coefficient, for all *Glomus* and control genomes **a)** *G. irregulare* DAOM 234179 run 1 **b)** *G. irregulare* DAOM 234179 run 2 **c)** *G. irregulare* DAOM 197198 **d)** *G. diaphanum*. Colors correspond to the number of reads that are represented by the respective datapoints (see legendas: for simplicity only one agenda is depicted per page). The shape of the datapoint refers to the reads set: circle, *Glomus* (corresponding to the respective graphs); square, *P. graminis*; diamond, *T. melanosporum*; triangle, *P. triticina.* 114

Figure 4-5 Rarefaction curves for single copy makers **a)** 1287 *G. irregulare* DAOM 234328 **b)** 1631 *G. irregulare* DAOM 234328 **c)** 2150 *G. irregulare* DAOM 197198 **d)** 628 *G. irregulare* DAOM 197198 **e)** 628 *G. irregulare* DAOM 234328 **f)** 628 *G. diaphanum* **g)** 663 *G. irregulare* DAOM 197198 **h)** 663 *G. irregulare* DAOM 234328 **i**) 663 *G. diaphanum*. The number of recovered alleles (y axis, grey line, 95% confidence intervals indicated by vertical lines) is compared to the maximum Chao1 value (Chao *et al.* 2005), which is the

estimated minimum richness for each group (solid black line, 95% confidence intervals in dotted lines). 121
Figure 4-6 Allele distributions for single copy markers that have more than two alleles which occur at frequencies higher than one **a)** 628 and **b)** 663..... 122

Abbreviations

AMF, arbuscular mycorrhizal fungi; bp, basepair; *c.*, *circa*; CBC, compensatory base changes; cDNA, complementary DNA; est, expressed sequence tag; FISH, Fluorescent In Situ Hybridization; gDNA, genomic DNA; GTP, guanosine-5'-triphosphate; ITS, Intergenic Spacer; IGH, intra-organismal genetic heterogeneity; LSU, large sub-unit; LGT, Lateral Gene Transfer; MOTU, Molecular Operational Taxonomic Unit; OTU, Operational Taxonomic Unit; PCR, Polymerase Chain Reaction; PID, percentage of identical positions; *PLS*, Polymerase1-like sequence; RT-PCR, Real Time Polymerase Chain Reaction; rDNA, ribosomal RNA-encoding DNA; SSU, small sub-unit; SD, standard deviation; Ta, annealing temperature; TE, transposable elements; VT, Virtual Taxon

1. Introduction

Arbuscular mycorrhizal fungi (AMF -phylum Glomeromycota) are extraordinary creatures. Earliest fossils resembling this group of organisms date back as far as the Ordovician, at least 455 million years ago (Redecker *et al.* 2000) and it is well possible that AMF evolved alongside the first land plants (Brundrett 2002). AMF associate with a wide range of land and aquatic plants (an estimated 60-80% according to Wang and Qiu (2006)) and they might have been crucial elements in the advance of the ancestors of modern plants from water to land (Malloch *et al.* 1980; Redecker *et al.* 2000; Wang *et al.* 2010). AMF are still very important to plants, which depend on AMF for various services from resistance to soil-borne pathogens (Azcon-Aguilar and Barea 1996) to improved fitness in polluted environments (Hildebrandt *et al.* 1999; 2007). AMF also have a positive effect on a wide range of plant and ecosystem performance parameters (Van der Heijden *et al.* 1998; Harrison 1999) and in turn benefit from their association with plants by obtaining carbohydrates from their host, so much so that they cannot survive independently (i.e. they are obligate biotrophs).

The symbiosis between plants and AMF is not a straightforward one to one relationship: it is probably more accurate to describe this interaction in terms of exchanges between communities. The responses of plants to this symbiosis are better studied than the fungal responses; plants show different growth responses when they are associated with different AMF (Munkvold *et al.* 2004; Koch *et al.* 2006) and with different AMF communities (Bever 2002; Vogelsang *et al.* 2006). A single plant can be colonized by several fungal species at the same time, which interact amongst themselves whilst at the same time interacting with the plant host (Van der Heijden *et al.* 2003; Kaiser and Charlesworth 2009). AMF, in their turn, show different sporulation rates when associated with different plants (Bever *et al.* 1996) and are able to associate with many different plant species simultaneously.

Also, they seem to influence the relationships *between* plants (Giovannetti and Gianinazzi-Pearson 1994; Dodd *et al.* 1996). The extensive mycelium of AMF forms, as it were, an underground network that potentially connects entire plant communities. To add another layer to the picture, certain bacterial symbionts seem to have a beneficial effect on important features of the mycorrhizal symbiosis, such as nutrient mobilization from soil minerals, fixation of atmospheric nitrogen, and protection of plants against root pathogens (Frey-Klett *et al.* 2007; Bonfante and Anca 2009). AMF are thus part of a complex system of interactions involving plants, fungi and bacteria, which ultimately determines many aspects of community and ecosystem functioning (figure 1-1).

The exact influence of AMF genetic diversity on individual plants and plant communities, however, is far from clear. A major impediment to the pursuit of this question has been the difficulty to recognize AMF diversity. Historically, AMF identification and taxonomy has been based on spore morphology but unfortunately accurate interpretation requires a lot of expertise. Spore characteristics suitable for taxon assignment are rather limited and there is therefore not always sufficient information to resolve questions of species identity (Sanders 2004). An obvious solution would be to employ genetic markers as tools for species recognition, classification and quantification. This is where one of the most interesting aspects of AMF biology, its genetics, becomes relevant.

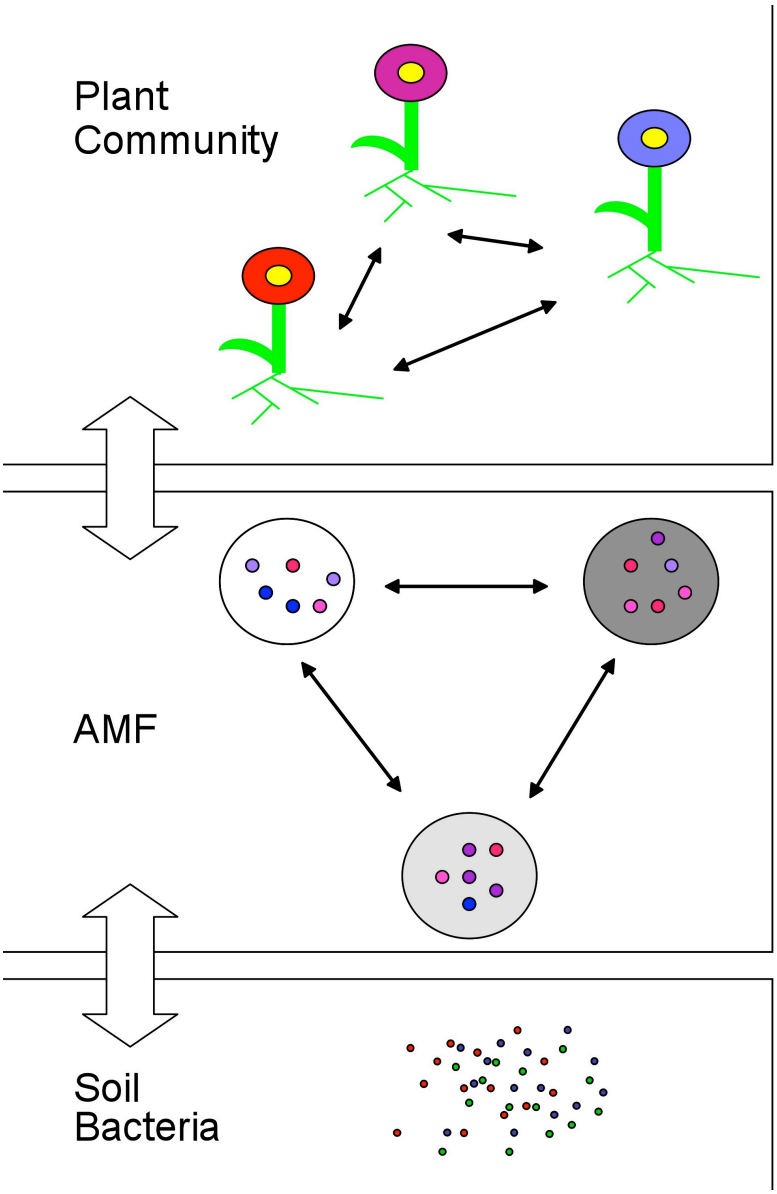


Figure 1-1 AMF are part of a complex system of interactions involving plants, fungi and bacteria, which ultimately determines many aspect of community and ecosystem functioning. AMF potentially exchange information among themselves (different isolates are indicated in light/dark grey and in white), with their host plants (upper panel), and with soil bacteria (lower panel). Furthermore, AMF connect individual plants via their hyphal systems and thus potentially play a role in information exchange *between* plants.

1.1. Literature review on AMF genetics

1.1.1. AMF are multigenomic

Arbuscular mycorrhizal fungi are coenocytic, i.e. there are no divisions (septa) between cells and the continuous cytoplasm contains multiple nuclei (figure 1-2a). A large amount of genetic variation has been reported within isolates[1] for ribosomal DNA and protein-coding genes. There is strong evidence suggesting that this variation is partitioned between different genomes: (1) intraspecific ITS polymorphism in *Scutellospora castanea* was structured within multinucleate spores (Hijri *et al.* 1999), (2) ITS and *BiP* alleles were shown to be located in separate nuclei using FISH (Fluorescent In Situ Hybridization) techniques (Kuhn *et al.* 2001; Kuhn 2003), (3) the reported high levels of variation for the *PLS* locus cannot be due to polyploidy –AMF seem to be haploid for all species investigated so far (Hijri and Sanders 2004), just like most other fungi, (4) at least for the protein-coding marker *PLS* and the LSU rDNA, observed within-isolate variation was specifically not due to gene duplications within nuclei (Hijri and Sanders 2005; Boon *et al.* 2010), (5) transcriptome data shows that polymorphism is widespread in the genome, and that >1 variant of each polymorphic gene is transcribed (Boon *et al.* 2010; Tisserant *et al.* 2012).

The above arguments lend strong support to the differentiated state of AMF nuclei within a single isolate. The following overview will attempt to shed light on the *extent* of intra-isolate polymorphism in AMF.

[1] An isolate is an AMF culture that was started from a single spore and is maintained in the laboratory.

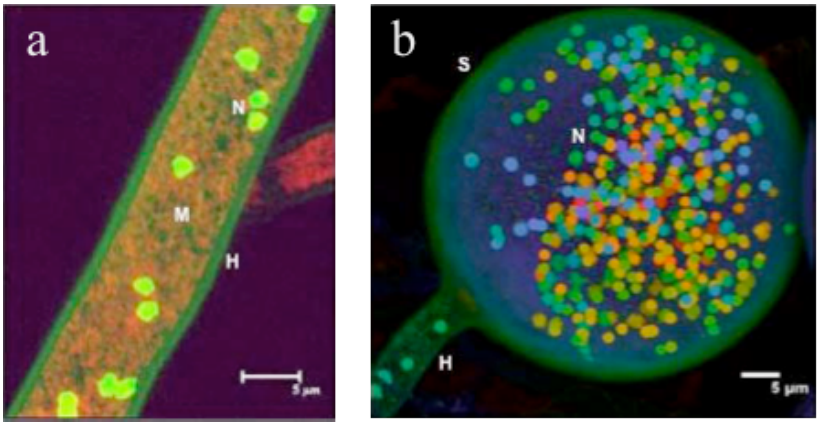


Figure 1-2: AMF are coenocytic and multinucleate **a)** typical AMF coenocytic hyphae (H) that lack septa with nuclei (N: green dots) and mitochondria (M: red dots) **b)** a live AMF multinucleate spore (S). These images were obtained by confocal microscopy, courtesy of Dr Hijri.

1.1.2. Intra-isolate variation in AMF RNA-encoding genes

Ribosomal tandem repeats are generally assumed to be kept genetically homogeneous within genomes through a process called concerted evolution (Arnheim 1983). Moreover, ribosomal DNA (rDNA) is subject to strong negative selection pressures, since even single base changes can result in the loss of complementarity and a non-functioning ribosome, particularly in regions that are structurally important to the secondary structure of the folded rDNA. Still, intragenomic rDNA diversity is far from uncommon in many taxa. Intragenomic sequence variability will accumulate if the rate of gene conversion is lower than the rate of mutation or if concerted evolution is slower than speciation. It is also possible that differences between mutation rates and DNA repair rates lead to an accumulation of sequence variants in genes with multiple copies (Elder and Turner 1995). Examples of intragenomic rDNA variation have been reported for all rDNA loci (i.e. Large Subunit or LSU, Small Subunit or SSU and the Intergenic Spacer or ITS) over a wide taxonomic range of organisms including the ITS regions in cactus (Bradley and Hillis 1997) and serviceberry bushes (Hugenholtz and Huber 2003), and across the entire rDNA cistron in salmonid fish (Gonzalez *et al.* 2005), humans

(Komatsoulis and Waterman 1997), dinoflagellates (Wang and Wang 1996; Thornhill *et al.* 2007; Seddas *et al.* 2008) and fungi (Simon and Weiss 2008).

Arbuscular mycorrhizal fungi show intra-isolate variation for the LSU (Clapp *et al.* 2001; Rodriguez *et al.* 2001), the SSU (Clapp *et al.* 1999; Schussler *et al.* 2001a) and for the ITS regions (Sanders *et al.* 1995; Lloyd-MacGilp *et al.* 1996; Hijri *et al.* 1999; Pringle *et al.* 2000; Kuhn *et al.* 2001). Given the proposed multigenomic state of AMF, this intragenomic variation can also be explained by the partitioning of different alleles *between* genomes of AMF isolates. Indeed, Kuhn *et al* (2001) used fluorescent *in situ* hybridization techniques to show that different ITS alleles are physically partitioned between nuclei. Finally, it seems that even tandem repeat copy numbers of RNA-encoding genes differ between isolates from the same population (Corradi *et al.* 2007).

There are two essential differences between intragenomic rDNA variation reported for AMF and rDNA variation in other organisms. First, the *location* of rDNA variation in AMF is potentially different. If we assume genetic homogeneity in other organisms, rDNA variation in these organisms is necessarily partitioned between tandem repeats *within* a genome. This assumption will be questioned in the discussion. In *Glomus,* we cannot assume genetic homogeneity between genomes. Thus, it is possible that variation occurs between *or* within genomes. Second, the *amount* of genetic variation in AMF is quite unusual. Levels of sequence divergence can be as high as 17% in AMF RNA-encoding genes (Boon *et al.* 2010) and tandem repeat copy numbers differ between isolates from the same population (Corradi *et al.* 2007). Especially copy number differences are considered to be major evolutionary changes that are usually associated with higher taxonomic levels in fungi such as species and genera (Simon and Weiss 2008). Thus, even though the occurrence of intra-isolate rDNA variation in AMF is not unprecedented in itself, there is no equivalent example in the literature for the location and extent of rDNA variation in AMF.

1.1.3. Intra-isolate variation in AMF protein-coding genes

The first experiments showing the occurrence of intra-isolate sequence variation in a protein-coding gene were completed by Kuhn *et al.* (2001) on a gene encoding the binding protein *BiP*. The authors obtained 15 variants of *BiP* from a single *Glomus irregulare* isolate, with up to 8% divergence between homologs. These variants must necessarily exist between nuclei, since *BiP* copy number was estimated to only three copies per genome in *G. irregulare,* isolate DAOM197198 (Corradi *et al.* 2007). The co-existence of different isoforms within a single isolate was also reported for ß-tubulins (Corradi *et al.* 2004b) and P-type IID ATPases, for which nucleotide variation of up to 3-5 % was reported (Corradi and Sanders 2006). Notably, for ß-tubulins and P-type IID ATPases, alleles did not only differ from each other by nucleotide substitutions but also through insertion-deletion (indel) variations, which are structural genomic changes that are considered to be rare events. P-type IID ATPases were also reported to vary in copy number between isolates (Corradi and Sanders 2006). However, because the absolute copy number was not estimated for these loci, it is unclear whether the reported variation in homologous sequences represents divergent orthologues (sequences located in different genomes) or paralogues (sequences located in the same genome).

1.1.4. The challenge of estimating genetic variation in AMF

Studies on genetic variation in AMF are much complicated by the large amounts of variation observed between nuclei, both within and between isolates. This variation can be so extreme that ribosomal sequences that have been retrieved from the same isolate are more dissimilar than sequences from different isolates *from completely different continents* (Boon *et al.* 2010). Extensive sequence and gene copy polymorphisms have even been observed between

isolates from the same field (Corradi and Sanders 2006; Corradi *et al.* 2007). There are other issues with estimating total genetic variation in the field, such as the delineation of the individual and the species. This issue is explored in more detail in section 5.4. The total amount of genetic variation within one isolate is hard to estimate, but would depend on the amount of nuclei within one hyphal system and the amount of genetic variation between these nuclei. At this point, very little data is available on these variables. Nevertheless, a typical spore can contain hundreds to thousands of nuclei (a number which is species-dependent, see Bianciotto and Bonfante (1992) and Becard and Pfeffer (1993)) and therefore the total amount of genetic variation found in an isolate is potentially considerate. It is as yet not clear how much genetic variation can be found in a single AMF isolate for any particular locus, however there are two studies that have so far addressed this question. Here, I will not discuss studies that are based on field studies, since these investigations face the added difficulty of defining 'the individual' and cannot be compared to studies that are performed on the laboratory definition of the AMF individual, which is the isolate.

The first evaluation of maximum sequence diversity is mentioned in Kuhn's doctoral thesis (Kuhn 2003). He estimated the maximum amount of variants of the *BiP* protein-coding gene that could be expected to occur within a single AMF isolate, based on a sampling effort of 166 clones. He used Monte Carlo simulations to show that the maximum number of variants should be between 28 and 70, depending on the underlying assumptions. Second, Zimmerman (2009) attempted to exhaustively sample diversity in clones from the protein-coding gene *PLS*. Even though she obtained 97 alleles from a total sampling effort of 159 cloned sequences, genetic variation was not sampled exhaustively.

With the rising availability of massive parallel sequencing techniques, recent studies have attempted to quantify AMF diversity in the field. The method of choice is pyrosequencing, which combines a considerable data quantity with relatively long

read lengths. The targeted region for estimating molecular diversity in the field is mostly the SSU rDNA (Öpik *et al.* 2009; Dumbrell *et al.* 2011; Moora *et al.* 2011), up to the extent that an ecological database devoted entirely to SSU sequences has been made available (Öpik *et al.* 2010). However, the LSU and ITS regions have also been employed for ecological inventories of AMF genetic diversity (Blaalid *et al.* 2011; Lekberg *et al.* 2012, respectively).

There are two major impediments for comparing the sequence diversity in these studies. First, the use of different loci makes direct comparisons meaningless, as levels of variation change with the marker in question. Second, there is as yet no consensus on what level of sequence divergence constitutes an OTU (Operational Taxonomic Unit), MOTU (Molecular Operational Taxonomic Unit) or VT (Virtual Taxon) in AMF. Cutoff points for taxon delineation vary between 97% and 98.5 %, and stringency criteria for filtering reads quality are not uniform. Section 5.4. provides a more detailed discussion on the taxon and the individual in AMF. For now, notwithstanding the abovementioned impediments, it is clear that molecular divergence between rDNA sequences in the field is high: only two out of five studies, both using the SSU rDNA, managed to exhaustively sample all variation (Öpik *et al.* 2009; Moora *et al.* 2011). Where, we then ask, does this enormous amount of variation come from?

1.1.5. The evolution of genetically differentiated genomes in AMF

The key to understanding genome diversity in AMF lies in two elements of their life history. First, AMF spores are multinucleate *at all life stages* (figure 2b). The large amounts of genetic variation that are found in AMF are passed on without a significant bottleneck of genetic variation between generations (Marleau *et al.* 2011). This bottleneck has long since been considered an essential characteristic of multicellular development, going all the way back to Weisman's

germ plasm theory (Weismann 1893). The role of the bottleneck of genetic variation in the broader context of evolutionary theory will be reviewed in more detail in the discussion. For now, I will focus on AMF. At sporulation, nuclei migrate from the hyphae into the spore with limited mitosis inside the spore (Jany and Pawlowska 2010; Marleau *et al.* 2011). The presence of different nuclei in the somatic cells, a state called heterokaryosis, is not unusual for fungi. AMF are distinct from other fungi because their multinucleate spores that do not undergo a bottleneck preserve the somatic mutations that occur in every genome during the lifespan of a nucleus.

The second important element of AMF life history is anastomosis, which is the fusion of hyphae of different fungal bodies. This process is relatively well studied in AMF, and it has been demonstrated that nuclei are exchanged after anastomosis and there are indications that nuclei can persist in the accepting fungus after anastomosis, if the donor and receptor are sufficiently related (Giovannetti *et al.* 1999; Giovannetti *et al.* 2001). However, little is known about the long-term fate of the transferred nuclei. Recent experiments on fusion and segregation of related AMF strains indicate that transferred nuclei persist in the receiving cytoplasm and can even change the original phenotype of the receiving hyphal system (Croll 2009; Angelard *et al.* 2010; Angelard and Sanders 2011).

On a population level, anastomosis in itself is not a source of genetic variation in AMF. However, anastomosis is an excellent mechanism to counter the effects of genetic drift, which, even in the absence of a significant bottleneck of genetic variation, would eventually lead to the extinction of genetic variation between AMF nuclei. Models that describe anastomosis in AMF confirm the importance of this process in the maintenance of genetic variation (Bever and Wang 2005; Bever *et al.* 2008).

Thus, even though ‘one individual, one genome’ might be a standard throughout biology, AMF clearly seem to have a problem with this adagio and

appear to have a unique mode of maintaining genetic variation. One might describe the organization of genetic variation in AMF as a population of nuclei that are bound by the same fungal body, i.e. the hyphal system. This thesis is an attempt to describe the evolutionary dynamics of this population.

1.2. Outline of research questions

The general aim of this thesis is to study the extent, maintenance and loss of genetic differentiation between nuclei within AMF isolates. Very little is known about genetic differentiation between nuclei in the AMF cytoplasm. This lacuna in our knowledge is partly due to technical limitations, and partly due to a lack of conceptual tools to analyze AMF genetics. The challenge of this thesis was to deal with these limitations, which will be enlarged upon in the general discussion (chapter 5).

Intra-isolate variation in AMF has been reported for rRNA genes (Sanders *et al.* 1995; Lloyd-MacGilp *et al.* 1996; Clapp *et al.* 1999; Hijri *et al.* 1999; Hosny *et al.* 1999; Pringle *et al.* 2000; Clapp *et al.* 2001; Kuhn *et al.* 2001; Rodriguez *et al.* 2004) and protein-coding genes (Kuhn *et al.* 2001; Kuhn 2003; Hijri and Sanders 2005; Boon *et al.* 2010), with alleles differing from each other by nucleotide substitutions and insertion-deletion (indel) variations. Even copy number polymorphism between isolates has been reported for some of the aforementioned loci (Corradi *et al.* 2007).

In this doctoral thesis, I ask whether (1) intra-isolate genetic variation persists in the AMF transcriptome (2) there is a loss of genetic variation at sporulation (3) genetic variation in AMF is wide spread through the genome.

1.2.1. Does intra-isolate genetic variation persist in the AMF transcriptome?

The occurrence of several gene copies in a genome can have very different evolutionary consequences. On the one hand, it can lead to intragenomic conflict. This intragenomic conflict can be solved through epigenetics, of which differential silencing of one or more gene copies is an example (Rapp *et al.* 2009). Therefore, if intragenomic conflict occurs between multiple copies of the same gene, evolutionary theory would lead us to predict that only a subset of this variation should be expressed in the transcriptome. On the other hand, of course, it is also possible that multiple gene copies do not lead to intragenomic conflict, but enhance or at least alter the function of the gene in question (Schrider and Hahn 2010). It is interesting to investigate these contrasting theories in AMF. Is the enormous genetic variation that we observe expressed or not?

Here, I am explicitly not concerned with mechanisms that increase genetic variation from a genome level to the transcriptome. Post-translational processes, such as alternative splicing and the very modular structure of proteins, have been extensively described and essentially increase diversity on a phenotypic level. Even though these mechanisms surely play a role in shaping the *Glomus* phenotype, here I am concerned with the reverse; how genotypic diversity inside the individual (here: the hyphal system) can lead to a functioning phenotype.

Most of what we know on intra-organismal transcriptome variation concerns ribosomal genes. This variation is often associated with the presence of pseudogenes (e.g. Thornhill *et al.* 2007), yet there are indications that this genomic variation plays a role in the functioning of the organism. For example, Kim *et al.* (2007) evaluated the importance of intragenomic variation in RNA-encoding genes on a functional level by isolating alleles from cDNA libraries. The authors report that in *Streptomyces* all of the heterogeneous rRNA molecules observed on a genomic level are expressed and assembled into ribosomes. Lopez-Lopez *et al.*

(2007) state that distinct ribosomes could play different functional roles in the cell. They describe how in the halophilic archeon *Haloarcula marismortui* different ribosomal RNA operons show distinct stabilities under varying temperatures, which is reflected in the secondary structure of these operons. In this case, intra-genomic variation in rDNA could represent functional divergence.

On a protein-coding gene level, I am not aware of any studies that investigate the persistence of intra-organismal variation in the transcriptome. In a way, we could state that gene duplications produce intra-organismal variation in the transcriptome. However, silencing events and unequal expression of gene variants occur very rapidly after gene duplication (Adams *et al.* 2003), probably to avoid abovementioned intragenomic conflicts. An interesting observation in this context is that Hosny *et al.* (1997) noted a percentage of methylated cytosines in the AMF *Scutellaspora castanea* that was much higher than values observed for other fungi, which possibly indicates a response to the many gene duplicates present in an AMF isolate–in spite of the fact that these copies are not located in the same nucleus.

Whilst both ribosomal and protein-coding genes of AMF have long been known to possess significant amounts of intra-isolate variation, it is not known whether this variation persists at the transcriptional level. In order to address this question, I investigated nucleotide polymorphisms in the 25S rRNA gene from *G. irregulare* and the *PLS* gene from *G. etunicatum*, from both genomic DNA and cDNA libraries (chapter 2).

1.2.2. Is there a loss of genetic variation at sporulation?

We know that at sporulation, multiple nuclei migrate into the developing spore. There seems to be no single nucleus stage in AMF life history. However, is the proportion of nuclei that migrate into a spore representative for the population

within the whole hyphal system? In other words, do AMF encounter a loss of genetic variation at sporulation? This question is important, since the answer to it will allow us to quantify a component of the rate at which genetic variation is lost in AMF. This rate, in its turn, will provide information on the role of genetic drift in AMF life history. Also, knowing how much variation is lost at sporulation will provide information on how often anastomosis or other processes such as recombination need to occur to replenish genetic variation. There are few precedents to this kind of investigation, but it might be useful to review what we know about genetic drift.

Genetic drift is the change of a gene variant, or allele, due to random sampling. The influence of drift on genome complexity is still under debate. Some authors have proposed that the process of drift leads to larger genome sizes (Lynch and Conery 2003; Lynch *et al.* 2011) whereas others conclude that genome size evolution is unlikely to be explained by population size only (Whitney and Garland 2010; Whitney *et al.* 2011) and even report an inverse correlation between drift and genome size in bacteria (Kuo *et al.* 2009). This discussion is relevant for AMF, since the presence of huge amounts of variation in the cytoplasm, albeit organized in separate nuclei, can be interpreted as a very large genome size. Genome size is hard to estimate in AMF. Because of its heterokaryotic state, the best estimate remains an average over multiple nuclei. However, the latest estimates of genome size in AMF are around 150 Mb (Martin *et al.* 2008; Sędzielewska *et al.* 2011), which is much larger than the previously reported 15Mb (Hijri and Sanders 2004). The presence of many repetitive elements in the *Glomus* genome (Yves Terrat, personal communication) further points to the possibility of a large genome size, since the two are correlated (Lynch and Conery 2003; but see Daubin and Moran 2004). If drift and large population sizes are correlated, then the size of the *Glomus* genome might give us information about the extent of drift that occurs in hyphal systems and at sporulation.

The most important part of the above argument is that genetic drift is inversely correlated to effective population size, and its effects thus become more important in small populations (Buri 1956; Kimura 1985; Ohta 1992). Effective population size, as defined by Sewall Wright, is "the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration" (Wright 1931, 1938). However, what is the effective population size of AMF? What is the number of breeding individuals? What is the individual in AMF? If it is the nucleus, then AMF effective population size may be enormous, with thousands to millions of nuclei per hyphal system and AMF hyphal systems in every ecosystem on the planet (Wang and Qiu 2006; Kivlin *et al.* 2011). If the breeding individual is the hyphal system, then there may be far fewer individuals in a population. And if the breeding individual is all hyphal systems that could potentially anastomose, there might be even fewer. This question will be explored in more detail in section 5.4.

1.2.3. How is genetic variation distributed between AMF genomes?

Single copy genes

Within a single AMF isolate, there are three possible ways for genetic variation to be distributed. These are detailed in figure 1-3. Imagine we study a particular locus in *Glomus irregulare* and find three alleles (represented by the colors red, blue and green). One possibility is that there is only one copy of the locus under study in the genome and alleles are partitioned between nuclei (figure 1-3a). A second option is that the locus occurs in multiple copies (paralogues) in the genome (figure 1-3b) and a third is that the entire genome is duplicated and the

fungus is polyploid (figure 1-3c). This latter possibility can be excluded at least for *G. irregulare*, which was shown to be haploid by Hijri and Sanders (2004) (but see section 5.2.1 for a discussion of the difficulties of applying molecular biology techniques to AMF). In short, if we do not know the copy number of the locus under study, we cannot distinguish intra- from inter-genome variation in AMF.

This difficulty in distinguishing genetic variation between and within genomes is a major hindrance to studying the evolution of the multigenomic structure of AMF. In order to study a particular locus, we need to study orthologous sequences, i.e. sequences that share the same evolutionary history. However, if a gene is duplicated in the same genome, these different copies (paralogues) do not share the same evolutionary history because they are not subject to the same selection pressures (Koonin 2005).

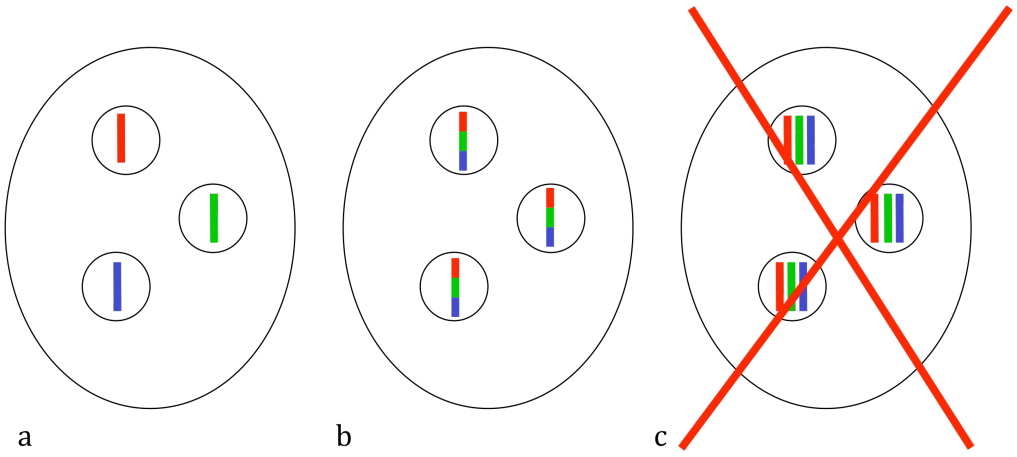


Figure 1-3: What kind of marker is needed? Standard molecular biology techniques do not allow us to distinguish situation **a)**, in which a gene occurs in single copy in genetically divergent nuclei, from situation **b)**, in which a gene occurs in multiple copies within genetically similar nuclei. A third possible scenario **c)**, in which the nuclei are polyploid, has been excluded by Hijri *et al.* (2004): *G. irregulare*, at least, is haploid.

To effectively study the evolution of genetic variation in AMF, we need be sure that we are dealing with orthologues and not with paralogues, because only the former share the same evolutionary history. Separating paralogues from orthologues is not straightforward in AMF because of its multigenomic structure. Unfortunately, we cannot trace the source of a particular allele with standard molecular biology laboratory techniques, without using microscopy (see section 5.2.1). The only way to be sure of the physical location of alleles with respect to each other is to study genes that occur in single copy in the genome.

It is possible to determine copy number for a locus by RT-PCR, even if certain caveats are in place (section 5.2.1). This has already been done for the protein-coding genes *PLS* (Hijri and Sanders 2005) and *BiP* (Corradi *et al.* 2007), and for the ribosomal genes 18S, 5.8S and 25S (Corradi *et al.* 2007). Up to date, there are no polymorphic protein-coding genes known that occur in singe copy in the AMF genome(s). Therefore, it is desirable to develop a set of single copy markers that will allow us to study gene and genome evolution in AMF.

To get an idea of the total genetic variation to be found *between* AMF genomes within an isolate we could exhaustively sequence all alleles for a gene that occurs in single copy in the genome. This would be extremely interesting because it will allow us to take a 'population' approach whilst studying a single isolate. By exhaustively sampling alleles that are cloned and sequenced from the same AMF isolate using the same primers, we can firstly reconstruct phylogenetic relationships between nuclei for this particular marker and get a rough idea of the possible divergence between nuclei of a single isolate. Secondly, it might be possible to study how genetic variation for this particular marker is maintained, by studying the interaction between 'migration' (i.e. anastomosis resulting in recombination or lateral gene transfer) and genetic drift (i.e. loss of genetic variation whilst passing through the spore stage) between isolates. Finally, using

molecular evolution tools, it will be possible to study the footprint of selection and/or demography on the sequences in our dataset.

Obtaining single copy markers will, moreover, allow us to employ another interesting approach. Multi-locus sequence typing (MLST) was initially proposed for the characterization of pathogenic microorganisms (Maiden *et al.* 1998) and entails sequencing internal fragments of 5-7 conserved house-keeping genes (~500 bp). Different sequences present within a bacterial species, or in our case, within an AMF isolate, are assigned as distinct alleles and, for each isolate, the alleles at each of the loci define the allelic profile or sequence type (ST). For example, an average of 30 alleles per locus, for seven loci, allows about 20 billion genotypes to be resolved. Importantly, MLST has also been used to study linkage disequilibrium, recombination and population structure in bacteria (Feil *et al.* 1999; Feil and Spratt 2001) and archaea (Papke *et al.* 2004) and has also been applied to eukaryotes (Bougnoux *et al.* 2002). An important constraint for our application of this method is that we need to be sure that the housekeeping genes in question are single-copy within the AMF genome (within one nucleus) so as to make sure our comparisons are legitimate. Therefore, using MLST to study AMF genomes fits seamlessly into the project to obtain a single-copy marker.

Genome simulations

Single copy genes are a way to investigate allele diversity between nuclei, which allows us to make evolutionary inferences about these genes. However, they cannot inform us on the extent of genome-wide diversity in AMF. For this, we need to use techniques that provide genome-wide information. Massive parallel sequencing techniques such as Solexa (Illumina), SOLiD (Applied Biosystems) and a variety of other sequencing approaches have become increasingly inexpensive and provide enormous amounts of genome-wide data. However, the drawback of these techniques is that read length is rather short. This is problematic for a genome with

many repeats, such as *Glomus irregulare*. Moreover, the high amounts of polymorphism between differentiated genomes in an AMF isolate can be potentially problematic. At the moment, the only technique that provides reads of sufficient length for successful alignment of AMF reads in contigs and scaffolds is 454 pyrosequencing. Pyrosequencing is based on detecting the activity of DNA polymerase with a chemiluminescent enzyme. The technique can be used to nonspecifically sequence entire genomes, which is called shotgun sequencing, or specific sequences, which is called amplicon sequencing (Margulies *et al.* 2005).

Nucleotide variation in a pyrosequencing run is sampled randomly, and this allows us to exploit sequence information and compare genome wide patterns of polymorphism. Nucleotide diversity patterns can be compared to simulated pyrosequencing runs from fungal genomes with comparable genome size, so as to investigate the unique features of AMF genome organization. When combined with a network-based clustering approach (Halary *et al.* 2009), it becomes possible to compare the distribution of cluster size, clustering coefficient and sequence similarity between selected fungal genomes. This approach was developed during my thesis research in collaboration with Dr. Sebastien Halary and Dr. Eric Bapteste, and is unique in offering a possibility to study genome wide patterns of diversity for non-model organisms for which no genome sequence is (yet) available.

1.3. Pyrosequencing

1.3.1. An overview of the technology

Pyrosequencing is based on the detection of polymerase activity by means of an enzymatic luminometric inorganic pyrophosphate (PPi) detection assay (ELIDA) (Nyren 1987). The PPi that is formed in the DNA polymerase reaction is converted to ATP, and this ATP production is monitored by the enzyme luciferase

from the firefly *Photinus pyralis* (Ronaghi *et al.* 1996). The incorporation of nucleotides by the polymerase during DNA synthesis can thus be monitored in real-time by fluorescent signals that correspond to different nucleotides. A modification of the technique that involved adding a nucleotide-degrading enzyme greatly improved the efficiency of the reaction (Ronaghi *et al.* 1998). Ever since, the technology has become widespread and plays an important role in the pantheon of massive parallel sequencing technologies that heralded the era of 'post-genomics'.

An important advantage of pyrosequencing is that the read lengths that are produced (approximately 450 bp for the GS FLX Titanium XLR70 system, according to the 454.com website in March 2012) are much longer than those of other sequencing technologies such as Illumina (approximately 100 bp for the HiSeq system, according to the Illumina.com website in March 2012). There are presently many other massive parallel sequencing technologies on the market, but only of them matches the read length of the GS FLX Titanium system. This technology is the SMRT system by Pacific Biosciences, which was introduced in 2010 and is a single molecule technique that can produce very long read lengths (approximately 3000 bp for the PacBio RS system, according to the pacificbiosciences.com website in March 2012).

Read length is a very important criterion for choosing the best massive parallel sequencing technology for AMF, for several reasons. First, we have no genome assembly for AMF, so that we cannot fit short read length to an already existing template. Most of our sequencing is, in effect, *de novo*. Second, because of the high genetic diversity in the AMF cytoplasm, it is very hard to assemble read sets. The longer the read, the easier it becomes to assemble the data. Third, because genetic diversity is not only found *within* but also *between* nuclei in the cytoplasm, read length is important to recognize differentiated versions of the same locus.

Nevertheless, read length needs to be balanced by accuracy. The GS GLX system suffers from the drawback that it does not deal well with homopolymers, which are repetitions of the same nucleotide. Moreover, sequence quality declines with read length, so that often in a 450 bp read only the first 250 bp really satisfies quality control criteria (see Huse *et al.* 2007; Gilles *et al.* 2011 for a more in-depth assesment of accuracy and quality of pyrosequencing technologies). In spite of these drawbacks, our reason for choosing the GS FLX system was that, at that moment, it was the only system that delivered sufficient read quantity and length to answer our research questions on genetic diversity in AMF.

1.3.2. Pyrosequencing simulations

Pyrosequencing runs deliver a wealth of data, especially for organisms for which relatively little sequence information is available. However, annotation of these runs is difficult in AMF and often only roughly 5 percent of pyrosequencing reads or contigs can be annotated. This makes it difficult to compare AMF genome diversity to other published genomes. To avoid this problem, we decided to develop an annotation-free comparative method that involves producing simulated pyrosequencing datasets that are based on the same parameters as the AMF pyrosequencing dataset, but are generated from different fungal genomes. To generate these simulated datasets, we used the program MetaSim (Richter *et al.* 2008).

How does MetaSim work? The input is simply the sequences of the DNA that one would submit to a pyrosequencing reaction. In our case, we downloaded the complete genome sequences of six fungi that had genome sizes in the upper and lower limit of the estimated genome size for *G. irregulare* DAOM197198 (15-150Mb). These genomes are detailed in Table 4-1. MetaSim generates a reads distribution from these genomes. This reads distribution is particular for every

sequence technology and is determined by a number of variables (Table 4-5). For pyrosequencing, the mean and standard deviation of this reads distribution are determined by the number of 'flows', according to the formula: flows=3*(mean+(8*stddev) (Felix Ott, personal communication). Subsequently, the mean and standard deviation of the reads distribution determine the number of 'flow cycles' according to the formula: flow cycles=(3*(mean+(8*stddev))/4. It is not clear from the program documentation whether the terms 'flow' and 'flow cycle' correspond to the flows and cycles of reagents we find in a 'real' pyrosequencing run.

During the simulation, MetaSim incorporates 'errors' in the reads that are typical for the kind of sequence technology chosen. This allows us to study AMF genetic diversity and at the same time to correct for the errors that are introduced during the pyrosequencing run, since these errors will be found in all reads distributions, both real and simulated.

The output of the program is a reads distribution that in reads amount and length distribution closely matches the original AMF pyrosequencing run. Subsequent analyses can then compare the patterns of diversity in the simulated fungal reads sets to those observed in AMF pyrosequencing data.

1.3.3. Pyrosequencing Analysis

All pyrosequencing analyses were performed in Mothur v.1.22 (Schloss *et al.* 2009), unless specified otherwise. Because of the high error rate associated with pyrosequencing technologies (see section 1.3.1), stringent checks need to be performed on the raw pyrosequencing data to remove low-quality reads and minimize sequencing errors which could have been introduced during the pyrosequencing process (Huse *et al.* 2007; Schloss *et al.* 2011). Reads to be eliminated include those that (i) do not perfectly match the adaptor and primer

sequences, (ii) have ambiguous bases, (iii) have a quality score below an average of 35 in a window of 50 bp, (iv) contain homopolymer lengths greater than 8 bp. Reads that pass quality control and differ by 1 bp between sequences were preclustered following Huse *et al.* (2010).

Chimeric molecules can form *in vitro* during PCR (Wang and Wang 1997) or pyrosequencing (Haas *et al.* 2011). To control for this artifactual genetic variation, we removed chimeric reads that did not match a database of previously obtained sequences with less than 90% bootstrap support using the program Chimeraslayer (Haas *et al.* 2011), as implemented in Mothur v.122. Chimeraslayer searches the ends of query sequences against a database of reference sequences. Candidate parent sequences of a chimera are selected if they form a best-scoring alignment with the query sequence. If possible, the alignment of the query sequence is improved using information from the candidate parent sequences. Chimeras that show greater sequence homology to a chimera formed *in silico* than between either of the parent sequences is flagged as a suspect sequencing artifact. As follows from this description, the role of the reference database is very important for the successful detection of sequencing artefacts.

The remaining clusters are henceforth referred to as ‘alleles’. We performed rarefaction analyses for the dataset. A clear asymptote in the rarefaction analysis is an indication that sampling was representative of the actual diversity at the locus. We calculate alpha-diversity indices, including the total number of alleles and the Chao1 diversity index (which is a measure of the minimum richness). To explore beta-diversity, we calculated the Jaccard similarity coefficient (jclass) based on the observed richness and the Yue & Clayton (thetaYC) measure of dissimilarity between the structures of two communities. The Yue & Clayton similarity coefficient compares the total number of OTU *i* in communities A and B to the sum of the relative abundance of OTU *i* in community A and the sum of the relative abundance of OTU *i* in community B.

The statistical significance of these latter two coefficients is calculated using the (un)weighted Unifrac distance metric (Lozupone and Knight 2005; Lozupone *et al.* 2006), which uses phylogenetic information to calculate a distance measure between separate samples. To determine whether samples are statistically different from each other, UniFrac employs Monte Carlo simulations to randomize the sample classification of each taxa and create a distribution of UniFrac distance values. With a distribution of UniFrac values, a p-value can be given to the observed distance.

Statically significant differences between samples (and community structures of samples) are determined using the nonparametric analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992; Anderson 2001; Martin 2002), a distance-based version of Bartlett's test for homogeneity of variance (HOMOVA) (Stewart and Excoffier 1996) and with a two-sampled Kolmogorov-Smirnov test.

2. Intra-isolate genome variation in arbuscular mycorrhizal fungi persists in the transcriptome

Eva Boon[1], Erin Zimmerman[1], B. Franz Lang[2] and Mohamed Hijri[1]

[1] Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, 4101 rue Sherbrooke Est, QC, H1X 2B2, Canada

[2] Robert Cedergren Centre for Bioinformatics and Genomics, Département de Biochimie, Université de Montréal, 2900 Boulevard Edouard-Montpetit, Montréal, Québec, H3T 1J4, Canada

This chapter was published in July 2010, in the *Journal of Evolutionary Biology* 23(7), 1519–1527

Data deposition: Nucleotide sequences were deposited in GenBank under accession numbers FJ235536-235574, FJ235605-FJ235623, GU992000-GU992199 and FJ743698-FJ743715.

Author contributions

Eva Boon obtained a part of the LSU rDNA sequences, collected and analyzed the data, and wrote the manuscript. Erin Zimmerman contributed to the *PLS* data. B. Franz Lang advised on the reconstruction of the LSU rDNA secondary structure and contributed to the manuscript. Mohamed Hijri was the principal investigator, was responsible for a part of the LSU rDNA sequences, the *PLS* data and pyrosequencing run, and participated in the design of the study. All authors read and approved the final manuscript.

2.1. Abstract

Arbuscular mycorrhizal fungi (AMF) are heterokaryotes with an unusual genetic makeup. Substantial genetic variation occurs among nuclei within a single mycelium or isolate. AMF reproduce through spores that contain varying fractions of this heterogeneous population of nuclei. It is not clear whether this genetic variation on the genome level actually contributes to the AMF phenotype. To investigate the extent to which polymorphisms in nuclear genes are transcribed, we analyzed the intra-isolate genomic and cDNA sequence variation of two genes, the large subunit ribosomal RNA (LSU rDNA) of *Glomus* sp. DAOM-197198 (previously known as *G. intraradices*), and the *POL1*-like sequence (*PLS*) of *Glomus etunicatum*. For both genes, we find high sequence variation at the genome and transcriptome level. Reconstructions of LSU rDNA secondary structure show that all variants are functional. Patterns of *PLS* sequence polymorphism indicate that there is one functional gene copy, *PLS*2, which is preferentially transcribed, and one gene copy, *PLS*1, which is a pseudogene. This is the first study that investigates AMF intra-isolate variation at the transcriptome level. In conclusion, it is possible that in AMF multiple nuclear genomes contribute to a single phenotype.

2.2. Introduction

Arbuscular mycorrhizal fungi are obligate symbionts of plant roots that improve host nutrient uptake and pathogen resistance (Smith and Read 2008). Plant growth response and productivity varies significantly with the composition of AMF communities (Van der Heijden *et al.* 1998), and the effects of AMF on plant performance are studied intensively. It is estimated that 60-80% of all land plants associate with AMF (Wang and Qiu 2006; Smith and Read 2008). The earliest fossils that resemble AMF spores are ~460 million years old, and various authors suggest that AMF may have played an essential role in the establishment of early land plants (Simon *et al.* 1993; Remy *et al.* 1994; Redecker *et al.* 2000).

Despite the obvious importance of AMF to plant functioning and ongoing efforts to understand this symbiosis, the magnitude, organization and expression of genetic variation in AMF is poorly understood. The study of genetic variation in these organisms is unusually complex due to their unique lifestyle and genome structure. First, from a technical perspective, AMF are difficult to handle because they are obligate biotrophs that are only cultivable in the presence of a host plant. It is therefore challenging to obtain sufficient amounts of contaminant-free nucleic acids for genetic analysis (Hijri *et al.* 2002). An *in vitro* culture system using transformed carrot roots has been developed to alleviate these problems (St.-Arnaud *et al.* 1996). Using this system, tissue cultures (or 'isolates') are started from single spores and can be maintained under sterile conditions. Second, the study of AMF genetics is conceptually challenging because of the large amount of genetic variation occurring within a single mycelium. AMF are coenocytic; that is, there are no septa between cells, and organelles freely move through the hyphal network. The partition of this genetic variation between nuclei has been a subject of controversy, featuring two competing hypotheses.

The first hypothesis states that genetic variation is contained within genetically identical polyploid nuclei (Pawlowska and Taylor 2004; Pawlowska 2005; Pawlowska and Taylor 2005). The second hypothesis describes nuclei as being haploid and heterokaryotic (Kuhn *et al.* 2001; Hijri and Sanders 2005). Here, heterokaryosis refers to the co-existence of many genetically differentiated nuclei within the same cytoplasm. The evidence in favor of the first hypothesis is contestable (Bever and Wang 2005) and recent studies favor the latter hypothesis (Corradi *et al.* 2007; Angelard *et al.* 2010; Cárdenas-Flores *et al.* 2010; Tisserant *et al.* 2012). In the present study, we assume the hypothesis of heterokaryosis.

In multicellular organisms, it is generally assumed that all cells share the same (nuclear) genome. Typically, this results from the transmission of a single progenitor genome, in the form of a spore or gamete, from parent to offspring. However, in organisms that do not pass through a single-genome stage, intra-organismal variation is inevitable and may even play an important role in clonal propagation (see Pineda-Krch and Lehtila 2004b, for a review). For example, heteroplasmy has been reported in both mitochondria (Rand 2001) and chloroplasts (Petit 2007), because cytoplasmic genomes do not necessarily pass through a single-copy stage in their life cycle. The potential for intra-organismal variation is closely linked to the magnitude of the genetic bottleneck an organism experiences during its life cycle. To our knowledge, arbuscular mycorrhizal fungi (AMF) are the only organisms which are multi-nucleate at all life stages, never passing through a uninuclear phase (Marleau *et al.* 2011).

Intra-isolate variation in AMF has been reported for rDNA (Sanders *et al.* 1995; Lloyd-MacGilp *et al.* 1996; Clapp *et al.* 1999; Hijri *et al.* 1999; Hosny *et al.* 1999; Pringle *et al.* 2000; Clapp *et al.* 2001; Kuhn *et al.* 2001; Rodriguez *et al.* 2004) and for protein-coding genes (Kuhn *et al.* 2001; Helgason *et al.* 2003; Kuhn 2003; Corradi *et al.* 2004b; Hijri and Sanders 2005; Corradi and Sanders 2006; Corradi *et al.* 2007; Boon *et al.* 2010). Even copy number polymorphism between

isolates has been observed (Corradi *et al.* 2007). While both ribosomal and protein-coding genes of AMF are known to possess significant amounts of intra-isolate variation, it is not known whether this intra-isolate variation persists at the transcriptional level. This is an important step towards understanding how a heterokaryotic population of nuclei gives rise to the AMF phenotype.

Why is it possible that the genetic variation we observe among nuclei is not expressed? First, the clonal mode of AMF reproduction could mean that recombination in AMF is a rare event (but see Gandolfi *et al.* 2003; Croll 2009; Croll and Sanders 2009). If recombination is rare, the failure to purge deleterious mutations could lead to the presence of pseudogenes. This process is known as Muller's ratchet (Muller 1964). Second, if multiple gene copies within the AMF mycelium (partitioned either within or between nuclei) behave as genes after a duplication event, they can lose functionality over time (Zhang 2003). For these two reasons, it is possible that the internuclear or intergenomic variation in AMF is largely comprised of pseudogenes, which are not transcribed and do not contribute to the phenotype.

The extent and organization of genetic variation between nuclei within AMF isolates has thus far been studied using two distinct approaches. First, the presence of different alleles, and variation in copy number between nuclei of a single isolate, has been visualized using FISH techniques for ribosomal (ITS, Kuhn *et al.* 2001) and protein-coding loci (*BiP*, Kuhn 2003). Second, by comparing copy number and allele number, it is possible to infer whether variation is partitioned between or within genomes of a given AMF isolate. The protein-coding gene *PLS* was studied using this latter approach (Pawlowska and Taylor 2004; Hijri and Sanders 2005). Two loci that vary between nuclei within the same isolate have been studied in detail: large subunit ribosomal DNA (LSU rDNA) and *PLS*. In this study, we compare the genetic diversity of these latter loci at the genome level with the genetic diversity at the transcriptome level. More specifically, we compare the

degree of within-isolate nucleotide polymorphism in genomic versus complementary DNA (gDNA vs. cDNA), for *Glomus sp.* (DAOM-197198) LSU rDNA and *Glomus etunicatum PLS* genes. If we find less variation on the transcriptome level, it would support the hypothesis that a portion of AMF genetic diversity is not expressed, and raise the possibility that much of AMF intra-isolate diversity consists of pseudogenes. We address two specific questions: (1) are all gene variants transcribed? (2) is there any evidence that alleles are under relaxed selective pressure?

2.3. Materials and Methods

2.3.1. DNA extraction and cDNA library construction

DNA and RNA were extracted with the Qiagen Plant DNA extraction kit (Qiagen, Canada) and RNAqueous®-Micro (Ambion, Canada), from freshly harvested spores and hyphae of *Glomus sp.* isolate DAOM-197198, previously known as *G. intraradices* (Stockinger et al. 2009) and *Glomus etunicatum* isolate NPI (Native Plants Incorporated). We did not use the same species for both markers because thus far *PLS* alleles have never been successfully amplified in *Glomus sp*. isolate DAOM-197198. cDNA synthesis was performed using RevertAid™ First Strand cDNA Synthesis Kits (Fermentas, Canada) and tested for gDNA contamination by comparing PCR fragment sizes of the rDNA intergenic spacer (IGS) and of (intron-containing) P-Type IID ATPases (Corradi and Sanders 2006). No gDNA contamination was detected in the cDNA libraries.

2.3.2. DNA amplification

The entire LSU rDNA was amplified from genomic DNA with the forward primer F1.LSU and the reverse primer R1.LSU (see table 2-1 for primer sequences), yielding fragments of 3056-3091 bp. A ~200 bp region was selected for further investigation and amplified from genomic DNA and cDNA using the same forward primer F1.LSU and the reverse primer R1.LSU.cDNA (see table 2-1). This section covers a region that contains both highly variable and conserved sections, and which corresponds to structure **a** in figure 1(a) in Schnare et al. (1996).

G. etunicatum cDNA was amplified by PCR using the primers Pol4 and Pol7 (Pawlowska and Taylor 2004). High fidelity *Pfu* DNA polymerase (Fermentas, Canada) was used for all PCR reactions, under standard conditions, with 35 cycles. Cloning was performed with the TOPO® TA Cloning Kit (Invitrogen, USA), and sequencing with universal primers was carried out at the Genome Québec Innovation Centre sequencing platform (Montréal, Québec).

Table 2-1 Primers used for LSU rDNA and *PLS* amplification

Primer name	Sense	5'-3' sequence
Pol4-A	Forward	GAATCCTTCCCAAATTGATCAGAATACTTGTT
Titanium adaptor sequence		CCATCTCATCCCTGCGTGTCTCCGACTCAG
Pol7-B	Reverse	TAATAATAAAAGCCTTTCAAAAAATCCATCAATA
Titanium adaptor sequence		CCTATCCCCTGTGTGCCTTGGCAGTCTCAG
F1.LSU	Forward	GCATATCAATAAGCGGAGGA
R1.LSU	Reverse	CGGTCTAAACCCAGCTCACG
R1.LSU.cDNA	Reverse	CTAATAGGGAACGTGAGCT

2.3.3. Testing for in vitro recombination and polymerase errors

If PCR is performed on a mixed template, *in vitro* recombination can bias polymorphism estimates. This is especially relevant when using *Pfu* polymerases and a higher number of PCR cycles (between 20-40) (Zylstra *et al.* 1998). We are aware that the heterokaryotic state of the AMF under study would provide a mixed template for the PCR reaction, and we thus evaluated the role of *in vitro* recombination on polymorphism in our dataset. We performed a PCR on a mixed template of equivalent amounts of two plasmids containing highly divergent *PLS* sequences (belonging to the groups *PLS*1 and *PLS*2), which served as our reference sequences. PCR was performed under the same conditions as those used for our main *PLS* dataset (gDNA and cDNA). We subsequently cloned and sequenced the PCR products, and compared the results with the initial two reference sequences. We recovered six *PLS*1 sequences, and eight *PLS*2 sequences. No recombination events and no polymerase errors were observed.

2.3.4. *PLS* pyrosequencing

In an attempt to exhaustively sample *PLS* allelic diversity, we sequenced single *G. etunicatum* spores using 454 technology. A single spore was picked, placed in 1 µl sterile water and crushed on the bottom of 0.2 ml tubes using a Pasteur pipette on which the tip had been melted into a ball. The entire genome of this spore was amplified using the GenomiPhi Whole-Genome Amplification Kit (GE Healthcare, Amersham, United Kingdom) according to the manufacturer's instructions. *PLS* variants were amplified using Dream*Taq* DNA polymerase (Fermentas, Burlington, Ontario) using primers Pol4 and Pol7, with added Titanium adaptor sequences for pyrosequencing. The reaction was performed in 50 µl volumes containing 0.2 mM dNTPs, 0.5 µmol of each primer and the PCR buffer. PCR was carried out for 34 cycles (95°C for 30 sec, 50°C for 30 sec, 72°C for 1

min; preceded by an initial 2 min denaturation at 90°C and followed by a 10 min hold at 72°C) on a Mastercycler EPgradient S (Eppendorf, Mississauga, Ontario). The PCR product was loaded on an electrophoresis gel to ensure successful amplification of the gene and the bands were cut from the gel and purified by 24 h incubation in 50 µl milliQ water. These purified samples were then sent to the Genome Québec Innovation Centre in Montréal for pyrosequencing using the GS FLX Titanium emPCR kit (Lib-A), with unidirectional reads (1/8 run per sample).

2.3.5. Sequence alignment and data analyses

Sequence alignments for both the LSU rDNA and *PLS* loci were performed with MAFFT (Schatz 2007) in the Jalview interface (Steinke *et al.* 2005), and then verified and refined by eye using Bioedit v7.0.5 (Hall 1999). For each sequence, indel composition was registered before excluding alignment gaps from further statistical analysis.

We used DNAsp v4.50.3 (Rozas *et al.* 2003) for polymorphism and genetic differentiation analyses, and calculated a Principal Component Analysis matrix in Jalview. Neighbor-Joining trees were inferred using MEGA v.4.0 (Tamura *et al.* 2007). A Neighbor-Net network (uncorrected p-distance) was constructed for the *PLS* locus using SplitsTree4 (Huson and Bryant 2006), testing statistical significance with bootstrap analyses (1000 replicates).

rDNA secondary structure was inferred by aligning LSU sequences of *Glomus sp.* (isolate DAOM-197198), with the secondary structures of *Tricholoma matsutake* (Hwang and Kim 2000) and *Saccharomyces cerevisiae* (U53879) from the Comparative RNA Web website (http://www.rna.ccbb.utexas.edu) as references. Pairings were identified both by sequence comparison at the primary sequence level and by searching for compensatory base pair mutations with CBCanalyzer (Falkenberg *et al.* 2007). Compensatory base pair mutations

preserve nucleotide interactions, and thus the secondary structure of the ribosome that is formed by these nucleotide interactions. For example, if an 'A' mutates into a 'G', then the 'U' that originally paired with the 'A' will change into a 'C' to preserve base pairing and thus secondary structure.

We investigated substitution patterns in the context of the stem-loop structure of both the complete 2053 bp LSU rDNA alignment and a 220 bp subset. We divided the LSU rRNA gene of *Glomus* sp. isolate DAOM 197198 into functional regions, following the description of the eukaryote LSU rRNA gene in Schnare et al. (1996) (see also table 2-4). To further examine the effect of nucleotide substitutions on the conservation of secondary structure, we calculated the amount of compensatory base changes (CBC) in these functional regions, using additional sequences from other isolates (additional sequences listed in table 2-5).

2.4. Results

2.4.1. Sequences obtained

For the LSU rRNA gene, we obtained 14 sequences from gDNA covering most of the gene's 5' region (2053 bp, accession numbers FJ235561-FJ235574). For the 220 bp subset of this region, we recovered 33 sequences from cDNA, (accession numbers FJ235536-FJ235560 and FJ743698- FJ743705) and 24 from gDNA (FJ235561- FJ235574 and FJ743706- FJ743715). Additional sequences were downloaded from GenBank; 18 from gDNA and 22 ESTs (table 2-6). One contig (gDNA) was obtained from the *Glomus* Genome Consortium database (contig 95064 of the October 16, 2008 assembly) (Martin *et al.* 2008). All sequences originated from *Glomus* sp. isolate DAOM 197198 and, once aligned,

resulted in two alignments; a 2053 bp alignment with 14 sequences, and a subset covering 220 bp with 114 sequences. A secondary structure was estimated covering the entire 5' end of the large subunit (available upon request).

For the *PLS* gene, we obtained 30 gDNA sequences by cloning (accession numbers FJ235575-FJ235604) and 19 cDNA sequences (accession numbers FJ235605-FJ235623). Pyrosequencing yielded 1250 sequences, of which the minimum length was 39 bp and the maximum length was 454 bp, with an average length of 130 bp. Of these, 200 sequences longer than 200 bp were retained (NCBI accession numbers GU992000-GU992199). New *PLS* sequences were pooled with 15 previously published sequences (accession numbers AY330581, AY394011–AY394024. All sequences came from *G. etunicatum* isolate NPI (Plantworks, Kent).

2.4.2. LSU rDNA polymorphism and genetic differentiation

Haplotype diversity H_d and pairwise nucleotide diversity π (Nei 1987) for the LSU rDNA are summarized in table 2-2. Both estimators are measures of sequence polymorphism. Pairwise nucleotide diversity π was also calculated along a sliding window for the 2053 bp alignment (figure 2-1).

Genetic differentiation was estimated using several statistical approaches. We used the nearest-neighbour statistic (S_{nn} statistic, Hudson 2000), which is a sequence-based estimator of particular use for datasets with high nucleotide diversity. The choice of a non-parametric measure of population differentiation such as S_{nn} is justified, since, in the case of AMF, we have very little information about the way genetic variation is generated and maintained. To obtain a second and independent estimate, genetic differentiation was also estimated using a Principal Component Analysis (results not shown). For either method, no genetic differentiation was detected within or between the 2053 bp and 220 bp alignments.

Table 2-2 LSU rDNA polymorphism estimates

Alignment length	# sequences	# alleles	H_d*	π†
2053 bp				
genomic	14	12	0.967	0.01872
220 bp	114	72	0.977	0.05362
genomic	43	37		
cDNA	71	60		

* haplotype diversity (Nei 1987)

† nucleotide diversity; average number of nucleotide differences per site between two sequences (Nei 1987)

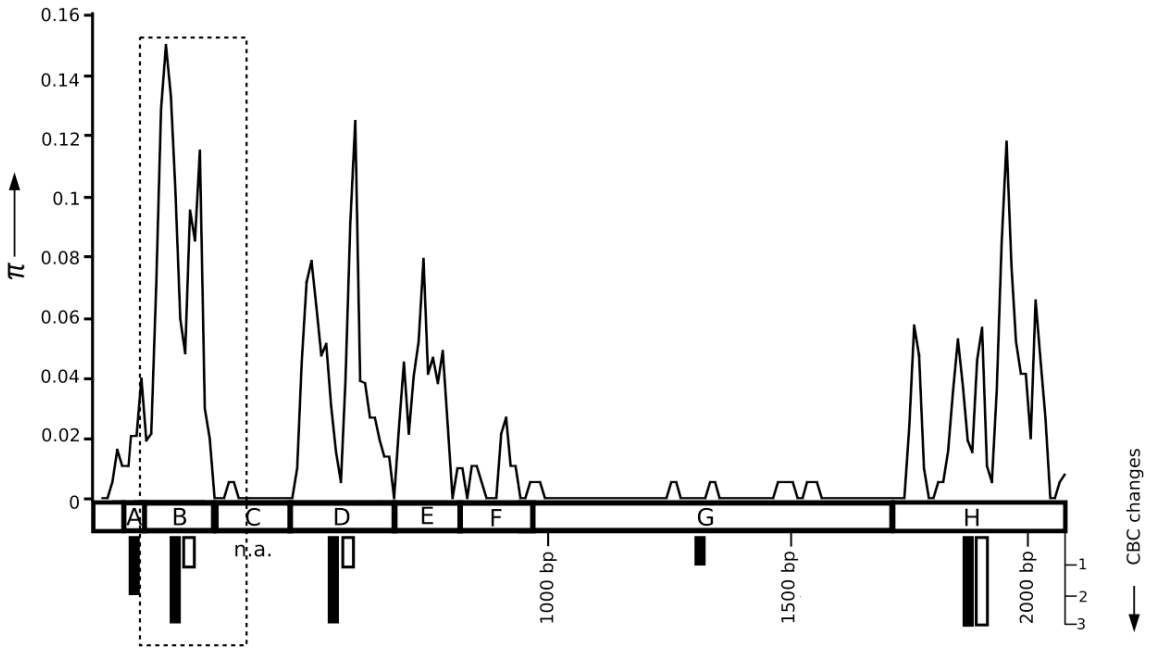


Figure 2-1 Within-isolate nucleotide diversity (π) (Nei 1987) along the 5' LSU rDNA of *Glomus irregulare* for the entire 2053 bp LSU rDNA, with a comparison of compensatory base pair changes within *Glomus sp.* isolate DAOM-197198 (white bars) and between isolate DAOM-197198 and other *G. irregulare* isolates (black bars). The dotted rectangle indicates the relative location of the 220 bp subset alignment.

2.4.3. *PLS* nucleotide polymorphism and population differentiation

By comparing cDNA and genomic sequences from *G. etunicatum*, we found that the *PLS* gene contains two introns. Our data was complemented with additional sequences from a previously published dataset (Pawlowska and Taylor 2004), and we observed no genetic differentiation between the two datasets.

At the *PLS* locus, 65 alleles were retrieved from 263 sequences. These alleles fall into two genetically distinct groups (S_{nn}=1.000, p-value 0.000) that correspond to two previously described gene variants, *PLS*1 and *PLS*2 (Pawlowska and Taylor 2004; Hijri and Sanders 2005). Patterns of sequence variation and pairwise nucleotide diversity (π) differ markedly between *PLS*1 and *PLS*2 sequences. For *PLS*2, areas of high genetic diversity are limited primarily to introns, whereas in the *PLS*1 group, exons show extensive polymorphism as well (table 2-3). The *PLS*1 group is more genetically diverse than the *PLS*2 group (π is 0.02802 and 0.01044, respectively). *PLS*1 and *PLS*2 allele frequencies change drastically between sequences that originate from gDNA and cDNA. For *PLS*1, we found 62 alleles in the gDNA and 2 in the cDNA, and for *PLS*2, 4 and 3 alleles, respectively (excluding gaps). Finally, 101 out of 217 *PLS*1 gDNA alleles have a deletion in the 1st exon, representing 62 unique events leading to a frameshift mutation. In 2 *PLS*1 alleles, we found a stop codon in the first exon. These observations are summarized in table 3 and in the Neighbor-Net network of Fig. 2-2.

Table 2-3 Polymorphism estimates for *PLS* (*G. etunicatum*)

		n *	# alleles	†H_d	π °	Frameshift mutations	Stopcodon
*PLS*1							
genomic	all	217	60	0.948	0.01987		
	exon				0.01484	101[1]	2
	intron				0.02738		
cDNA		3	2	0.667	0.02083	3[2]	
*PLS*2							
genomic	all	27	6	0.649	0.01101	0	
	exon				0		
	intron				0.01839		
cDNA		16	3	0.242	0.00569	0	

[1]at least 65 independent mutational events

[2]three frameshift mutations in same sequence; reading frame not disturbed

* n : number of sequences

† H_d : haplotype diversity

° π : pairwise nucleotide diversity (Nei 1987)

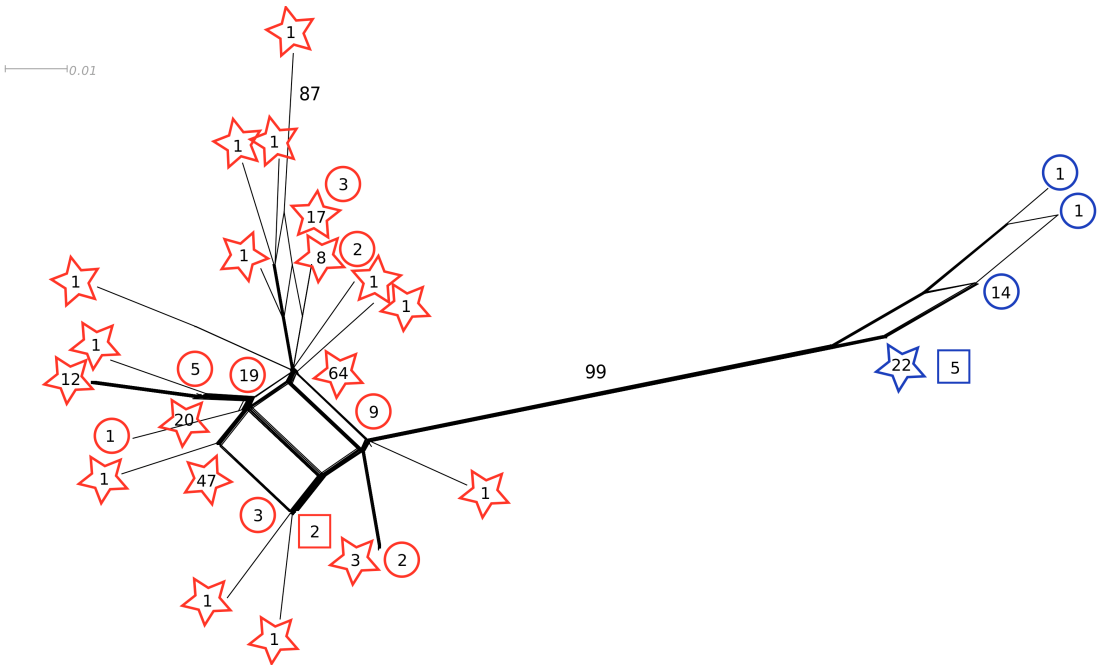


Figure 2-2 A Neighbor-Net network for the *PLS* locus, based on uncorrected P-distances. Only exons were taken into account. Reticulation of the branches indicates uncertainties due to homoplasy (recombination or multiple hits). Sequences that belong to the *PLS*1 sequence type are represented in red and sequences that belong to the *PLS*2 sequence type are represented in blue. The shape of the symbol corresponds to the mode of sequencing and/or origin of the template; squares are sequences from the cDNA library, circles are gDNA sequences obtained by cloning and stars are gDNA sequences obtained by pyrosequencing. Numbers in the symbols correspond to the frequency of the allele in question. Bootstrap support values >70 are depicted above branches (1000 replicates). Scale bar indicates substitutions per site. The shading on the edges of the network reflects confidence values associated with each split

2.4.4.LSU rDNA secondary structure

Ribosomal regions A, B, D and H, which are known to be variable and/or dispensable in other species (details listed in table 2-4), also show high genetic variation in our alignment (figure 2-1). In general, genetic diversity was higher in regions that formed loops, and lower in stem regions. For regions C and F, no complementary base pair changes (CBCs) were calculated, because region C interacts with the small subunit of the ribosomal DNA for which pairing sequences were unavailable, and region F appears to be a Glomeromycota-specific insertion. No CBC changes were observed for region E.

2.5. Discussion

2.5.1.High genetic variation persists at the transcript level

The analyses of both LSU rDNA and *PLS* sequences confirm high intra-isolate genetic polymorphism at the genome level, as shown previously for both the LSU rDNA (Rodriguez *et al.* 2004) and *PLS* (Pawlowska and Taylor 2004). The present study is the first to show that genetic variation persists at the transcript level (table 2-2 and 2-3).

The high levels of intra-isolate sequence polymorphism may raise concerns about the frequency of PCR errors, which could artificially increase sequence variability. Nevertheless, the reported error rate (2.6×10^{-6} errors per nucleotide per cycle) for *Pfu* polymerase (Lundberg *et al.* 1991) is much lower than the observed number of nucleotide differences per nucleotide per sequence (π) (Nei 1987), which was 0.01872 for the ribosomal gene in the *Glomus* sp. isolate DAOM-197198, and 0.02802 and 0.01044 for *PLS*1 and *PLS*2 in *G. etunicatum,* respectively. Our estimates of nucleotide diversity are quite conservative,

considering that the PCR approach with conserved primers we used will retrieve only the most abundant alleles in the isolates under study. Moreover, the sequences retrieved from GenBank were not genetically distinct from our own sequences, suggesting that our dataset is unbiased. Finally, specifically designed tests did not detect *in vitro* recombination or *Pfu* error (see Materials and Methods). Considering these arguments, we believe that the genetic diversity we found within AMF isolates is not an artifact. To investigate whether the variant alleles evolve under selection, we explored sequence conservation and patterns of polymorphism for the two loci.

2.5.2. LSU rDNA sequence conservation: a population of sequences under selective constraints

Conserved regions in the *Glomus* sp. LSU rDNA nucleotide sequence correspond well to regions of structural conservation known for *S. cerevisiae* (Schnare *et al.* 1996) (figure 2-1). Also, regions B and H, which were experimentally shown to be dispensable in yeast (Musters *et al.* 1989), show high levels of variation in our *Glomus* sp. isolate DAOM 197198. The number of complementary base pair changes (CBC) follows this pattern, since regions B and H have up to three CBCs, whereas the structurally important region G, which contains the GTP-associated centre of the large subunit, only contains a single CBC (figure 2-1). The same pattern of conserved and variable regions is observed for different isolates (see table 2-5 for accession numbers).

Moreover, it is interesting to note that the number of CBC changes is higher *within* isolates than *between* isolates. This may indicate a more restrained gene flow between different isolates due to differences in anastomosis (hyphal fusion) frequency between different AMF isolates and species (Giovannetti *et al.* 1999; Giovannetti *et al.* 2001; Croll 2009; Lake 2009).

We found 114 LSU rDNA alleles in the isolate DAOM-197198 (table 2-2). To investigate how this variation is partitioned within the cytoplasm, we estimated copy number per genome using absolute quantitative real-time PCR (see section 2.7.1 of the supplementary materials). LSU rDNA copy number is 31.27 ± 0.25 SE, which means that the high number of alleles for LSU rDNA cannot be explained exclusively by polyploidy, as is also the case for the *PLS* loci (Hijri and Sanders 2005). This observation lends further support to the hypothesis that nuclei in AMF contain genetically divergent genomes (Kuhn *et al.* 2001; Hijri and Sanders 2005).

We are aware that the aforementioned copy number per genome is merely an average over the entire nuclear population within this isolate. It is theoretically possible that copy number per nucleus varies widely, and that most variant alleles are still contained within the same nucleus. However, considering that (i) the standard error for our copy number estimate is very low, and (ii) concerted evolution of rDNA genes will tend to homogenize gene copies within a genome (Smith 1976; Arnheim 1983), it is still unlikely that polyploidy alone can explain our observations. Still, only visualization techniques such as FISH (Fluorescence *In Situ* Hybridization) can conclusively show how genetic variation is partitioned between nuclei. FISH has indeed been performed on two AMF protein-coding genes, demonstrating intra-isolate genetic variation for these two loci (Kuhn *et al.* 2001; Kuhn 2003).

2.5.3. *PLS* groups show contrasting patterns of diversity

We confirm the presence of two genetically divergent groups of *PLS* alleles, *PLS*1 and *PLS*2, that were reported previously in *G. etunicatum* (Pawlowska and Taylor 2004; Hijri and Sanders 2005). Allele frequencies are markedly different between *PLS*1 and *PLS*2, both in the genome and in the transcriptome (table 2-3 and figure 2-2). We are confident that this difference in allele frequencies between

*PLS*1 and *PLS*2 reflects the genetic variation in the genome and is not due to sampling bias or technical artifacts, since we used two different approaches to obtain our sequences. These approaches include conventional cloning and subsequent Sanger sequencing, as well as massive parallel pyrosequencing using Titanium technology. Both datasets yielded the same skewed allele frequencies.

We propose that the two *PLS* alleles arose by gene duplication, and that the resulting paralogues underwent different evolutionary trajectories, based on the following evidence. First, Hijri and Sanders (2005) proposed that *PLS* occurs in two copies, based on RT-PCR experiments. Second, genetic diversity is up to 4 times higher among *PLS*1-type alleles than among *PLS*2-type alleles, indicating relaxed selection pressures for *PLS*1 (table 2-3). Third, all *PLS*2 alleles found in the gDNA were found back in the cDNA, whereas only 3.3% of *PLS*1 alleles were found back in the cDNA libraries. Finally, 101 frameshift mutations and 2 stop codons occurred in the *PLS*1 gDNA alleles. We propose that *PLS*2 represents the functional gene copy, whereas *PLS*1 shows signs of relaxed selection pressures, and may in fact be a pseudogene.

2.6. Concluding remarks

Genetic variation among AMF nuclei within a mycelium has been demonstrated at a genomic level (Hijri and Sanders 2005). This study is the first to show that this variation is transcribed. Our work demonstrates the presence of many LSU rDNA alleles within a single AMF isolate, all of which show structural conservation. We present evidence for a duplication of the protein-coding gene *PLS*, of which one duplicate shows signs of relaxed selective pressures. For *PLS*, genetic variation at the genome level is not necessarily representative of genetic variation in the transcriptome. Our findings are consistent with previous observations of large amounts of genetic variation for protein-coding genes (Hijri

and Sanders 2005), ITS sequences (Hijri *et al.* 1999; Kuhn *et al.* 2001), and non-coding sequences (Hijri *et al.* 2007) within the genome of *Glomus* sp. isolate DAOM197198 (Martin *et al.* 2008). The question remains as to how these unique organisms function with tens to hundreds of divergent alleles within a single mycelium.

2.7. Supplementary Information

Table 2-4 Description of LSU regions (structure information is based on Schnare et al.(1996))

Code	Position in alignment	description
A	0-115	Isolated hairpin of variable length in eukaryotes; artificial extension of the hairpin at this position has no effect on ribosome production or function (Musters *et al.* 1989)
B	116-230	Variable region with in the middle a phylogenetically conserved structure as identified by Schnare *et al.*(1996)
C	231-406	This region pairs with the small subunit (5.8S)
D	407-633	'545 region' This is one of the most highly variable regions in eukaryotic large subunit rRNA, ranging from 8 to 865 nt in length. This region is 102 nt in *M.racemosus*, the closest relative to *G. irregulare* for which a secondary structure has been published (Ji and Orlowski 1990)
E	634-767;	This insert seems unique to *Glomus*; a search in GenBank yielded no significant similarities. Using the software mfold (Zuker 2003) did not detect any helical structures.
F	768-942	'650 region'As in most eukaryotes, this region contains 2 hairpin structures
G	943-1722	non variable region; according to Schnare et al (1996) this section codes for the GTP associated centre of the large subunit and is highly conserved
H	1723-2053	variable end region. It has been suggested that some of the hairpins in this region are dispensable (Musters *et al.* 1989)

Table 2-5 Additional *G. irregulare* (unless specified otherwise) genomic sequences from GenBank used for inter-isolate comparison in the LSU rDNA analysis

Accession Number	isolate	Used in alignment	Geographic origin
gil121488322-gil121488331	DAOM197198	0-1284	Point Rouge, Québec
gil82398345	DAOM197198	97-1215	Point Rouge, Québec
gil121488398-gil121488403	MUCL41833	0-1294	Canary Islands
Gil121488388-gil121488397	MUCL41835	0-1293	Contaminant pot culture
gil82398570	4695rac-11G2	1-2053	Point Rouge, Québec
gil121488334-gil121488335	MUCL43203	0-1282	Quetico, Ontario
gil121488342-gil121488344	MUCL43204	0-1283	Canada
gil121488348		0-1279	Canada
gil121488362	MUCL43206	0-1291	St André Avelin, Québec
gil121488363			
gil121488368		0-1292	St André Avelin, Québec
gil121488370		0-1304	St André Avelin, Québec
gil121488378	MUCL43207	0-1290	Unknown
gil121488349-gil121488350	MUCL43205	0-1278	Québec, Canada
gil121488352		0-1284	Québec, Canada
gil121488354		0-1263	Québec, Canada
gil121488358-gil121488359		0-1295	Québec, Canada
gil121488383	MUCL43196	0-1292	Contaminant pot culture
gil121488386-gil121488387		0-1287	Contaminant pot culture
gil82398394	*G. mosseae*	1-2053	AFTOL-ID 139

Table 2-5 Additional *G. irregulare* LSU sequences from GenBank

Accession number	DNA source
DQ273790	Genomic
AJ854577	Genomic
AJ854574-AJ854583	Genomic
AF396797	Genomic
AF396787	Genomic
AF396792	Genomic
X99640	Genomic
AY138146	Genomic
AJ416418	Genomic
BE603763	EST[a]
BE603772	EST
BE603780	EST
BE603790	EST
BE603801	EST
BE603803	EST
BE603804	EST
BE603810	EST
BE603819	EST
BE603850	EST
BE603862	EST
BE603872	EST
BE603875	EST
BE603882	EST
BE603885	EST
BE603889	EST
BE603904	EST
BE603906	EST
BE603930	EST
BE603945	EST
BE603951	EST
BE603955	EST

2.7.1. Absolute quantification of LSU rDNA copy number

Methods

Quantification of LSU rRNA copy number was performed as described previously (Hijri and Sanders 2004, 2005; Corradi *et al.* 2007). A qPCR assay was developed for a conserved 78 bp fragment of the LSU rDNA, using Primer Express Software Version 2 (Applied Biosystems), with as a forward primer rLSU_conF, 5'-ACGCTCATTTAGATACCACAAAAGGT and a reverse primer rLSU_conR, 5'-TTAGCGGATTCCGACTTCCA. The sequence of the TaqMan probe was 5'-TTAGTTCATCTAGACAGCAGGACGGTGGCC, which was labelled with FAM at the 5' end and BHQ-1 at the 3' end. Fluorescence data was collected using the ABI PRISM 7500 Sequence Detection System (SDS; Applied Biosystems). In each real-time PCR experiment, we employed plasmid DNA containing known amounts of the LSU rDNA as a standard. Two-fold serial dilutions (ranging between 1.37×10^6 and 175.36×10^6 plasmid molecules) of the purified plasmid DNA were included in each experiment to generate the standard curve. Real-time PCR on the plasmid DNA was performed in 3 replicates, with 6 dilutions per replicate. The Ct values were then plotted against the Log of the copy number of the plasmid molecule containing the LSU rDNA insert (figure 2-3). Copy number was established for each sample of target DNA using the plasmid DNA regression line as a standard (Hijri and Sanders 2004, 2005; Corradi *et al.* 2007) and an estimated genome size of 15.74 Mb of DNA per nucleus (Hijri and Sanders 2004). PCR efficiency was calculated by converting the slope produced by the linear regression of the curves to percentage efficiency using the formula: Efficiency = $-1 + 10^{(-1/\text{slope})}$.

Results

In the *G. irregulare* (DAOM-197198) isolate, the mean number of LSU rDNA copies per nucleus is 31.27 (± 0.25 SE, n = 3), as estimated by real-time PCR and assuming a genome size of 15.74 Mb (Hijri and Sanders 2004). Note that because of the multinucleate organization of AMF, this copy number estimate represents an average over the nuclear population within this *G. irregulare* isolate.

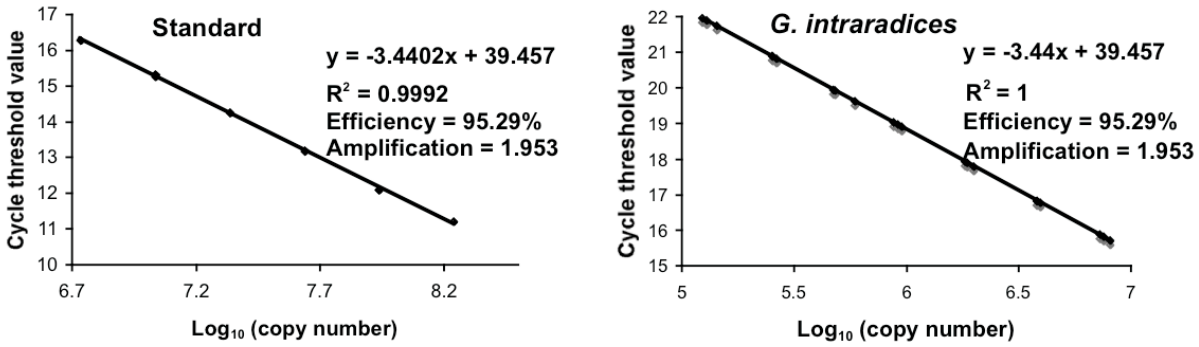


Figure 2-3 Absolute quantification of rDNA copy number. Results of real-time quantitative PCR showing linear regressions of the cycle threshold (CT values) and the Log concentration of **a)** plasmid DNA that was used as standard and **b)** LSU rDNA of *Glomus sp*. isolate DAOM-197198. Three replicate amplifications were performed for each sample. Data points shown in the graphs represent the CT values of the three replicates

3. Allelic diferences among sister spores suggest genetic drift at sporulation in an arbuscular mycorrhizal fungus

Eva Boon[†], Erin Zimmerman[†], Marc St-Arnaud and Mohamed Hijri

Université de Montréal, Département de Sciences Biologiques, Institut de Recherche en Biologie Végétale, 4101 rue Sherbrooke Est, QC, H1X 2B2, Canada

[†]These authors contributed equally to this work

This chapter was submitted to *BMC Evolutionary Biology in March 2012*

Data deposition: Nucleotide sequences were deposited in GenBank under accession numbers GQ325050-GQ325231

Author contributions

Eva Boon analyzed and interpreted the pyrosequencing data, created the figures and wrote the manuscript with assistance from Erin Zimmerman, who performed the experimental setup and cloned and sequenced the *PLS* data. Marc St-Arnaud was involved in the design of the experimental setup. Mohamed Hijri was the principal investigator, was responsible for the *PLS* pyrosequencing run, and participated in the design of the study. All authors read and approved the final manuscript.

Abstract

Arbuscular mycorrhizal fungi (AMF) are root-inhabiting fungi that form mutualistic symbioses with their host plants. AMF are made up of coenocytic networks of hyphae through which nuclei and organelles can freely migrate. Mature AMF spores contain hundreds of nuclei obtained directly from the parent mycelium. To our knowledge, AMF are the only known organisms that do not pass through a uninucleate stage in any part of the life cycle, which allows nuclei to diverge into genetically distinct nuclei within the same cytoplasm. In this study, we investigated the presence and size of the genetic bottleneck for a low-copy Polymerase1-like gene, *PLS.* We inoculated tomato plants with one spore each of a single *Glomus etunicatum* strain. *PLS* amplicon products, from eight selected daughter spores from three tomato cultures, were sequenced by Roche FLX Titanium sequencing. In parallel, nucleotide diversity of 182 Sanger-sequenced *PLS* sequences from the same fungal isolate was investigated to compare Sanger-derived allele distributions to the pyrosequencing results. Our objectives were (1) to estimate what genetic diversity is passed on to a single spore, (2) to determine whether this diversity is less than the total amount of variation found in all spores, and (3) to investigate whether there is any differential segregation of genetic variation. Our results indicate that spores contain only a subset of the total genetic variation of the pooled spores, and contain structurally different populations of the *PLS* allele. We conclude that a potential genetic bottleneck occurs at sporulation in AMF, resulting in differences in genetic variation between sister spores. We suggest that the effects of this bottleneck are countered by anastomosis (hyphal fusion) between related hyphae, a well-described phenomenon in AMF.

3.1. Introduction

Arbuscular mycorrhizal fungi (AMF) are root-inhabiting fungi that form mutualistic symbioses with plants and are grouped in the phylum Glomeromycota (Schussler *et al.* 2001b; Smith and Read 2008). They improve nutrient uptake in their host plants and buffer the plant against both abiotic and biotic stresses (St-Arnaud and Vujanovic 2007). As a consequence of these characteristics, AMF significantly increase plant growth rates, with benefits varying depending on the composition of both the fungal and plant communities (Van der Heijden *et al.* 1998). AMF are of great potential interest to agriculture, and in recent years much progress has been made to understand the peculiar genetics and reproductive biology of these organisms.

Arbuscular mycorrhizal fungi are made up of vast, branching networks of hyphae. These hyphae are coenocytic, or lacking discrete cellular divisions; hyphal walls form long, tube-like structures through which cytoplasm and organelles can freely migrate. At sporulation, spores are formed as outgrowths from the parent hyphae, and large numbers of nuclei migrate directly into the spores (Marleau *et al.* 2011).

Mature AMF spores therefore contain hundreds of nuclei obtained directly from the parent mycelium, making these fungi, to our knowledge, the only known organisms which do not pass through a uninucleate stage in any part of their life cycle (Bever *et al.* 2008; Sanders and Croll 2010; Marleau *et al.* 2011). This is significant, because the uninucleate stage acts as a genetic bottleneck that is hypothesized to prevent genetic conflict by ensuring that each daughter cell possesses only one copy of the original genome.

The prevention of genetic conflicts is at the heart of many theories of individuality (Buss 1987; Michod and Roze 2001; Queller and Strassmann 2009).

Nevertheless, strict soma-germline division is applicable only to metazoans. Many other multicellular organisms, such as plants, fungi and algae, do not have such a clear-cut soma-germline division and consequent bottleneck in their life history. Accordingly, the extent of intra-organismal genetic heterogeneity among multicellular organisms may be vastly underestimated (Santelices 1999; Pineda-Krch and Lehtila 2004a; Santelices 2004b).

In an organism that lacks a uninucleate stage, genome polymorphisms due to somatic mutations are allowed to pass on to the next generation, resulting in individuals that can theoretically contain any number of differentiated genomes (Sanders 2002). Numerous studies now suggest this to be the case in AMF, with evidence for intra-isolate variation having been found in rDNA (Sanders *et al.* 1995; Clapp *et al.* 1999; Pringle *et al.* 2000; Kuhn *et al.* 2001; Rodriguez *et al.* 2001; Boon *et al.* 2010), protein-coding genes (Kuhn *et al.* 2001; Helgason *et al.* 2003; Corradi *et al.* 2004a; Corradi and Sanders 2006; Corradi *et al.* 2007; Boon *et al.* 2010) and even in the transcriptome (Boon *et al.* 2010; Tisserant *et al.* 2012).

What evidence is there for the heterokaryotic state in AMF? An early study by Kuhn *et al*. (2001) used double-target fluorescent DNA-DNA *in situ* hybridization to visualize two divergent sequences of the ITS2 region, T2 and T4. These two markers were found to occur in varying frequencies in different nuclei, indicating non-identical genomes. Later work by Hijri and Sanders (Hijri and Sanders 2005) showed that *PLS* (*Polymerase 1*-like sequence), a marker previously found to occur in at least 13 variants within an isolate of *Glomus etunicatum* (Pawlowska and Taylor 2004), is a two-copy gene in that species. This again pointed to heterokaryosis and the lack of a strong genetic bottleneck in the AMF life history. The most recent work on *PLS* has shown that the variation seen at the genomic level persists in the cDNA, meaning the effect is not simply attributable to a high number of pseudogenes in the AMF genome (Boon *et al.* 2010).

If the number of dissimilar *PLS* alleles that are distributed randomly throughout the mycelium is larger than the average nuclear content of a daughter spore, it is unlikely the daughter spore will receive the full genetic complement of the parent and genetic drift becomes a possibility. Here, we define drift as the change in allele frequency from parent mycelium to spore, due to random sampling of alleles. Without replenishment of genetic variation, drift leads to a loss of nuclear variants with each new generation and leads to a loss of heterokaryosis over time. This process was modeled by Bever and Wang (Bever and Wang 2005; Bever *et al.* 2008), who found that the only means of countering this genetic loss is through the action of anastomosis, in which hyphal strands, either within or among mycelia, fuse and nuclei are exchanged.

Here, we explore the extent of within-isolate genetic polymorphism and test for the presence of genetic drift in the AMF *Glomus etunicatum*. This was accomplished by inoculating tomato plants with individual spores corresponding to a single mycorrhizal strain (figure 3-1). Eight daughter spores were isolated from three such cultures; from each, we amplified and sequenced a fragment from a Polymerase1-like gene, *PLS*. As mentioned above, several markers are found to be polymorphic between nuclei in AMF. We chose the *PLS* marker because it is present in low copy numbers in *G. etunicatum* genomes (Hijri and Sanders 2005), and we would therefore be less likely to confound intra-genome with inter-genome variation. Amplicon products from the selected daughter spores were then sequenced using Roche FLX Titanium sequencing. In this way, genetic variation within each spore could be compared to that of spores from the same plant, or to all spores together, to estimate the size of the potential bottleneck in genetic variation that occurs during sporulation in AMF.

Our specific objectives were (1) to estimate what proportion of genetic diversity is passed on in a single spore, (2) to determine if this proportion is less

than the total amount of variation found in AMF, and (3) to investigate whether there is any differential segregation of genetic variation.

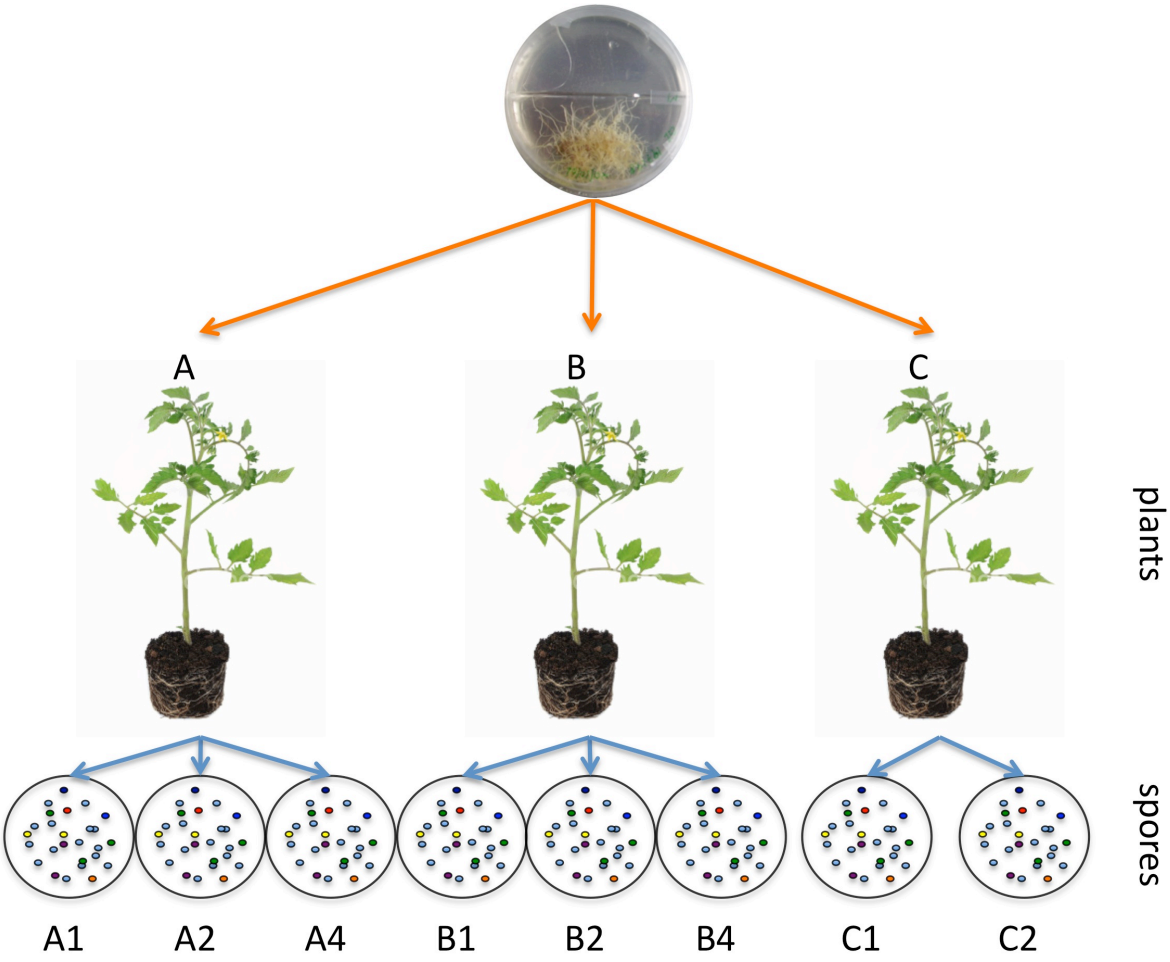


Figure 3-1 Diagram of the experimental setup showing three parent spores taken from a single isolate of *G. etunicatum* grown in pot culture to produce eight progeny spores for pyrosequencing.

3.2. Materials and Methods

3.2.1. Establishment of single spore pot cultures

Spores of *Glomus etunicatum* isolate NPI (Native Plants Incorporated, Salt Lake City) were isolated from calcined clay through wet sieving and rinsed in sterile water. Spores that appeared opaque and dark brown in color were considered healthy and chosen preferentially for use in the experiment.

One hundred and fifty 115 ml cone-tainers (Stuewe & Sons, Inc., Corvallis, OR, USA) were filled with a sterilized mixture of 1:1:1 (v/v) perlite, sand, and field soil collected from the Montréal Botanical Garden. In each cone-tainer, a 1000 μl micropipette tip with the smaller end trimmed off was pushed into the substrate such that the larger end was level with the surface. A germinated tomato seedling *Solanum lycopersicum* L. cv. Micro-Tom (Meissner *et al.* 1997) with a single *G. etunicatum* spore placed upon the root was then inserted into each micropipette tip, helping to keep the spore in close proximity to the root, and topped with more sterile soil mixture. Cone-tainers were then maintained in a growth chamber (see table 3-5 for environmental conditions).

Following a six-month growth period, 25 plants were chosen at random for spore collection. The roots of these plants were cut into small pieces and wet sieved along with the soil core to isolate spores. Spores were separated from debris of similar size by centrifuging at 1620 g for 2 min in a discontinuous density gradient with a 60% (w/v) sucrose fraction. The spores were collected from the gradient interface with a Pasteur pipette and thoroughly washed with distilled water before being stored in distilled water at 4°C.

3.2.2. Gene amplification and pyrosequencing of individual spores

Two to three spores from each of three replicate host plants were isolated for further analysis. These spores were individually picked and placed in 1 μl sterile water in a 0.2 ml tube, then crushed using fine forceps, which were flame-sterilized between samples. The entire DNA content of each of the spores was amplified using the GenomiPhi Whole-Genome Amplification Kit (GE Healthcare, Canada) according to manufacturer instructions. *PLS* sequences were then amplified using Dream*Taq* DNA polymerase (Fermentas) using primers Pol4-A (GAATCCTTCCCAAATTGATCAGAATACT TGTT) and Pol7-B (TAATAATAAAAGCCTTTCAAAAAATCCATCAATA), with added pyrosequencing adaptor sequences at the 5' ends (forward: CCATCTCATCCCTGCGTGTCTCCGACTCAG, reverse: CCTATCCCCTGTGTGCCTTGGCAGTCTCAG) (Pawlowska and Taylor 2004). The reaction was performed in 50 μl volumes containing 0.2 mM dNTPs, 0.5 μmol of each primer and the PCR buffer. PCR was carried out for 34 cycles (95°C for 30 sec, 50°C for 30 sec, 72°C for 1 min; preceded by an initial 2 min denaturation at 94°C and followed by a 10 min hold at 72°C) on a Mastercycler EPgradient S (Eppendorf). The PCR product was checked on an electrophoresis gel to ensure successful amplification of the gene, then purified using a MinElute PCR Purification Kit (Qiagen) according to manufacturer instructions. These purified samples were then sent to the Genome Québec Innovation Centre (McGill University, Montréal) for pyrosequencing using the GS FLX Titanium emPCR kit (Roche 454 Life Science) with lib-L chemistry (1/8 run per sample).

3.2.3. Pyrosequencing analysis

All analyses were performed in Mothur v.1.22 (Schloss *et al.* 2009), unless specified otherwise. Stringent checks were performed on the raw pyrosequencing

data to remove low-quality reads and minimize sequencing errors which could have been introduced during the pyrosequencing process (Huse *et al.* 2007; Schloss *et al.* 2011). Eliminated reads included those that (i) did not perfectly match the adaptor and primer sequences, (ii) had ambiguous bases, (iii) had a quality score below an average of 35 in a window of 50 bp, (iv) contained homopolymer lengths greater than 8 bp. Reads that passed quality control and differed by just 1% were preclustered following Huse *et al.* (2010). Chimeric molecules can form *in vitro* during PCR (Wang and Wang 1997) or pyrosequencing (Haas *et al.* 2011). To control for this artifactual genetic variation, we removed chimeric reads that did not match a database of previously obtained Sanger sequenced *PLS* sequences (Boon *et al.* 2010) with less than 90% bootstrap support using the program Chimeraslayer (Haas *et al.* 2011), as implemented in Mothur v.122.

The remaining clusters are henceforth referred to as 'alleles'. We performed rarefaction analyses for the entire dataset and for the 'plant' and 'spore' subsets. A clear asymptote in the rarefaction analysis is an indication that sampling was representative of the actual diversity at the *PLS* locus. We calculated alpha-diversity indices for plants and spores, including the total number of alleles and the Chao1 diversity index (which is a measure of the minimum richness). To explore beta-diversity, we calculated the Jaccard similarity coefficient (jclass) based on the observed richness and the Yue & Clayton (thetaYC) measure of dissimilarity between the structures of two communities. The statistical significance of these latter two coefficients was calculated using the (un)weighted Unifrac distance metric (Lozupone and Knight 2005). Statically significant differences between spore and plant communities (and community structures) were determined using the nonparametric analysis of molecular variance (AMOVA) (Excoffier *et al.* 1992; Anderson 2001; Martin 2002), a distance-based version of Bartlett's test for homogeneity of variance (HOMOVA) (Stewart and Excoffier 1996) and with a two-sampled Kolmogorov-Smirnov test.

3.2.4. Cloning and Sanger sequencing of *PLS* alleles

Spores of *G. etunicatum* isolate NPI grown in leek pot cultures (Smith and Dickson 1997) in a greenhouse were removed from soil via wet-sieving as described above. Genomic DNA was obtained *en masse* from approximately 150 spores using a DNeasy Plant Mini Kit (Qiagen) according to the manufacturer's directions. The DNA was then used as a template to amplify *PLS* using high fidelity *Pfu* DNA polymerase (Fermentas) and specific primers FwdPOL4 and RevPOL7 (see above). The reaction was performed in a 50 µl volume containing 1.25 units *Pfu*, 0.2 mM dNTPs, 0.5 µM of each primer and the PCR buffer. PCR was carried out for 34 cycles (94°C for 30 sec, 54°C for 30 sec, 72°C for 1 min; preceded by an initial 2 min denaturation at 94°C and followed by a 10 min hold at 72°C) on a Mastercycler EPgradient S (Eppendorf). After the PCR product was checked on an electrophoresis gel to ensure successful amplification, the amplified gene was cloned using a CloneJET PCR Cloning Kit (Fermentas), according to the manufacturer's instructions. Three hundred and six bacterial colonies containing the *PLS* insert were subcultured on Luria-Bertani (LB) medium (Bertani 1951) containing 100 mg/l ampicillin, and sequenced at the Genome Québec Innovation Centre (McGill University, Montréal) using pJET1.2 forward and reverse sequencing primers (Fermentas).

3.2.5. Sanger sequencing analysis

Sequence chromatograms were visualized using the program Finch TV (version 1.4.0, Geospiza Inc.); those that were low-quality, incomplete, or contained plasmid sequences were removed from the analysis. Sequences were aligned using MEGA4 (Tamura *et al.* 2007), corrected by hand, and deposited in GenBank with accession numbers GQ325050-GQ325231. The online program

DNAcollapser (Villesen 2007) was used to collapse the sequences to distinct variants.

3.3. Results

3.3.1. Pyrosequencing

A total of 131,797 raw sequence outputs were recovered. After stringent quality checks, a total of 14,320 reads were retained, which represents approximately 11% of the original data. The retained alignment covers roughly 200 bp of the *PLS* gene, including the end of the first exon, the entire first intron, and the beginning of the second exon, according to the structure of the gene as previously described in Boon *et al.* (2010).

Using rarefaction analyses, we determined that the genetic diversity in the alignment under study, after quality control and clustering, had not been exhaustively sampled. Considering that the intron represents most of the genetic diversity (figure 3-4), we excluded the intron from the alignment. Even with this shortened alignment the saturation curve did not reach the lower confidence interval for the estimated richness according to the Chao1 estimate for any of the spores, spores pooled by plant or the entire dataset (figure 3-3). We performed the remaining analyses on the exonic sequences.

The 14,320 reads were clustered at 1% (i.e. sequences with 1 bp difference for an exon-only alignment of 100 bp), following Huse *et al*. (2010). Every cluster was consequently considered an allele for the *PLS* locus (table 3-1). The number of alleles, the bias-corrected Chao1 richness estimator and the Simpson diversity index are summarized in table 3-1.

Table 3-1 Alpha-diversity estimates for the 1st (partial) and 2nd exonic sequence of *PLS.* Alpha diversity was calculated for individual spores, spores grouped per plant and for all spores combined.

		# reads before quality control	# reads after quality control	% reads retained	# alleles	Chao1	Chao1 95% lci	Chao1 95% hci	bottleneck[1] (%) size	bottleneck[1] (%) 95% lci	bottleneck[1] (%) 95% hci
	all	131797	14320	11	356	813	664	1036	56	46	66
Plants	A	65274	5767	9	209	538	398	780	61	48	73
	B	49704	5928	12	196	361	292	479	46	33	59
	C	16819	2625	16	120	231	180	328	48	33	63
Spores	A1	11242	1052	9	68	154	105	267	56	35	74
	A2	43001	3811	34	152	314	240	452	52	37	66
	A4	11031	904	2	87	209	147	333	58	41	74
	B1	22711	3585	32	135	218	179	293	38	25	54
	B2	5051	1536	7	82	170	124	267	52	34	69
	B4	21942	807	16	79	229	150	398	66	47	80
	C1	11920	2032	9	98	191	145	282	49	32	65
	C2	4899	593	5	43	118	72	237	64	41	82

[1]Reduction of number of alleles in comparison to the Chao1 index for the group, its lower confidence interval (lci) and higher confidence interval (hci)

Significance of beta-diversity estimators is summarized in table 3-2. Genetic diversity between all spores was significantly different for the UniFrac, AMOVA and HOMOVA tests. Significance between spores was dependent on the test applied, with weighted UniFrac scores significant between all spores, no significant differences between spores using AMOVA and for the HOMOVA test only comparisons with spore A1 were significant. Beta-diversity indices for plants were significantly different from each other applying the weighted UniFrac test, but not for AMOVA or HOMOVA tests (table 3-3). Applying a Kolmogorov-Smirnov test, only the comparison of allele distributions between spores A1-B4 was significantly different, and none of the allele distributions between plants were significantly different (tables 3-2 and 3-3).

Table 3-2 P-values of three different tests that evaluate whether *PLS* alleles show the same structure of genetic diversity between spores

Comparison	p-values[1]			Kolmogorov-Smirnov[5]
	UNIFRAC[2]	AMOVA[3]	HOMOVA[4]	α =0.01
A1-A2-A4-B1-B2-B4-C1-C2	<0.0010	0.002	0.002	n.a.
A1-A2	<0.0010	0.998	0.003	0
A1-A4	<0.0010	1	0.001	0
A1-B1	<0.0010	1	0.001	0
A1-B2	<0.0010	1	0.001	0
A1-B4	<0.0010	1	0.001	7
A1-C1	<0.0010	1	0.001	0
A1-C2	<0.0010	0.026	0.001	0
A2-A4	<0.0010	0.371	0.957	0
A2-B1	<0.0010	0.042	0.092	0
A2-B2	<0.0010	0.249	0.387	0
A2-B4	<0.0010	0.204	0.549	0
A2-C1	<0.0010	0.243	0.494	0
A2-C2	<0.0010	0.01	0.007	0
A4-B1	<0.0010	0.994	0.007	0
A4-B2	<0.0010	0.957	0.103	0
A4-B4	<0.0010	0.876	0.386	0
A4-C1	<0.0010	0.888	0.398	0
A4-C2	<0.0010	0.006	0.115	0
B1-B2	<0.0010	0.972	0.543	0
B1-B4	<0.0010	0.976	0.033	0
B1-C1	<0.0010	0.798	0.013	0
B1-C2	<0.0010	0.011	0.028	0
B2-B4	<0.0010	0.94	0.123	0
B2-C1	<0.0010	0.852	0.042	0
B2-C2	<0.0010	0.006	0.122	0
B4-C1	<0.0010	0.825	0.961	0
B4-C2	<0.0010	0.041	0.121	0
C1-C2	<0.0010	0.024	0.082	0

[1]Pair-wise error rate (Bonferroni): 0.002; [2]UNIFRAC tests describes whether the communities have the same structure by chance (Lozupone and Knight 2005; Lozupone *et al.* 2006); [3]Analysis of Molecular Variance (AMOVA) (Excoffier *et al.* 1992; Anderson 2001; Martin 2002) determines whether the genetic diversity within each community is significantly different from the average genetic diversity of both communities pooled together (Schloss 2008); [4]Homogeneity of Molecular Variance (HOMOVA) (Stewart and Excoffier 1996) tests whether the genetic diversity between spores is homogeneous; [5]two-sample Komogorov-Smirnov test with the null hypothesis that the allele distributions of the spores under comparison are the same. This column specifies the number of alleles for which the null hypothesis is rejected at α=0.01)

Table 3-3 Statistical tests for the difference between 'plants' community parameters

	p- value[1]			
	A-B-C	A-B	A-C	B-C
UNIFRAC[2]				
Sharedchao[5]	1	<0.001	<0.001	<0.001
Jclass[6]	1	<0.001	<0.001	<0.001
Thetayc[7]	0.46	<0.002	<0.001	<0.001
Amova[3]				
Sharedchao[5]	0.34			
Jclass[6]	0.17			
Thetayc[7]	0.52			
Homova[4]				
Sharedchao[5]	0.34	n.a.[9]	n.a.	n.a.
Jclass[6]	0.82	n.a.	n.a.	n.a.
Thetayc[7]	0.70	n.a.	n.a.	n.a.
Kolmogorov-Smirnov[8]	n.a.	n.s.[10]	n.s.	n.s.

[1]Experimental-wise error rate: 0.05;

[2]UNIFRAC (Lozupone and Knight 2005; Lozupone *et al.* 2006);

[3]Analysis of Molecular Variance (AMOVA) (Excoffier *et al.* 1992; Anderson 2001; Martin 2002);

[4]Homogeneity of Molecular Variance (Stewart and Excoffier 1996);

[5]shared Chao-1 richness estimate for an OTU definition (Chao *et al.* 2005);

[6]Jaccard index describing the dissimilarity between two communities;

[7]Yue & Clayton measure of dissimilarity between the structures of two communities;

[8]two-sample Komogorov-smirnov test with the null hypothesis that the allele distributions of the spores under comparison are the same.

[9]not applicable

[10]not significant

Despite the filtering of chimeric sequences based on a database of previously sequenced *PLS* alleles, we detected 6 recombination events, applying the estimator *R*m as implemented in the software DNAsp v5 (Rozas *et al.* 2003). To further investigate the genealogical relations between alleles, we estimated the best model of codon evolution in jModelTest v.0.1.1 (Posada 2008) and constructed a Maximum Likelihood tree in PhyML v3.0 (Guindon *et al.*), under the Hasegawa-Kishino-Yano +G model (Swofford *et al.* 1996), performing 1000 bootstrap replicates to test the significance of branch lengths. The alleles fell in five different groups with bootstrap support higher than 70. 346 out of 356 alleles belonged to group E, which represents 14,308 sequences (figure 3-2). The ten remaining alleles represented only thirteen sequences, from spores A2 and A4 and C2.

Finally, again using the software DNAsp, we compared the translated amino acid sequences and found 920 stop codons and 36 frameshift mutations in the total dataset. Of the 13 sequences that did not belong to group E, 10 were found to contain stop codons. All results are summarized in table 3-4.

Table 3-4 Changes on the amino acid level. Letters refer to groups in figure 3-2.

	group	Stopcodon alleles	Total # sequences
Frameshift	All	11	36
Stopcodon	All	53	920
	A	2	2
	B	2	4
	C	2	3
	D	0	0
	E	47	910

3.3.1. Sanger sequencing

In total, 306 clones were sequenced, with 185 yielding sequences that cover the entire targeted area. The remainder produced partial sequences. When the same first region of the second exon was evaluated at the same clustering level that had been employed for the pyrosequencing data (1%), the Sanger sequences yielded 4 alleles that were all present at high frequencies in the pyrosequencing

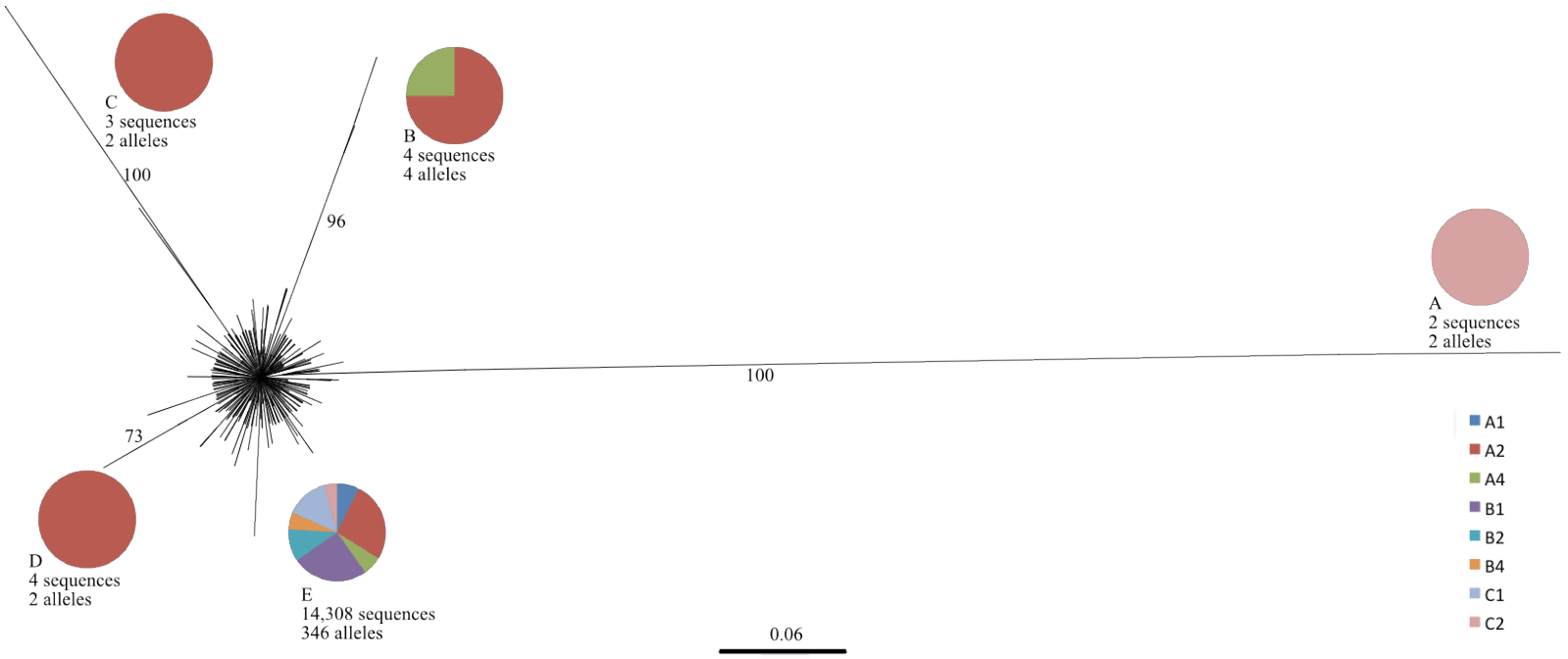


Figure 3-2 Phylogenetic analysis on genetic divergence between alleles (n=356). Evolutionary relations between alleles were inferred by using the Maximum Likelihood method based on the Hasegawa-Kishino-Yano model (Hasegawa *et al.* 1985) with a discrete gamma distribution (5 categories (+G, parameter = 11.0677)). Only bootstrap values higher than 70 are depicted. Scale bar indicates number of substitutions per site. Positions with less than 95% site coverage were eliminated, in total 92 nucleotide sites were used.

3.4. Discussion

Our first objective in this study was to estimate the proportion of genetic diversity for the marker *PLS* that is passed on in a single *G. etunicatum* spore. Comparing total allele diversity to that found per spore, estimated by means of the Chao1 statistic, the loss of allelic diversity can be as high as 41-82% for C2, the spore with the lowest allele diversity, and 37-66% for A2, the spore with the highest allele diversity. Placing a bottleneck of this magnitude (40%) in the context of the model of Bever *et al.* (2008) we find that without anastomosis, most allele diversity is expected to be lost in 10 generations, whereas with 2-25% anastomosis most allele diversity is conserved.

We further attempted to determine whether the proportion of genetic diversity transmitted is less than the total amount of variation found in the parent isolate, as approximated by the allele diversity from all spores pooled together. In other words, would a single progeny spore be likely to contain all the parental variation? Based on the smallest Chao1 estimates of between 72 and 237 alleles for spore C2, an individual spore would not be expected to contain all the genetic diversity of the parent isolate in our experiment. Even the allelic diversity estimates for spore A2, which is between 240 and 452 alleles, is smaller than the estimated minimum number of alleles from all spores pooled together (664 to 1036 alleles). Therefore, in our experimental setup a spore always contains less alleles than the original isolate.

Finally, we investigated whether differential segregation of genetic variation occurs. We tested this by comparing allele distributions between spores, using a range of statistical methods to test the significance of the differences between distributions (tables 3-2 and 3-3). If allele distributions are significantly different from each other, and no subsequent anastomosis takes places, differential segregation occurs between allele populations in the parent and progeny.

Significance of the tests varied with the estimator and the spore under consideration. The most abundant alleles were shared by all spores (figure 3-2) and were also found in the reads obtained via Sanger sequencing. However, we cannot reject the possibility that differential segregation may occur in the absence of anastomosis.

The approach taken in analyzing our data was quite conservative, both in terms of quality control on sequencing reads and the level of clustering employed to assign reads to alleles. The depth of our sampling was estimated with the Chao1 index, which is based on singletons and doubletons in the data (Chao *et al.* 2005). Therefore, it is possible that this conservative approach might have led to our failure to exhaustively sample all allele variation at the *PLS* locus on the spore and plant level. However, the confidence intervals of the Chao1 estimator do allow us to detect a significant difference between the total amount of allele diversity at the spore level vs. the total amount of allele diversity observed in all spores together. We therefore feel confident that a more exhaustive sampling of allele diversity will not compromise our conclusions based on the present dataset. Future studies on the subject of allele diversity in AMF, however, should take advantage of improvements in sequencing technology to sample AMF diversity with a greater depth and statistical power than we were able to achieve here.

This study provides strong indications that a single spore is not sufficient to transfer all genetic variation that was present in the parent cytoplasm to the next generation. This means that *G. etunicatum* is potentially exposed to a loss of genetic variation through drift. Recent research has demonstrated that genetic drift may occur in AMF isolates. *Glomus irregulare* isolate DAOM 197198 has been maintained in several different *in vitro* culture collections since 1992. Cardenas-Flores *et al*. (2010) found that subcultures from the same isolate which had been maintained in separate laboratories for 69 generations, anastomosed at a much lower rate than spores from the same isolate and the same subculture.

Segregation of genetically differentiated nuclei could lead to different genomic contents between subcultures from the same isolate. Decreased rates of anastomosis are observed between isolates that are genetically more different (Giovannetti *et al.* 1999; De la Providencia *et al.* 2005; Croll 2009; Marleau *et al.* 2011). If segregation of nuclei at sporulation is coupled to the absence of nuclear exchange by anastomosis, this would lead to rapid genetic differentiation between subcultures and could effectively extinguish intra-isolate genetic variation within a few generations as predicted by Bever and Wang (2005; 2008).

In spite of the extraordinary amounts of allelic diversity that was found in the spores, we propose that the large majority of *PLS* alleles that we investigated are under functional constraints, for the following reasons. First, the exon of the *PLS* gene shows considerably less nucleotide diversity than the intron (see supplementary materials). Second, only 14-15% of alleles were found to contain stopcodons or frameshift mutations (table 3-4). Finally, most alleles were found not to be different from each other, based on a maximum likelihood analysis (figure 3-2). The alleles that did vary were present at very low frequencies and mostly contained stopcodons, which means these alleles are not transcribed. The large amounts of allelic diversity in *G. etunicatum*, which translates into little amino acid diversity, confirms previous reports (Boon *et al.* 2010).

In the broader context of the heterokaryote state of *G. etunicatum*, it seems that the functioning of *PLS* within *G. etunicatum* is not influenced by the presence of multiple alleles. The sequence conservation on the amino acid level shows that the large amount of allele diversity is not translated into different amino acid sequences. However, this might not be the case for all loci that show genetic differentiation between nuclei. In the abovementioned case of diminished anastomosis rates between subcultures from the same isolate DAOM197198, genetic segregation can have a significant influence on a phenotypic trait, in this case anastomosis rate (Cárdenas-Flores *et al.* 2010). An additional piece of

evidence of the life history consequences of genetically differentiated nuclei in AMF is the observation that spores of *G. irregulare* which contain less than a critical amount of nuclei are unable to germinate (Marleau *et al.* 2011). Regardless of how the subtle interactions of differentiated genes and genomes within the same cytoplasm in AMF may play out, the major finding of this study is that a spore does not contain all genetic variation that can be found in a hyphal system. This observation might mean that AMF isolates, which are derived from single spores, might not display the full phenotypic spectrum of a hyphal system in the field.

3.5. Supplementary Materials

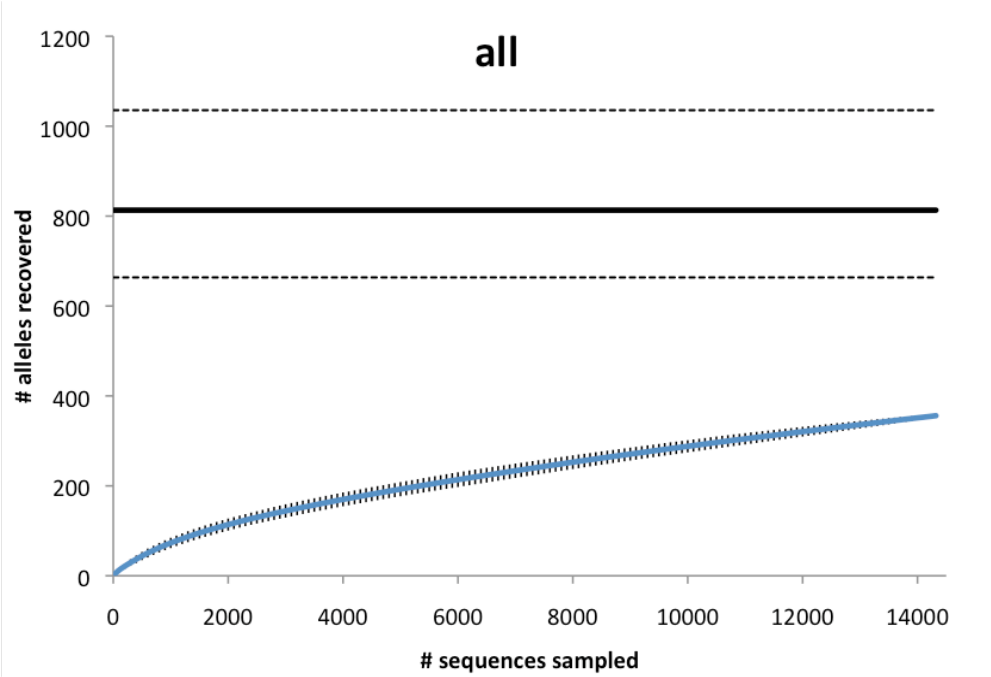
3.5.1. Tomato seedling growth conditions

Table 3-5 Environmental conditions under which tomato seedlings were grown

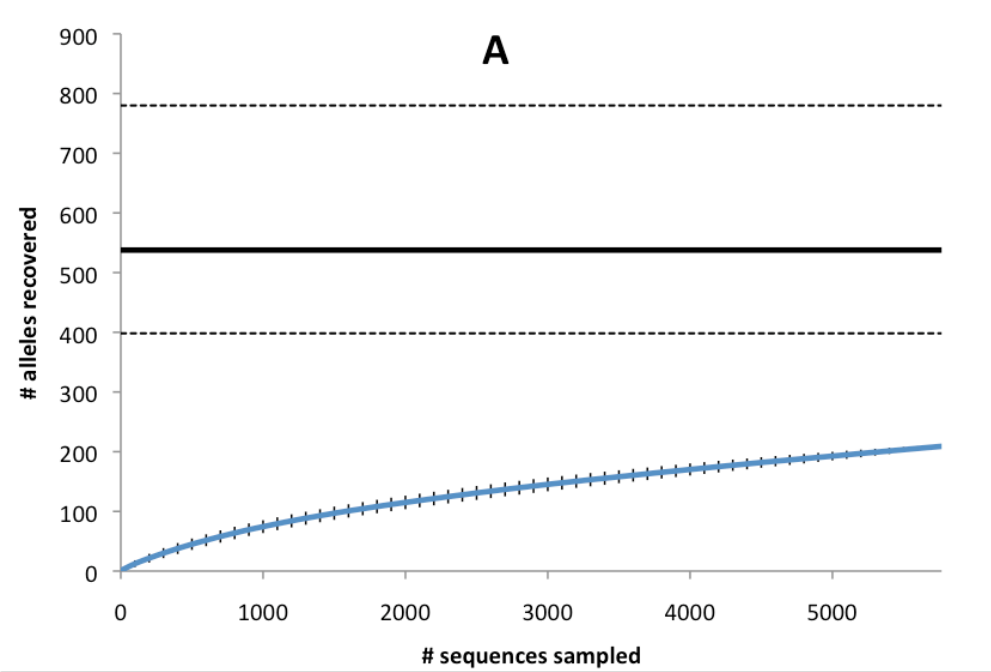
Time of Day	Temp. (°C)	Humidity (%)	Fluorescent bulbs	Incandescent bulbs
5h30	20	60	0	2
6h00	23	70	2	2
7h00	23	70	4	4
19h00	20	60	2	2
21h30	18	55	0	2
22h00	18	55	0	0

3.5.2.Rarefaction analyses

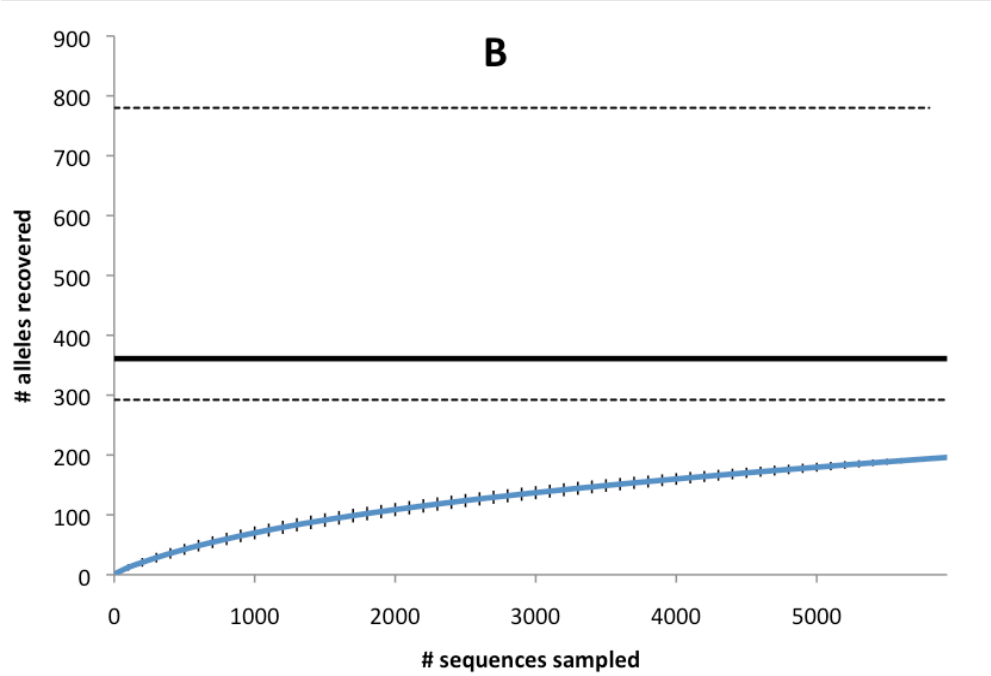
a)



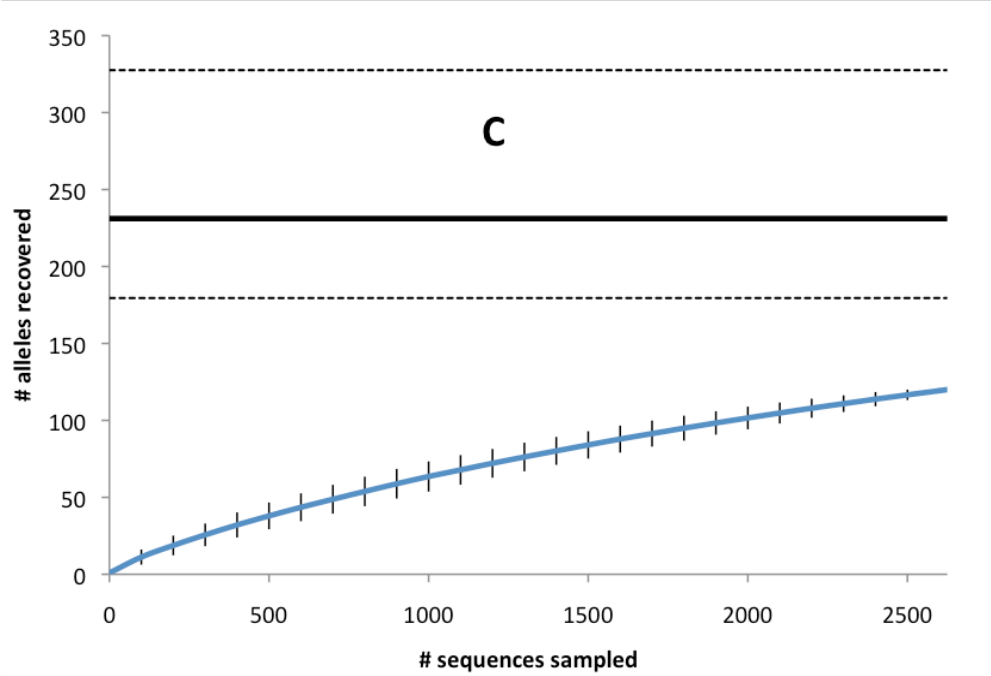
b)



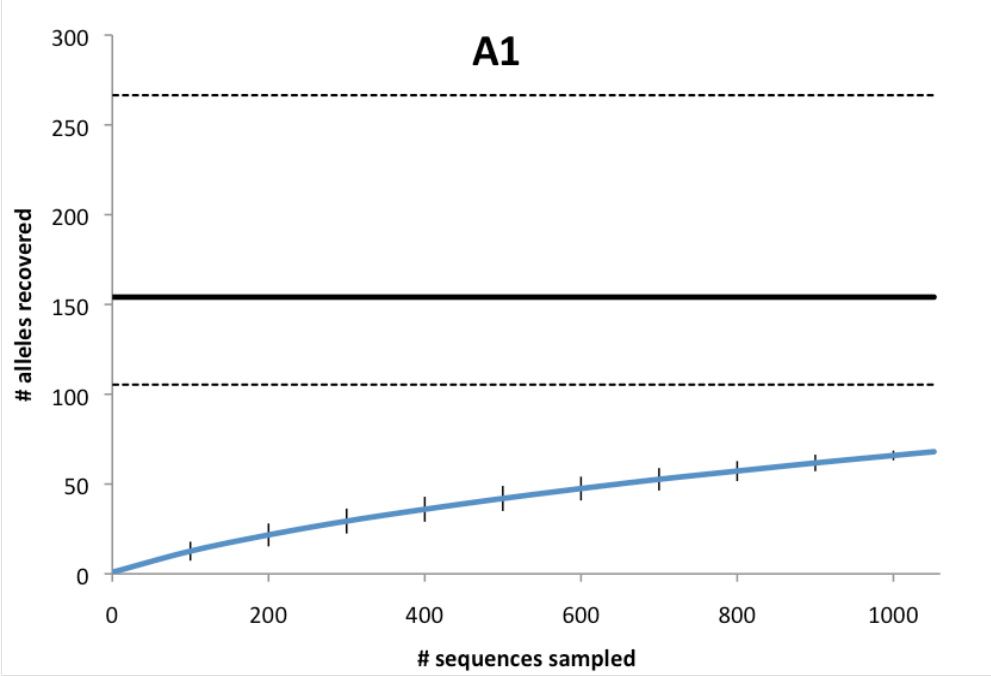
c)


B
900
800
700
600
500
400
300
200
100
0
alleles recovered
0
1000
2000
3000
4000
5000
sequences sampled

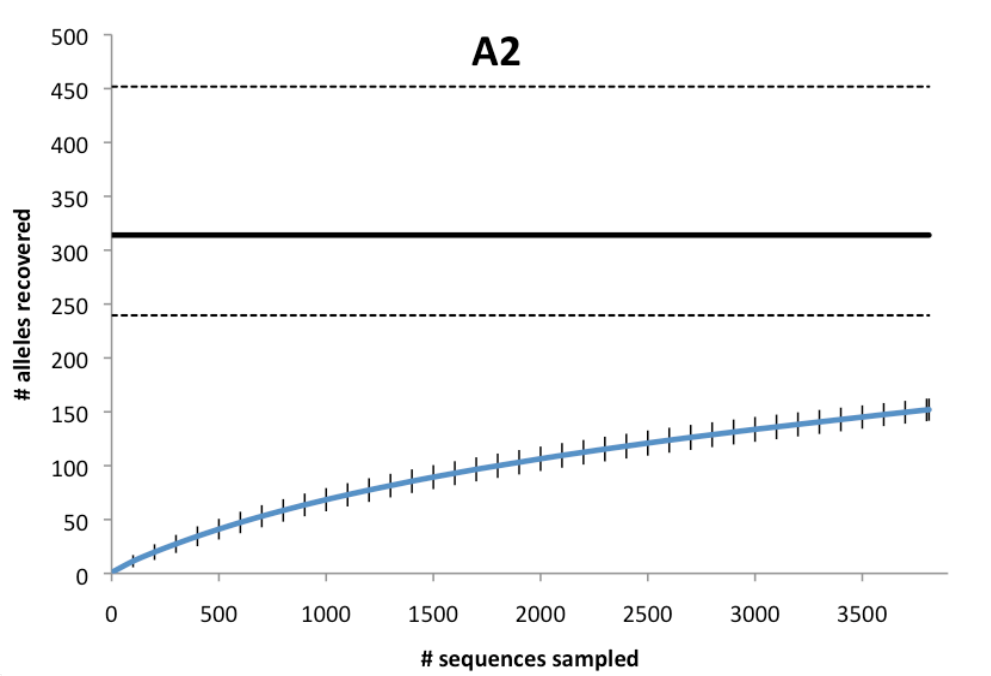
d)


C
350
300
250
200
150
100
50
0
alleles recovered
0
500
1000
1500
2000
2500
sequences sampled

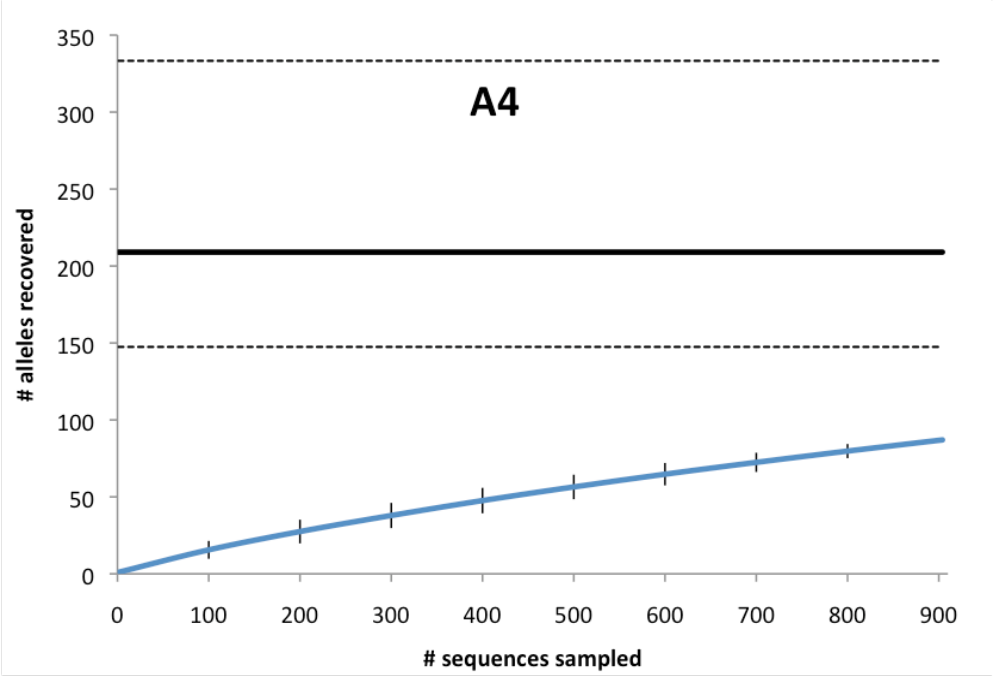
e)


A1
alleles recovered
sequences sampled
0
50
100
150
200
250
300
0
200
400
600
800
1000

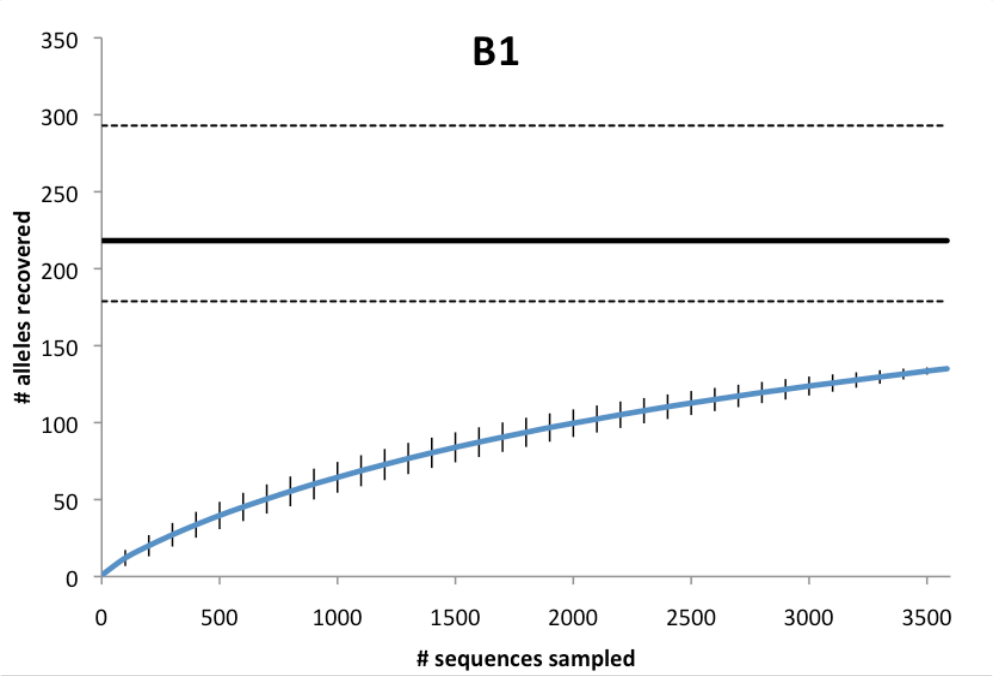
f)


A2
alleles recovered
sequences sampled
0
50
100
150
200
250
300
350
400
450
500
0
500
1000
1500
2000
2500
3000
3500

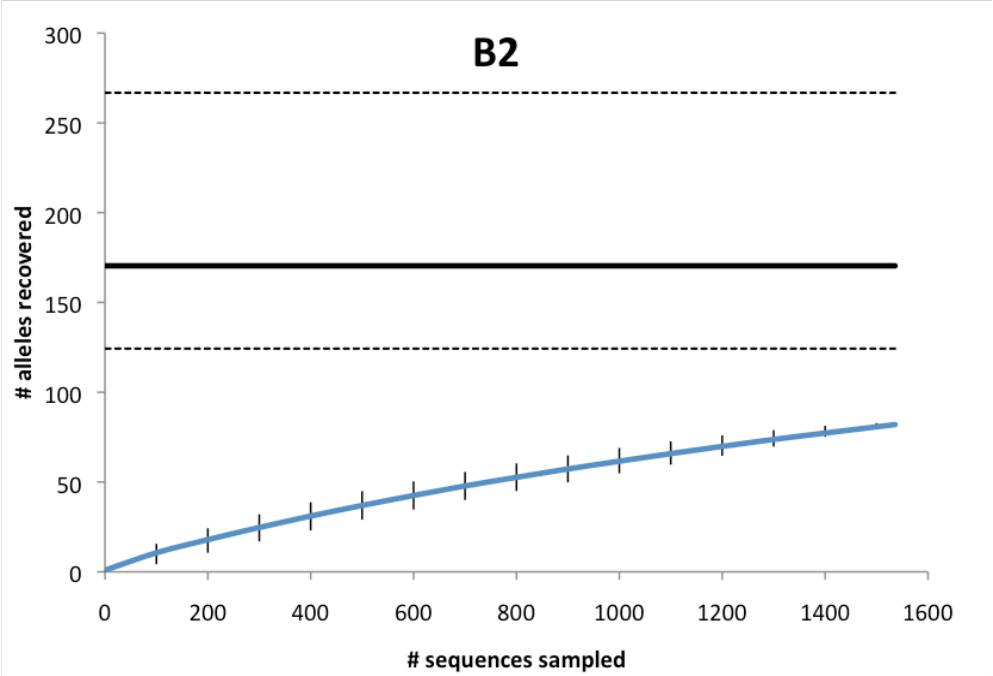
g)


A4
alleles recovered
sequences sampled
0
50
100
150
200
250
300
350
0
100
200
300
400
500
600
700
800
900

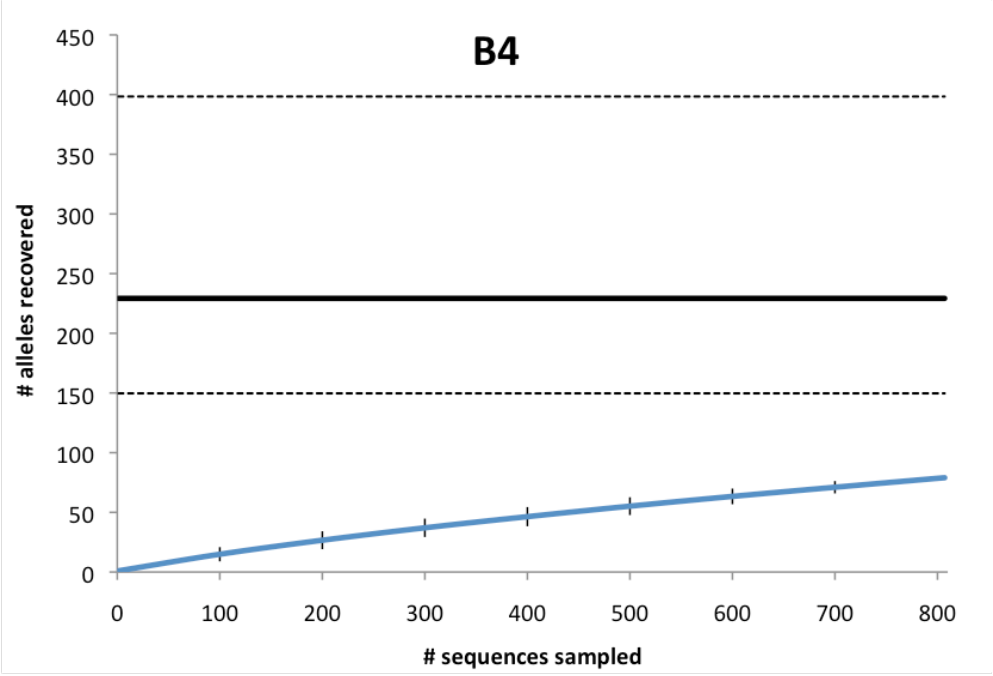
h)


B1
alleles recovered
sequences sampled
0
50
100
150
200
250
300
350
0
500
1000
1500
2000
2500
3000
3500

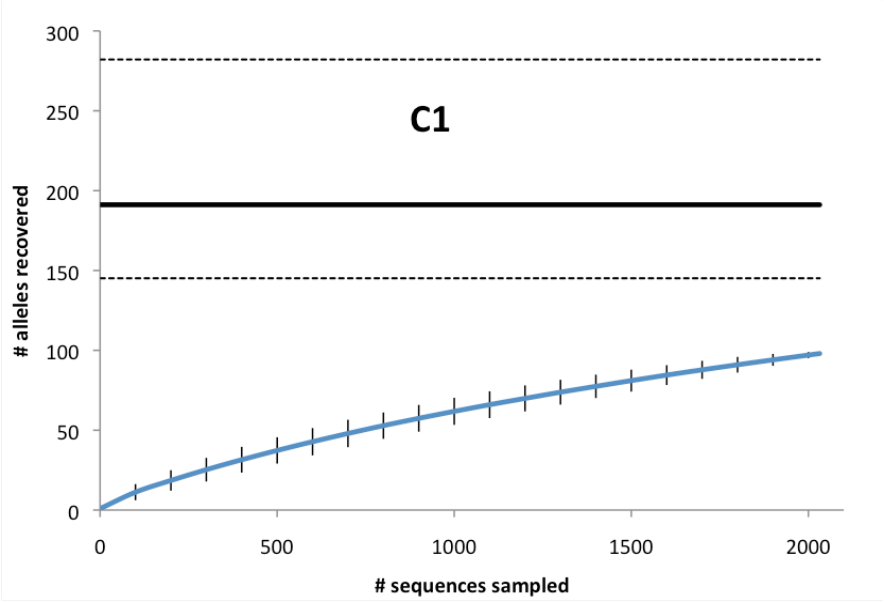
i)


B2
alleles recovered
sequences sampled
0
50
100
150
200
250
300
0
200
400
600
800
1000
1200
1400
1600

j)


B4
alleles recovered
sequences sampled
0
50
100
150
200
250
300
350
400
450
0
100
200
300
400
500
600
700
800

k)



l)

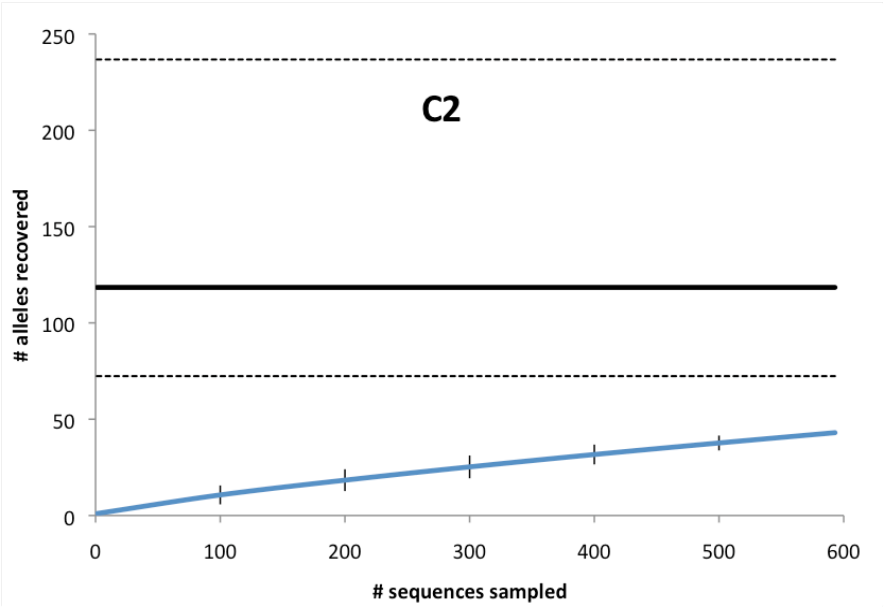


Figure 3-3 Rarefaction analyses for **a)** all spores pooled together (all) **b-d)** spores pooled by plant (A-B-C) and **e-l)** individual spores (A1-A2-A4-B1-B2-B4-C1-C2). The number of recovered alleles (y axis, grey line, 95% confidence intervals indicated by vertical lines) is compared to the maximum Chao1 value (Chao *et al.* 2005), which is the estimated minimum richness for each group (solid black line, 95% confidence intervals in dotted lines).

3.5.3.Nucleotide diversity along the *PLS* fragment

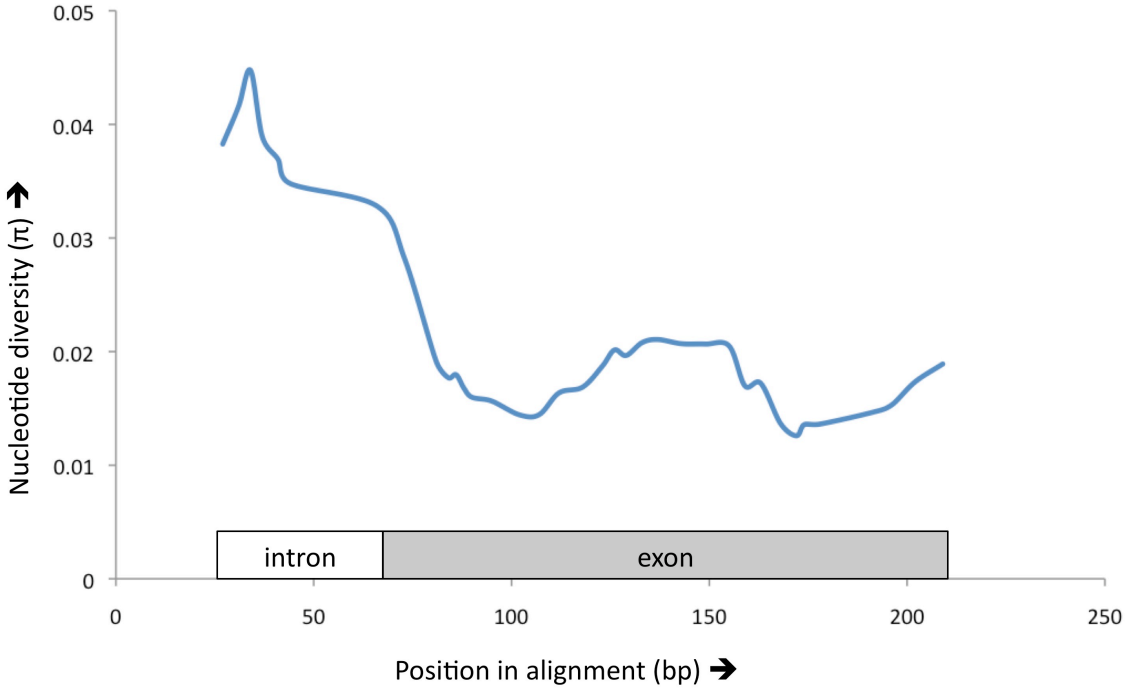


Figure 3-4 Nucleotide diversity along the fragment of the Polymerase-1 like gene under study.

4. Estimating genome diversity in arbuscular mycorrhizal fungi: the case for a pangenome in AMF

Eva Boon[1], Sébastien Halary [1], Eric Bapteste[2], Mohamed Hijri[1]

[1] Institut de Recherche en Biologie Végétale, Département de Sciences Biologiques, Université de Montréal, 4101 rue Sherbrooke Est, QC, H1X 2B2, Canada

[2] Université Pierre et Marie Curie (UPMC), Centre National de la Recherche Scientifique (CNRS), Unité Mixte de Recherche 7138, 75005 Paris, France

This chapter will be submitted to *Proceedings of the National Academy of Sciences*

Author contributions

Eva Boon developed, tested and pyrosequenced the single copy markers, analyzed the output from the clustering approach and wrote the manuscript. Sébastien Halary was involved in the experimental setup of the clustering analysis and performed the calculations. Eric Bapteste designed and orchestrated the bioinformatics workflow to select potential single copy markers in *Glomus*, was involved in the experimental setup of the clustering analysis and in writing the manuscript. Mohamed Hijri was the principal investigator and was responsible for the non-amplicon pyrosequencing runs. All authors read and approved the final manuscript.

4.1. Abstract

Arbuscular mycorrhizal fungi (AMF) are root-inhabiting fungi whose hyphal networks form symbioses with plants. The growth advantages they confer to their host plants are of great potential benefit to sustainable agricultural practices. However, measuring genetic diversity in arbuscular mycorrhizal fungi (AMF) is a major challenge, not only for ecologists in the field but also for molecular biologists in the lab. AMF show extremely high levels of genetic variation for some loci, which has lead to the proposition that they contain hundreds of genetically divergent nuclei that share the same cytoplasm, i.e. they are heterokaryotic coenocytes. No reproductive stage has as yet been observed in AMF, yet evidence is accumulating that the observed high levels of diversity could be maintained by the exchange of nuclei between hyphal systems and recombination. The extent of polymorphism in AMF genomes is unclear since no polymorphic marker has been sampled exhaustively. Here, we adopted two complementary approaches to estimate genome diversity in AMF. First, we use data from whole-genome shotgun pyrosequencing runs of three isolates from two different AMF species (*Glomus irregulare* and *G. diaphanum*) to estimate overall genetic diversity, based on a conservative network-based clustering approach. The similarity distributions of these clusters are compared to simulated pyrosequencing runs on *Candida albicans, Candida dubliniensis, Puccinia graminis, Puccinia triticina* and *Tuber melanosporum*. Second, we present five new single copy genomic markers for *G. irregulare* and *G. diaphanum*. Polymorphism of a single copy marker can be directly interpreted as polymorphism between nuclei, which we sampled by massive parallel amplicon sequencing for two *G. irregulare* isolates and one *G. diaphanum* isolate. Both approaches that we employed to study genetic diversity in AMF revealed high levels of genetic variation. Our analyses demonstrate that AMF isolates show marked differences in genome wide diversity patterns with respect to

the control genomes and extend the number of AMF loci that report high genetic diversity among nuclei. We propose that it is impossible to describe AMF genome diversity with a single genome. Their unique genome organization makes AMF the first eukaryote with a pangenome.

4.2. Introduction

The idea of genome homogeneity within multicellular individuals is inextricably linked to theories of individuality, where conflict at lower levels of biological organization needs to be resolved before an additional level can evolve (Buss 1982; Maynard Smith and Szathmáry 1995). Here, we loosely define the multicellular individual as a functionally integrated assemblage of cells that share the same evolutionary fate, and which can also be referred to as 'organism' (but see Pepper and Herron 2008 for a detailed overview of definitions of individuality). This link of genome homogeneity with individuality is based on the assumption that intra-organismal genetic heterogeneity (IGH) leads to conflict within the individual. IGH can indeed be detrimental to the multicellular organism, as advances in cancer genomics clearly show (Summers *et al.* 2002). However, recent reviews on IGH in non-model organisms question the ubiquity of the genetically homogeneous organism (Pineda-Krch and Lehtila 2004a; Folse and Roughgarden 2010).

A group of organisms that undoubtedly evoke questions about the criteria of individuality are arbuscular mycorrhizal fungi (AMF), which are clonally reproducing, root-inhabiting fungi that form their own phylum, the Glomeromycota (Smith and Read 2008). AMF improve nutrient uptake in their host plants, and buffer the plant against both abiotic and biotic stresses (Van Der Heijden and Sanders 2002). AMF significantly increase plant growth rates, with benefits varying depending on the composition of the AMF and plant community (Van der Heijden *et al.* 1998; Avio *et al.* 2006). AMF are of great potential interest to agriculture, and in recent years much progress has been made to understand the genetics and biology of these organisms (Bonfante and Anca 2009; Sanders and Croll 2010). In AMF, an unknown number of genetically divergent nuclei share the same cytoplasm, i.e. AMF are heterokaryotic coenocytes. This peculiar genomic

organization is the result of an unusual life history, in which genome variation is never reduced to less than roughly 65 nuclei per spore (Marleau *et al.* 2011).

High levels of genetic variation within the AMF cytoplasm lead to practical as well as conceptual challenges. Both kinds of challenges are inextricably linked, since failure to recognize the former will comprise the latter, and vice versa. First, the limitations of the molecular biology toolbox pose a practical challenge to studying the real extent of IGH in AMF. Most molecular biology approaches for the study of multicellular organisms assume genome homogeneity between cells. For example, a PCR is usually performed on a template of DNA that, to be present in sufficiently high concentrations, originated from multiple cells. If an amplicon is found polymorphic within an individual, this could be due to either a gene or a to genome duplication. However, it is equally possible that the different amplicons originate from differentiated genomes. Second, conceptual challenges include defining `the individual' and determining the relevant criteria for species delineation. The problem of species delineation is analogous to the problem of defining 'the individual', but is not directly influenced by IGH. However, it is impossible to group individuals into species based on molecular data, if we fail to fully sample genetic diversity in a genetically heterogeneous individual.

As of now, we have no clear picture of the organization and extent of AMF genetic diversity. Even the total genome size is largely undetermined for most *Glomus* species. The genome size of the only well-studied isolate, *G. irregulare* DAOM 197198, was reported to vary between 15Mb (Hijri and Sanders 2004) and 150Mb (Martin *et al.* 2008; Sędzielewska *et al.* 2011). Even though it has recently become possible to sequence small amounts of DNA or even single cells (Raghunathan *et al.* 2005; Lasken 2007), this approach is labor intensive and costly, demanding resources that are not available to most researchers. An expanded array of methods is required to study genome structure and organization of non-model organisms, with a focus on inquiries into genetic heterogeneity.

Here, we demonstrate how relatively cost-effective and easily applicable methods inspired from environmental (meta)genomics can be used within the cytoplasm of each AMF isolate, to provide novel estimates of diversity in an organism with genetically differentiated genomes. We adopt two complementary approaches to study the organization of genome diversity in AMF. First, we use data from whole-genome shotgun pyrosequencing runs of two *Glomus irregulare* isolates, DAOM 197198 and DAOM 234179 (abbreviated as Gi, Gi1 and Gi2, respectively), and one *G. diaphanum* isolate (abbreviated as Gd), all together referred to as '*Glomus'*, to estimate the distribution of sequence differentiation as measured by average % identity of shortest sequences within clusters of overlapping homologous reads, following Halary *et al.* (2009). This evolutionary network analysis allows us to compare *Glomus* clusters to simulated whole-genome shotgun pyrosequencing runs of fully sequenced fungal genomes with the closest available genome size. Second, we study genome differentiation in more detail by targeting polymorphic loci present in single copy in the *G. irregulare* and *G. diaphanum* genomes. We adopted a comparative bioinformatics approach to develop these markers, by screening ORFs in 15 fully sequenced fungal genomes (table 4-6, supplementary materials). Candidate loci were amplified using RT-PCR to confirm copy number. Since genetic variation between sequences that are present in single copy in the genome necessarily represents genetic variation between different nuclei within the same hyphal system, this approach allowed us to study inter-genomic genetic variation for specific loci. Together, these two approaches represent the first large-scale study to tackle the extent and organization of genetic variation in a multigenomic organism.

4.3. Materials and Methods

4.3.1. Genome simulations and clustering

Pyrosequencing of *Glomus* strains

Approximately one million sterile spores of *Glomus irregulare* DAOM-197198 (Mycorhise® ASP) were provided by Premier Tech Biotechnologies (Rivière-Du-Loup, QC, Canada) in a liquid suspension of 4000 spores per ml. This suspension was filtered on a sterile plastic 35 μm sieve. The spores were checked for root contamination under a binocular microscope. The fungal material of isolates *G. irregulare* DAOM 234179 and *Glomus diaphanum* DAOM 229456 were obtained from in vitro cultures. Spores and hyphae were freshly harvested by dissolving the Gellan-Gum matrix in which cultures were grown using a solution containing 0.0083 N sodium citrate and 0.0017 N citric acid. Spores and hyphae were gently crushed in a 1.5 ml microtube using a sterilized pestle. DNA was extracted using DNeasy Plant Mini kit (Qiagen), according to manufacturer instructions. The purified DNA was then sent to the Genome Québec Innovation Centre (McGill University, Montréal) for pyrosequencing using the GS FLX Titanium whole genome shotgun kit (Roche 454 Life Science), employing a full run for each sample

Genomes and pyrosequencing simulations

The aim of our simulation approach was to generate datasets that, in terms of number of reads, length distribution and technical bias, closely resemble the outcome of the pyrosequencing runs that generated our *Glomus* data. Performing pyrosequencing simulations allowed us to start our comparative clustering analysis

with exactly the same amount of reads for all samples, and to compensate for possible technical biases introduced during sequencing, such as erroneous nucleotide substitutions (Huse *et al.* 2007; Huse *et al.* 2010). We chose control genomes that were in the same order of magnitude as the predicted *Glomus irregulare* genome size, i.e. 15 Mb or 150 Mb, respectively. For similar fungal genomes in the 15 Mb estimate, we chose genomes of *Candida albicans* (strains wo1 and sc5314) and *Candida dubliniensis* (abbreviated as wo1, sc and dub, respectively). For genomes in the estimated upper range of the *G. irregulare* genome size of 150 Mb, our simulations were based on the *Puccinia triticina*, *Puccinia graminis* and *Tuber melanosporum* genomes (abbreviated as tri, gra and tub, respectively). Information regarding these genomes is summarized in table 4-1. Genomes were downloaded from either NCBI or directly from the sequencing centre or consortium (see table 4-1). Pyrosequencing simulations for all control genomes were performed in the sequencing simulator Metasim v0.9.1 (Richter *et al.* 2008). Simulation parameters included the number of reads and the number of flow cycles, which determines the mean and standard deviation of the reads distribution according to the formula cycles=3*(mean+(8*stddev) (Felix Ott, personal communication). Simulation details are provided in the supplementary materials, table 4-5.

Evolutionary Network Analysis

We clustered sequences with segments presenting significant similarity for simulated and *Glomus* reads. All sequences sharing at least 25% homology and 75 identical nucleotides, with a BLAT (Kent 2002) e-value cutoff of 1e-20, were grouped together using ENG (Halary et al, in prep). These clusters are described by two variables. The first variable is the “percentage of identical positions on the shortest sequence of an aligned reads pair”, from now on abbreviated as PID (figure 4-2a). For example, if we find a BLAT hit of 100 bp with 50% identity

between a 150 bp and a 200 bp sequence in a cluster, the number of identical bases is 100 bp*50%=50 bp, and these 50 bp represent 30 % of the smallest sequence, thus our PID is 30 % for that pair of sequences. This approach yields a highly conservative average percentage identity between sequence pairs in a cluster.

The second variable is the clustering coefficient, which corresponds to the number of connected reads pairs in the cluster, divided by the maximum number of possible connections. For example, if a cluster with four reads is transformed into a network with four nodes that are connected by five edges, the maximum number of edges is n*(n-1)/2 where n is the number of nodes. In our example, the maximum number of edges is six. Thus, the clustering coefficient is 5/6 = 0.83 (figure 4-2b). The more the clustering coefficient approaches zero, the less connected is the network.

Estimating the size of the pangenome

The reads from a given pyrosequencing run can give valuable information about the genome size of the *Glomus* isolate in question. A cluster of reads can be represented as an alignment, a 'contig', if we disregard the possibility that reads could have originated from separated differentiated genomes within the *Glomus* cytoplasm. The sum of all contigs and singletons (which are reads that could not be assigned to a cluster) will give an estimate of the *Glomus* genome size, if we do not take into account possible undersampling of the isolate's genome in our pyrosequencing run.

Table 4-1 Genomes used for pyrosequencing simulations

Species (strain)	genome size (Mb)[1]	genes[1]	status[1]	coverage[1]	source	accession no.
Candida albicans (sc5413)	27.56	14,217	assembly	10.4	NCBI[3]	AACQ01000001-AACQ01000413
Candida albicans (wo1)	21.68	-	complete	10	NCBI[3]	AAFO01000001-AAFO01000086
Candida dubliniensis (CD36)	14.62	6,095	complete	11	NCBI[4]	FM992688-FM992695
Puccinia triticina *(*1-1 BBBD Race 1)	126.64	-	assembly	20	PG[2]	ADAS01000001-ADAS01038776
P. graminis f. sp. *tritici* (CRL 75-36-700-3, race SCCL)	88.64	21,073	assembly	7.88	PG[2]	AAWC01000001-AAWC01004563
Tuber melanosporum (Mel28)	120	7,496	assembly	10	NCBI	CABJ01000001-CABJ01004455

[1]Information from NCBI website: http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi

[2] *Puccinia* Group Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/)

[3]*Candida* Sequencing Project, Broad Institute of Harvard and MIT (http://www.broadinstitute.org/

[4]Department of Molecular and Cell Biology, University of Aberdeen, School of Dental Science, Trinity College, Dublin, Wellcome Trust Sanger Institute (Jackson *et al.* 2009)

4.4. Development and amplification of single copy markers

Development of single copy markers

We focused our search for single copy markers on the genomes of *G. irregulare and G. diaphanum*, since these were the first taxa for which pyrosequencing runs were available. Proteins from fifteen fungal genomes, which were all available fungal genomes fully sequenced at time of analysis (table 4-6) were aligned using Phylogenie.pl (a bioinformatics pipeline provided by Dr. Philippe Lopez) to generate alignments of gene families. We used custom-made perl scripts to find all ORFs that were (1) present in all 15 fungal genomes and (2) only occurred once in all genomes. Gene families that responded to these criteria were extracted and their alignment was refined using HoT (courtesy of Dr. Oliver Deusch). Alignments were used to construct Neighbour-Joining (NJ) and Maximum Likelihood (ML) trees. We estimated the best model of codon evolution in jModelTest v.0.1.1 (Posada 2008) and applied the WAG+Gamma 4 categories model of nucleotide substitution. These trees were scanned to define gene families in which (1) fungi were monophyletic, and (2) were found in a canonical position with respect to other taxa (following the phylogeny published in James *et al.* 2006). Reads from the *G. irregulare* DAOM197198 and *G. diaphanum* pyrosequencing runs described above were then aligned against these likely vertically inherited highly conserved single copy genes. Alignments for which homologues could be found for both *G. irregulare and G. diaphamum* were retained. Alignment quality for selected markers was visually evaluated, applying unambiguity of alignment, sequence length and conservation as criteria to design PCR primers for candidate single copy markers in *Glomus*. Amplification of PCR primers was tested in the laboratory. After this confirmation, fluorescence intensity of markers was compared

to the fluorescence intensity of a known single copy marker Rad15 (Hijri and Sanders 2005; Corradi *et al.* 2007) using real time PCR (RT-PCR). Markers for which fluorescence intensity (Ct) values were lower than Rad15 during the same experiment (in three replicates) were considered to occur in single copy in the genome. RT-PCR experimental details are provided in the supplementary materials.

Amplicon pyrosequencing of single copy markers

Potential single copy marker sequences were amplified using Dream*Taq* DNA polymerase (Fermentas) using primers listed in table 4-7, with suitable adapter, key and MID sequences added. DNA from the strains *G. irregulare* DAOM 197198, *G. irregulare* DAOM 234328 and *G. diaphanum* was extracted as described in Boon *et al.* (2010) The reaction was performed in 50 µl volumes containing 1 ng DNA, 1,25 U Taq polymerase (Fermentas), 0.2 mM dNTPs, 0.4 µM of each primer and the PCR buffer. PCR was carried out for 40 cycles (94°C for 30 sec, Ta (see table 4-7 for Ta per primer pair) for 30 sec, 72°C for 1 min; preceded by an initial 3 min denaturation at 95°C and followed by a 10 min hold at 72°C, on a Mastercycler EPgradient S (Eppendorf). The PCR product was checked on an electrophoresis gel to ensure successful amplification of the gene, and then purified using a MinElute PCR Purification Kit (Qiagen) according to manufacturer's instructions. These purified samples were pooled by molecular weight and sent to the Genome Québec Innovation Centre (McGill University, Montréal) for pyrosequencing using the GS FLX Titanium emPCR kit (Roche 454 Life Science) with lib-L chemistry.

Analysis of pyrosequencing results for single copy markers

All analyses were performed in Mothur v. 1.22 (Schloss *et al.* 2009), unless specified otherwise. Stringent checks were performed on the raw pyrosequencing data to remove low-quality reads and minimize sequencing errors which could have been introduced during the pyrosequencing process (Huse *et al.* 2007; Schloss *et al.* 2011). Eliminated reads included those that (i) did not perfectly match the adaptor and primer sequences, (ii) had ambiguous bases, (iii) had a quality score below an average of 35 in a window of 50 bp, (iv) contained homopolymer lengths greater than 8 bp. Reads that passed quality control and differed by just 1 bp were preclustered following Huse *et al.* (2010). Chimeric molecules can form *in vitro* during PCR (Wang and Wang 1997) or pyrosequencing (Haas *et al.* 2011). To control for this artifactual genetic variation, we removed chimeric reads that did not match a database of previously obtained (Sanger sequenced) sequences with less than 90% bootstrap support using the program Chimeraslayer (Haas *et al.* 2011), as implemented in Mothur v.122. For the remaining reads (henceforth referred to as 'alleles'), we performed rarefaction analyses for every marker. A clear asymptote in the rarefaction analysis is an indication that sampling was representative of the actual diversity at the *PLS* locus. We calculated the total number of alleles and the Chao1 diversity index, which is a measure of the minimum richness in a sample (Chao *et al.* 2005).

4.5. Results

4.5.1. Genome simulations and clustering

Results from the clustering analysis are summarized in table 4-2. Even though the initial total number of reads was exactly the same for every *Glomus* and simulated reads set, the number of clusters varied widely. *Glomus* reads sets and reads sets from the larger control genomes (*P. triticina*, *P. graminis* and *T. melanosporum*) yielded high numbers of clusters, with a smaller number of reads per cluster. Reads from the smaller control genomes (*C. albicans* sc 5314, *C. albicans* wo1 and *C. dubliniensis*) clustered together in small numbers of clusters containing many reads. These clusters from the smaller control genomes effectively resolved into contigs representing most of the original genome. Average number of reads per cluster for the *Glomus* reads sets and reads sets from the larger control genomes were several orders of magnitude smaller than average number of reads per cluster for the smaller control genomes.

The mean PID per cluster, which is the percentage of identical positions on the shortest sequence of an aligned reads pair, was higher for *Glomus* pyrosequencing runs. Clustering coefficients were invariably highest for *Glomus* pyrosequencing runs, followed by the larger control genomes, with the lowest clustering coefficients for the smaller control genomes.

Table 4-2 Results from clustering analyses

strain	total # reads	# clusters	# reads in clusters	# singletons	average # reads/ cluster	average % ID /pair of homologous reads in the cluster (SD)	average PID (SD)	clustering coefficient
G. irregulare DAOM 234179 run 1								
Gi1	485491	37729	422770	62721	11	98 (1)	76 (11)	0.87
gra	485491	38620	416586	68905	11	92 (1)	51 (11)	0.72
tri	485491	33862	331616	153875	10	92 (1)	50 (7)	0.82
tub	485491	49868	429214	56277	9	92 (1)	51 (9)	0.8
dub	485491	17	476957	8534	28056	92 (2)	49 (19)	0.03
sc	485491	340	476582	8909	1402	92 (2)	50 (18)	0.28
wo1	485491	30	477027	8464	15901	92 (2)	49 (19)	0.05
G. irregulare DAOM 234179 run 2								
Gi2	639222	40221	581333	57889	14	98 (1)	76 (13)	0.84
gra	639222	33046	566891	72331	17	92 (1)	51 (13)	0.66
tri	639222	36429	460538	178684	13	92 (1)	50 (8)	0.79
tub	639222	52391	594448	44774	11	92 (1)	51 (10)	0.74
dub	639222	1	630117	9105	630117	92 (2)	47 (18)	0
sc	639222	100	629982	9240	6300	92 (2)	49 (18)	0.16
wo1	639222	12	630106	9116	52509	92 (2)	48 (18)	0.07
G. irregulare DAOM 197198								
Gi	398817	44989	290560	108257	6	98 (1)	76 (7)	0.9
gra	398817	53315	260521	138296	5	92 (1)	60 (6)	0.84
tri	398817	33851	198476	200341	6	92 (1)	59 (4)	0.87
tub	398817	44780	264612	134205	6	92 (1)	60 (4)	0.89
dub	398817	11795	380878	17939	32	92 (2)	61 (17)	0.37
sc	398817	11934	380010	18807	32	92 (2)	61 (16)	0.42
wo1	398817	10932	381249	17568	35	92 (2)	61 (17)	0.35
G. diaphanum								
Gd	1078190	46087	1000094	78096	22	98 (1)	74 (14)	0.77
gra	1078190	20125	973054	105136	48	92 (2)	53 (16)	0.53
tri	1078190	37849	807607	270583	21	92 (1)	52 (11)	0.72
tub	1078190	33908	581299	496891	17	92 (1)	52 (11)	0.51
dub	1078190	1	1060812	17378	1060812	92 (2)	49 (18)	0
sc	1078190	33	1060947	17243	32150	93 (2)	51 (19)	0.18
wo1	1078190	12	1060793	17397	88399	92 (2)	50 (19)	0.06

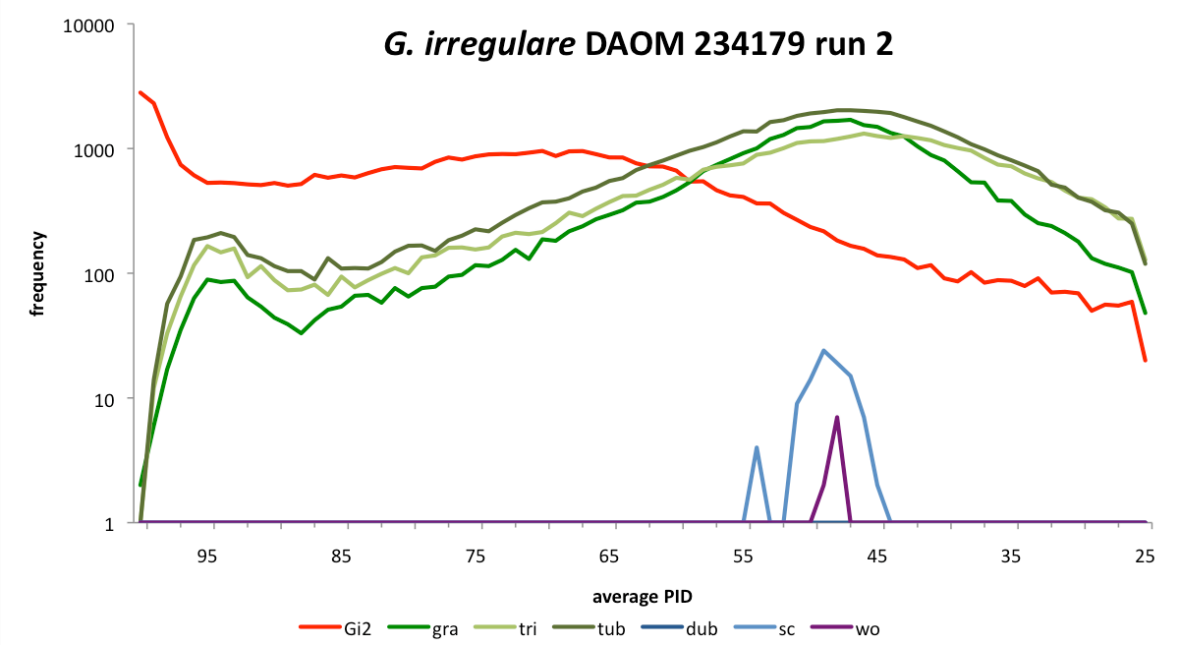
We plotted the frequency distribution of average PID in a similarity cluster from *Glomus* and from each (simulated) pyrosequencing run (figure 4-1). *Glomus* pyrosequencing runs showed an excess of clusters between 90-100% PID in comparison to the control genomes. Frequency distributions of the larger control genomes (*P. triticina*, *P. graminis* and *T. melanosporum*) consistently showed a small peak around 95 % PID and a larger peak around 45 % PID with lower kurtosis than the 95% PID peak. Frequency distributions of the smaller control genomes (*C. albicans* sc 5314, *C. albicans* wo1 and *C. dubliniensis*) invariably showed a more narrow PID range, between 45-55 % PID, reflecting the paucity of clusters for these control genomes. A sole exception was the reads distribution for control genomes for the *G. irregulare* DAOM 197198 run, the reads set with the smallest number of reads, which extended between 25-100 % PID (figure 4-3).

We also plotted clustering coefficients against average % PID in a similarity cluster (figure 4-4). All distributions showed a typical sigmoid curve. The *Glomus* read sets, however, generally showed a higher % PID for the same clustering coefficient. The clusters with the most reads (2,001-10,000; purple symbols and 10,001-70,000; orange symbols) showed the highest clustering coefficients for all *Glomus* and control reads sets. The clusters with the lowest and highest numbers of reads (<600, 601-1,000 and > 70,000 reads, respectively) showed the lowest clustering coefficients.

Estimating the size of the pangenome

The sum of all contigs and singletons for *G. irregulare* DAOM 234179 run 2 is 178 Mb. The sum all contigs and singletons for *G. diaphanum* is 163,7 Mb.

a)



b)

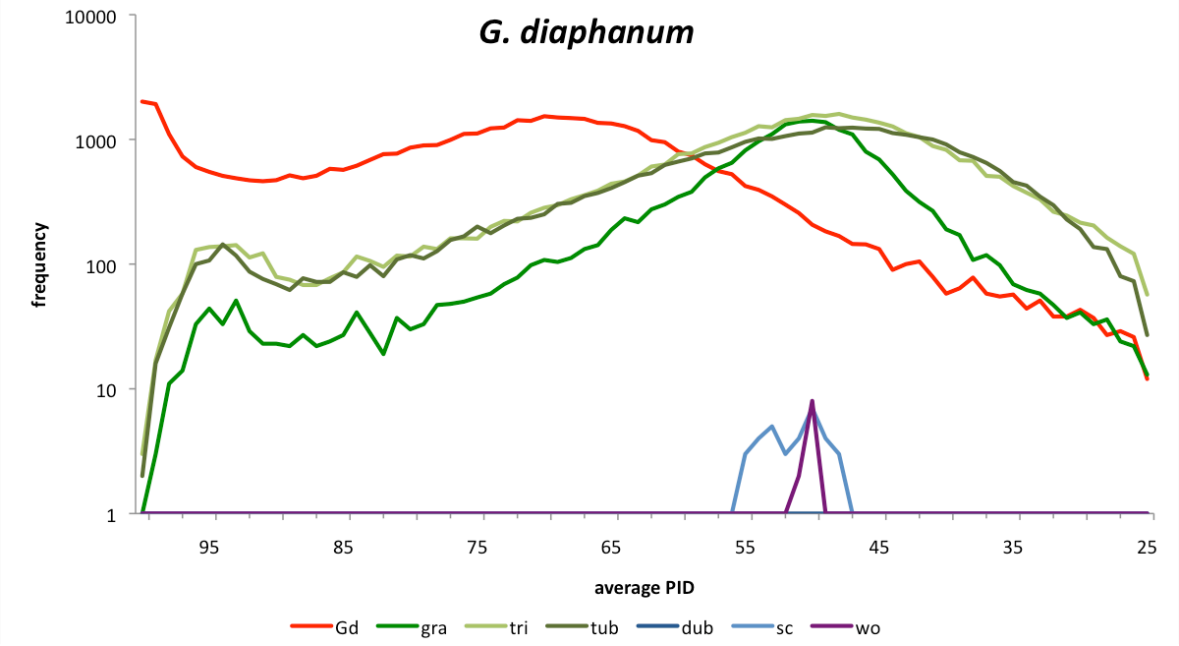


Figure 4-1 Frequencies of similarity clusters compared to average PID per cluster for all *Glomus* and control genomes **a)** *G. irregulare* DAOM 234179 run 2 **b)** *G. diaphanum*.

4.5.2. Single copy markers

A hundred (100) potential single copy markers were selected from the fungal genome alignments. Thirty-six (36) ORFs were selected based on the quality of the alignment, for which PCR primers were developed and amplification was tested in the laboratory. Ten (10) markers amplified successfully and after RT-PCR verification of relative copy number where then sent for pyrosequencing. After analysis of the pyrosequencing data, five markers were retained. Per marker, between 90 and 1123 sequences that fulfilled stringent quality criteria were recovered (table 4-3). All markers were polymorphic, yielding between 2 and 203 alleles. Frequency distributions for all alleles per marker are reported in the supplementary materials. Based on rarefaction analyses (supplementary materials) and the Chao1 index (Chao *et al.* 2005), genetic diversity was at least partially sampled exhaustively for three markers, 1287, 2150 and 628. Only markers 628 and 663 amplified successfully for all strains (*G. irregulare* DAOM 197198, *G. irregulare* DAOM 234328 and *G. diaphanum*). Markers were annotated applying three psi-Blast iterations with a minimum cutoff at 10E-25 (table 4-4).

Table 4-3 Amplification and genetic diversity of single copy markers

marker name	all					*Gi* A[1]		*Gi* B[2]		*Gd*		final alignment (bp)
	seqs	alleles[3]	Chao1[4]	Chao1 lci	Chao1 hci	seqs	alleles[3]	seqs	alleles[3]	seqs	alleles[3]	
1287	299	4(3)	7	4	28	x	x	299	4(3)	x	x	68
1631	556	17(15)	122	56	298	3	1	553	16(15)	x	x	99
2150	768	2	2	2	2	768	2	x	x	x	x	21
628	90	11(8)	39	18	116	13	7(6)	14	3(1)	63	3(2)	175
663	1123	203 (103)	294	258	352	247	80 (61)	448	114 (81)	428	88 (65)	197

[1] *G. irregulare* DAOM 197198; [2] *G. irregulare* DAOM 234328

[3] brackets indicate subset of total number of alleles that occur only once in the dataset

[4] Chao *et al.* 2005

'x': no amplification

Table 4-4 Psi Blast results for single copy markers

marker	psi-Blast result
1287	RNAse L inhibitor-type ATP binding cassette protein
1631	actin-related protein (ARP) 2/3 complex subunit 2
2150	Inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase
628	40S ribosomal protein S2
663	translation elongation factor EF1-α

4.6. Discussion

In this study, we demonstrate that AMF show unique patterns of genetic variation that can best be explained by the presence of genetically differentiated nuclei within the same cytoplasm. This intra-organismal genetic heterogeneity (IGH) can be described as a pangenome, for which analyses of genetic variation that are based on the assumption of genome homogeneity within the multicellular individual are not appropriate. Here, we propose an approach to describe genetic variation in a pangenome. Our method is based on two independent but complementary analyses comprising (i) a description of the levels of polymorphism among single copy genes and (ii) a comparative network-based clustering approach based on pyrosequencing data.

We developed and pyrosequenced five novel single copy markers, which represent segments of protein-coding sequences. When a single copy gene shows polymorphism within a DNA sample from the same isolate, this strongly suggests that this marker is variable between nuclei, since each nucleus harbors its own copy of the gene. We report high amounts of allele diversity within single copy markers, varying from 2 to 203 different alleles when all alleles are counted and from 2 to 103 different alleles if singletons are excluded (table 4-3 and figure 4-5). These allele counts are most likely underestimates of the total allele diversity that can be found in the isolates used in this analysis, for two reasons. First, only three of the rarefaction analyses that estimated allele diversity for the three *Glomus* isolates reached a plateau of diversity, as defined by the lower confidence interval of the Chao1 index of diversity (markers 1287, 2150 and 628). Second, quality checks were extremely stringent, and reads differing by one basepair were clustered together. This conservative approach increases the chance of lumping two different alleles together, but is necessary to avoid the inflated diversity counts that are commonly associated with pyrosequencing technology (Huse *et al.* 2010).

Our observations reveal extensive genomic heterogeneity in *Glomus* and join previous studies in a growing body of evidence for genetic differentiation of AMF nuclei. Judging from our single copy markers, differentiation is not homogenous between loci: some markers yielded relatively few alleles (marker 2150), whereas others revealed hundreds of different alleles that occurred more than once in the dataset (marker 663). This difference might be due to differential selection on the loci themselves or on the adjacent genome regions of the single copy markers. For example, the closest psi-BLAST hit to marker 2150, inositol hexakisphosphate and diphosphoinositol-pentakisphosphate kinase, plays a role in the production of phytic acid, which is the principal storage form of phosphorus in many plant tissues (Ockenden *et al.* 2004). This gene has never been described in AMF and might play an important role in AMF metabolism. In contrast, the closest psi-BLAST hit to the highly polymorphic marker 663 is elongation factor EF1-alpha, which plays a role in translation, and for which multiple copies and pseudogenes are reported in humans (Opdenakker *et al.* 1987; Madsen *et al.* 1990), ciliates (Bergemann *et al.* 1996) and spiders (Hedin and Maddison 2001). Alternatively, variation between markers could be explained by genetic drift (chapter 3, this thesis).

By means of the clustering analysis, we expanded our analysis of IGH in *Glomus* to random fragments of each isolate's pangenome. Genetic diversity in the *Glomus* pangenome was compared to that of six control fungal genomes for which the same pyrosequencing runs were simulated. These comparisons were based on the same amount of reads, with the same average reads length and standard deviation. We observed that clusters from *Glomus* genomes have higher average PIB (table 4-2) than clusters of reads from control fungi. Furthermore, the distributions of sequence identities for *Glomus* and simulated genomes (figure 4-1) show that *Glomus* genomes generally have more clusters with higher sequence identities (PID). The excess of 'strongly similar' regions (PID > 95%) in *Glomus* isolates suggests that the genetic organization in *Glomus* differs from that in the

control fungi. We propose that *Glomus* nuclei are not only genetically variable for some markers, but also partly clonal. In this case, conserved genome regions in *Glomus* would be characterized both by a high average PID value and high clustering coefficient. Indeed, for the same average PID, clusters from *Glomus* genomes have higher average clustering coefficients than clusters from the control genomes (figure 4-4). Thus, nuclei in one *Glomus* isolate contain more highly similar regions, and a greater redundancy than the control fungi.

Importantly, in the distributions of sequence identities, simulated reads from larger (gra, tri, tub) and smaller (dub, sc, wo1) genomes showed markedly different behavior. For the small control genomes, all simulated reads resolve into a single cluster (a contig that effectively becomes a genome assembly) as the sequencing coverage increases (table 4-2). For the large control genomes and *Glomus*, the reads show no sign towards resolution into a single contig. This suggests that each Glomus isolate harbors a population of genomes, with a total number of nucleotides at least as large as that of the large control genomes. This collection of genomes can be viewed as a pangenome, and its minimal length can be computed. The *Glomus* pangenome size that we deduced from our clustering analyses is *circa* 178 Mb for *G. irregulare* and 163.7 Mb for *G. diaphanum*. These estimates are closer to the upper limit of the *G. irregulare* genome size that has been published so far, 150 Mb (Martin *et al.* 2008; Sędzielewska *et al.* 2011), than to a previously published lower estimate of 15 Mb (Hijri and Sanders 2004). An easy way to show whether the *Glomus* pangenome is even larger than the size of large fungal genomes is by increasing pyrosequencing coverage for a given isolate. We predict that large fungal genomes will all group into a single cluster. In contrast, the *Glomus* pangenome is impossible to collapse, since variation between genomes cannot be summarized in a single genome sequence.

Glomus genomes are genetically differentiated *and* interdependent, which justifies the term ‘pangenome’. Genetically differentiated nuclei in AMF are

functionally integrated and the population of genetically differentiated nuclei within the AMF cytoplasm acts synergistically to produce the *Glomus* phenotype. There are four observations that support this interpretation. First, at no point in their life cycle are AMF reduced to a single genome (Jany and Pawlowska 2010; Marleau *et al.* 2011, chapter 3, this thesis). Second, *Glomus irregulare* spores do not germinate under a certain number of nuclei per spore (which is roughly 65 nuclei for G. irregulare, Marleau *et al.* 2011). Third, genetic polymorphism is expressed in the transcriptome (Boon *et al.* 2010; Tisserant *et al.* 2012). Fourth, the high amounts of genetic variation in AMF isolates plays a role in the ability of AMF to adapt to a wide range of host plants (Angelard *et al.* 2010).

There is an important consequence to the pangenome organization we propose for the *Glomus* isolates in this study: the definition of the 'individual' in *Glomus*. The ubiquity of genetic homogeneity as a defining aspect of individuality has already been criticized elsewhere (Santelices 1999; Pineda-Krch and Lehtila 2004a; Pepper and Herron 2008; Folse and Roughgarden 2010). Here, we argue that IGH can be used as an indicator for the possible presence of a 'superorganism', a term which we propose to describe a biological individual that can only survive with a population of differentiated genomes in the cytoplasm. Here, IGH is the condition *sine qua non* to the survival of this population of differentiated genomes. The term 'superorganism' is meant to provide a contrast to individuals in which IGH is present, but not essential to its survival. In other words, it would live with a genetically homogenous nuclear content. This distinction has the advantage of providing a simple test for the importance of IGH in an individual: if it can survive in a genetically homogenous state, it is not a superorganism. In *Glomus*, spores that contain too little nuclei (and thus too little genomic diversity) do not germinate. This is a strong indication that *Glomus* is a superorganism.

An important corollary of the status of *Glomus* as a superorganism is that the genome diversity inside its cytoplasm cannot be described with a single

genome sequence. The *Glomus* genome can only be correctly described as a pangenome, and should be studied as such. A particularly exciting avenue of investigation into the functioning of *Glomus* as a superorganism is the role of anastomosis in the maintenance of genetic variation, which could be tested by a Multi Locus Sequence Typing (MLST) approach on multiple *Glomus* isolates. This approach will make it possible to study linkage and recombination without a priori assumptions on the level of selection or the unit of evolution. Microbial ecology has already developed many metagenomics tools that can be used to study evolution and function of a community of which we cannot trace the individual actors (Maiden *et al.* 1998; Handelsman 2004; Halary *et al.* 2009; Schloss *et al.* 2009; Huson *et al.* 2011; Shade and Handelsman 2011). We hope that this present study will put *Glomus* students on the same road.

4.7. Supplementary material

4.7.1. Pyrosequencing simulations

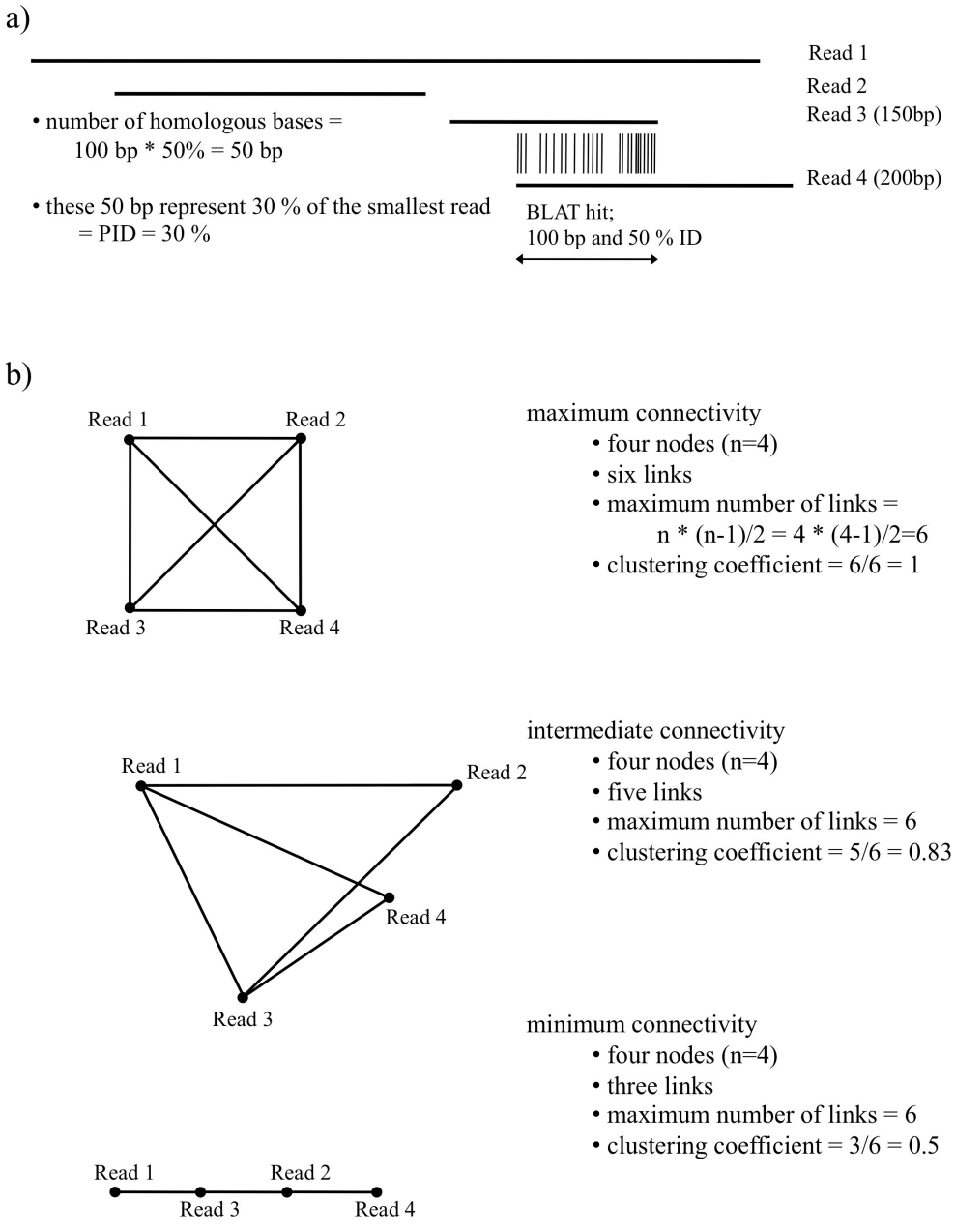


Figure 4-2 Two variables employed to describe similarity clusters, **a)** the PID is the percentage of identical positions on the shortest sequence of an aligned reads pair and **b)** the clustering coefficient, which is the number of aligned reads pairs in the cluster, divided by the maximum number of possible pairs.

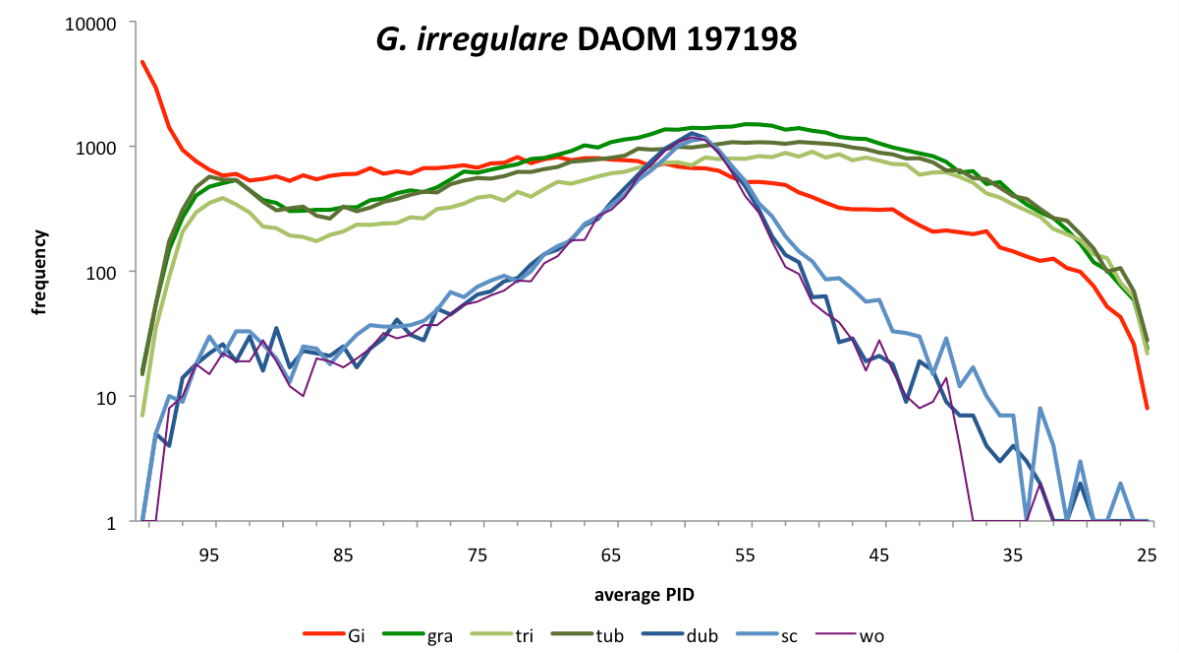
Table 4-5 Parameters for *metasim* genome simulations

strain	# reads from pyrosequencing run	mean length of reads (bp)	standard deviation of reads (bp)	# flows[1]	flow cycles[2]
G. irregulare DAOM197198	398817	216	73	2392	598
G. irregulare DAOM 234179 run 1	485491	365	145	4585	1146
G. irregulare DAOM 234179 run 2	639222	371	142	4525	1131
G. diaphanum	1078190	336	128	4083	1021

[1]flows=3*(mean+(8*stddev)

[2]#cycles=(3*(mean+(8*stddev))/4

a)



b)

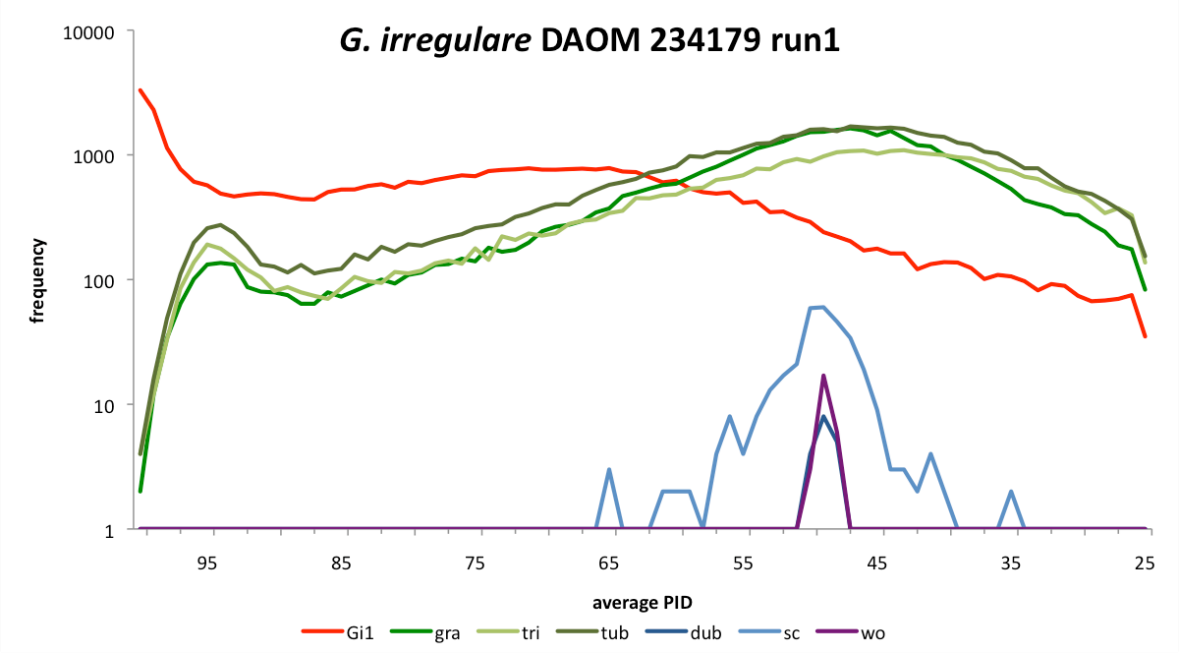
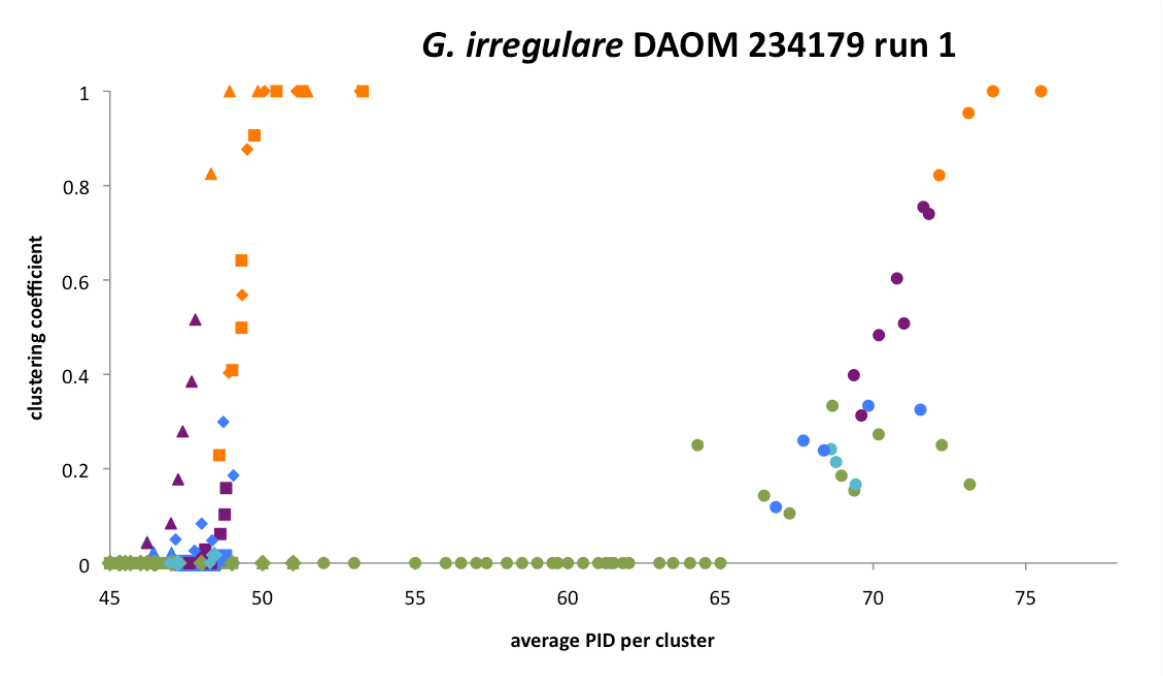
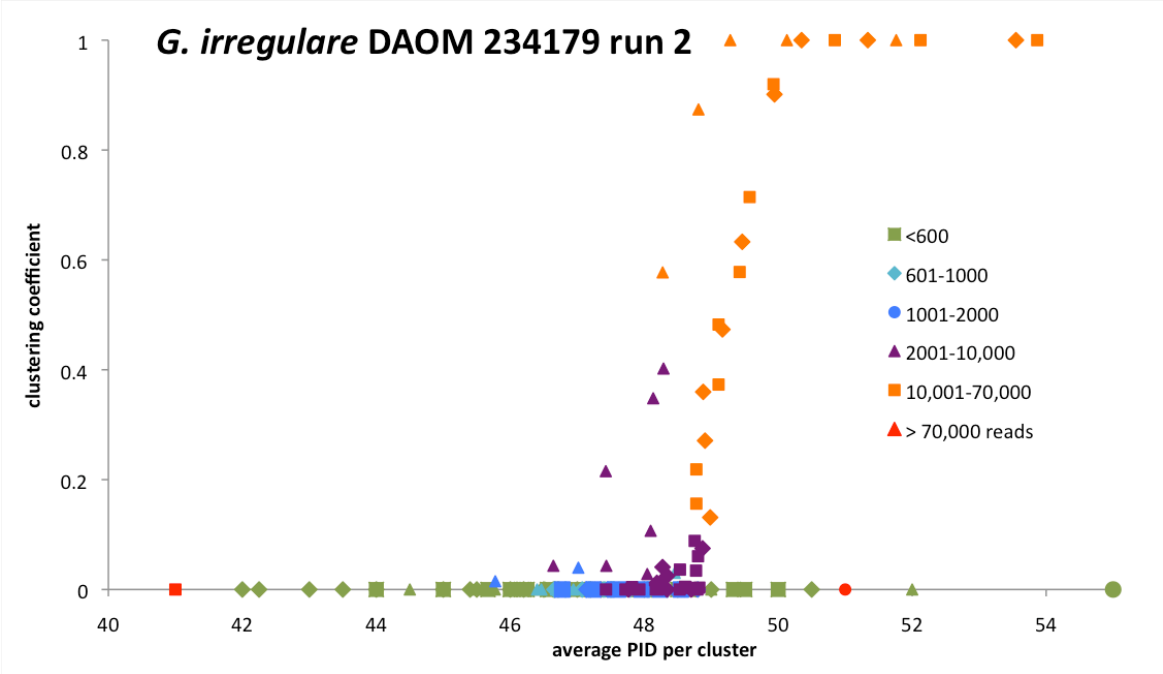


Figure 4-3 Frequencies of similarity clusters compared to average PID per cluster for *Glomus* and control genomes **a)** *Glomus irregulare* DAOM 197198 and **b)** *G. irregulare* DAOM 234179 run 1.

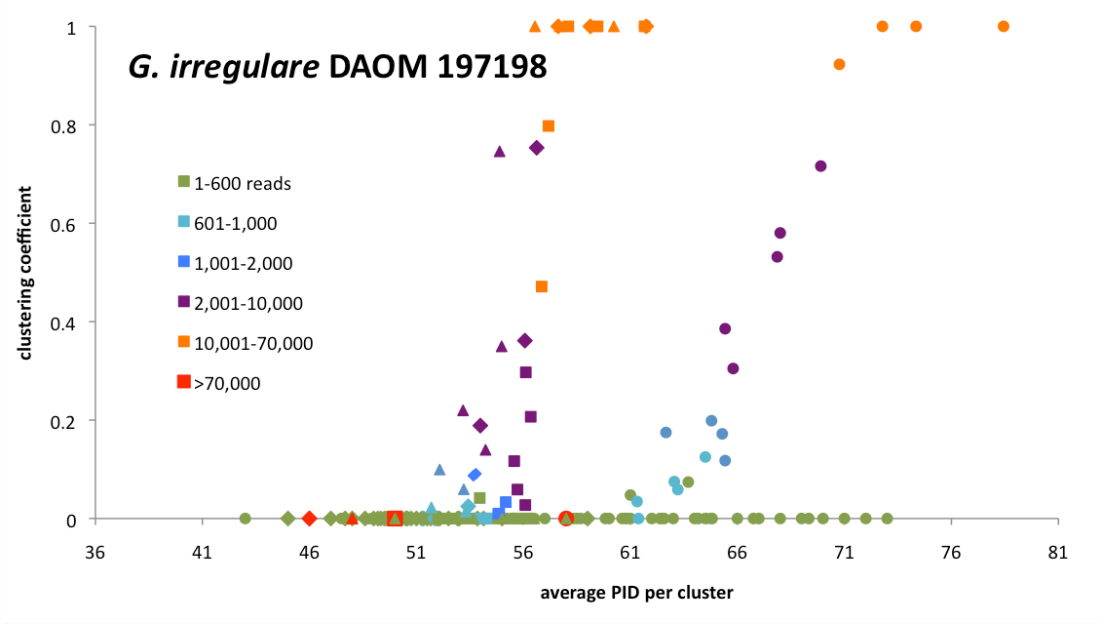
a)


G. irregulare DAOM 234179 run 1
clustering coefficient
average PID per cluster
0
0.2
0.4
0.6
0.8
1
45
50
55
60
65
70
75

b)


G. irregulare DAOM 234179 run 2
clustering coefficient
average PID per cluster
<600
601-1000
1001-2000
2001-10,000
10,001-70,000
> 70,000 reads
0
0.2
0.4
0.6
0.8
1
40
42
44
46
48
50
52
54

c)



d)

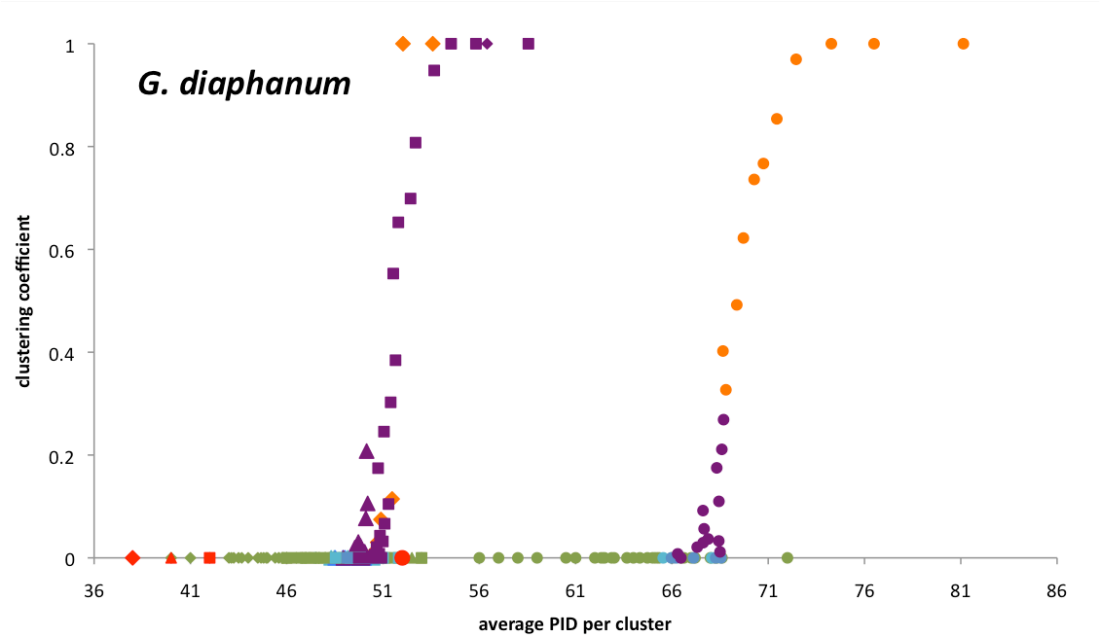


Figure 4-4 Scatterplot between PID and clustering coefficient, for all *Glomus* and control genomes **a)** *G. irregulare* DAOM 234179 run 1 **b)** *G. irregulare* DAOM 234179 run 2 **c)** *G. irregulare* DAOM 197198 **d)** *G. diaphanum*. Colors correspond to the number of reads that are represented by the respective datapoints (see legendas: for simplicity only one agenda is depicted per page). The shape of the datapoint refers to the reads set: circle, *Glomus* (corresponding to the respective graphs); square, *P. graminis*; diamond, *T. melanosporum*; triangle, *P. triticina.*

4.7.2. Single copy markers

Aligning 14 fungal genomes

Table 4-6 Fungal species used for initial alignment of genomes

species	phylum
Aspergillus fumigatus	Ascomycete
Aspergillus nidulans	Ascomycete
Aspergillus niger	Ascomycete
Aspergillus oryzae	Ascomycete
Candida glabrata	Ascomycete
Debaryomyces hansenii	Ascomycete
Kluyveromyces lactis	Ascomycete
Magnaporthe grisea	Ascomycete
Neurospora crassa	Ascomycete
Saccharomyces cerevisiae	Ascomycete
Schizosaccharomyces pombe	Ascomycete
Yarrowia lipolytica	Ascomycete
Asbya gosipii	Ascomycota
Cryptococcus neoformans	Basidomycete
Encephalitozoon cuniculi	Microsporidia

PCR amplification

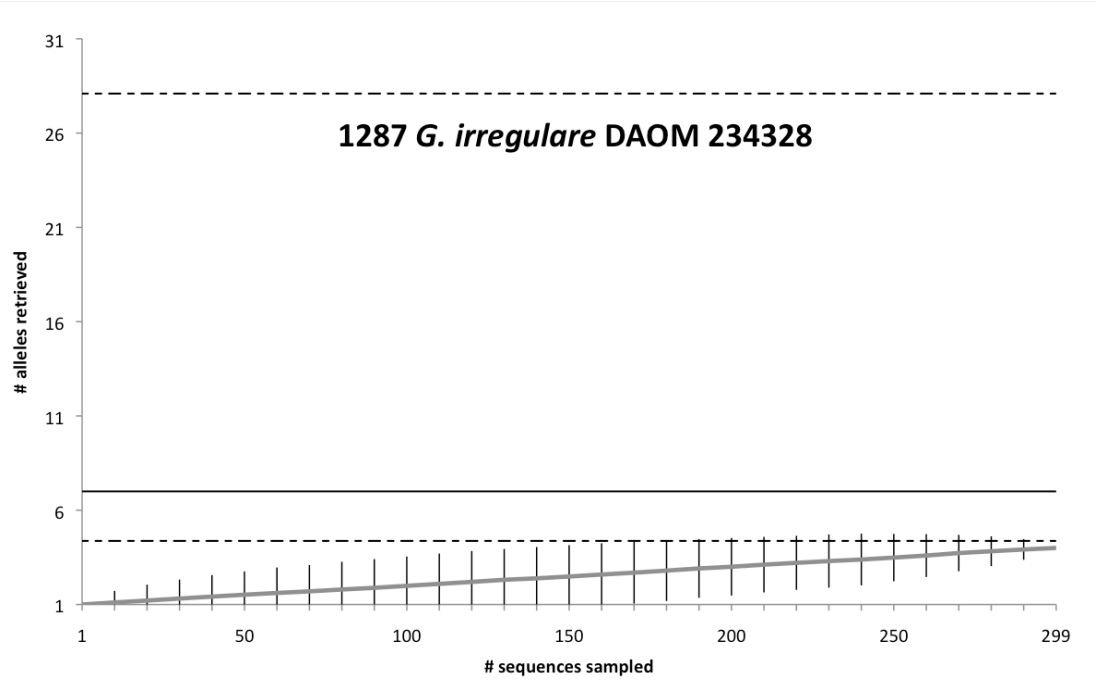
Potential single copy marker sequences were amplified using Dream*Taq* DNA polymerase (Fermentas) using primers listed in table 4-7. DNA from the strains *G. irregulare* DAOM 197198, *G. irregulare* DAOM 234328 and *G. diaphanum* was extracted as described in Boon *et al.* (2010) The reaction was performed in 50 *µ*l volumes containing 1 ng DNA, 1,25 U Taq polymerase (Fermentas), 0.2 mM dNTPs, 0.4 *µ*M of each primer and the PCR buffer. PCR was carried out for 40 cycles (94°C for 30 sec, Ta (annealing temperature, see supplementary materials for Ta per primer pair) for 30 sec, 72°C for 1 min; preceded by an initial 3 min denaturation at 95°C and followed by a 10 min hold at 72°C, on a Mastercycler EPgradient S (Eppendorf).

Table 4-7 Primer sequence, annealing temperatures and amplicon length for five single copy markers

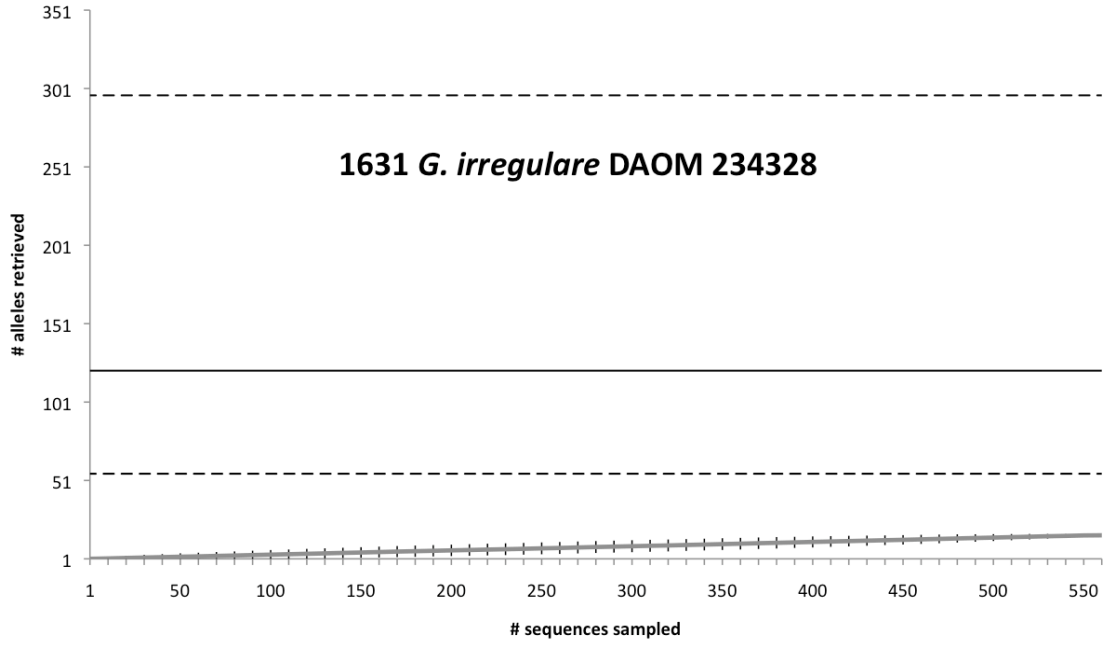
marker	direction	Primer sequence (5'-3')	Annealing temperature (°C)	Length (bp)
1287	forward	CTGAGCGGCGGCGAACTGCAGC	49	340
	reverse	CAGATCATGTTCCACCAC		
1631	forward	AATGATTTTGTTAGATTATC	49	500
	reverse	GTAAAGATTACAGTGACTCTATCC		
2150	forward	GTCAAGTGGGGTGGTGAG	48	310
	reverse	TTTGCCGCATTACTATCATCA		
628	forward	GTTCCTGTTACGAAACTTGGG	45	470
	reverse	GCCAGTAACCTTGCAAGGAACAGT		
663	forward	ATTGGACATGTAGATCACGGGAAA	48	310
	reverse	ATGTTCACGAGTCTG		

Rarefaction curves for single copy markers

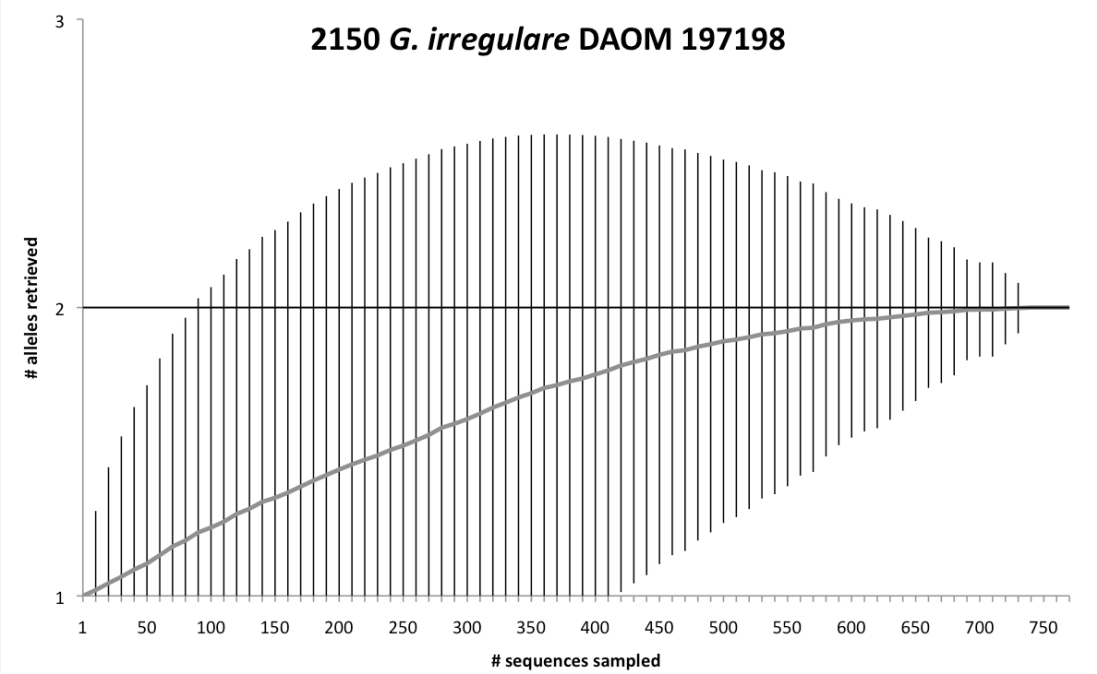
a)



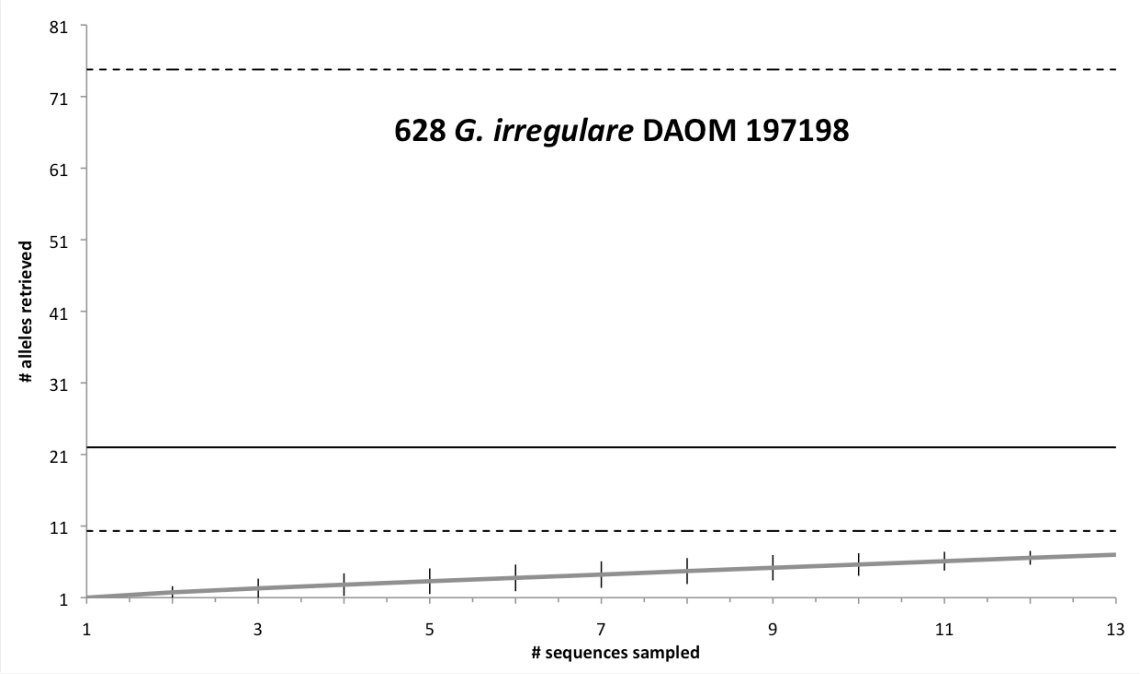
b)



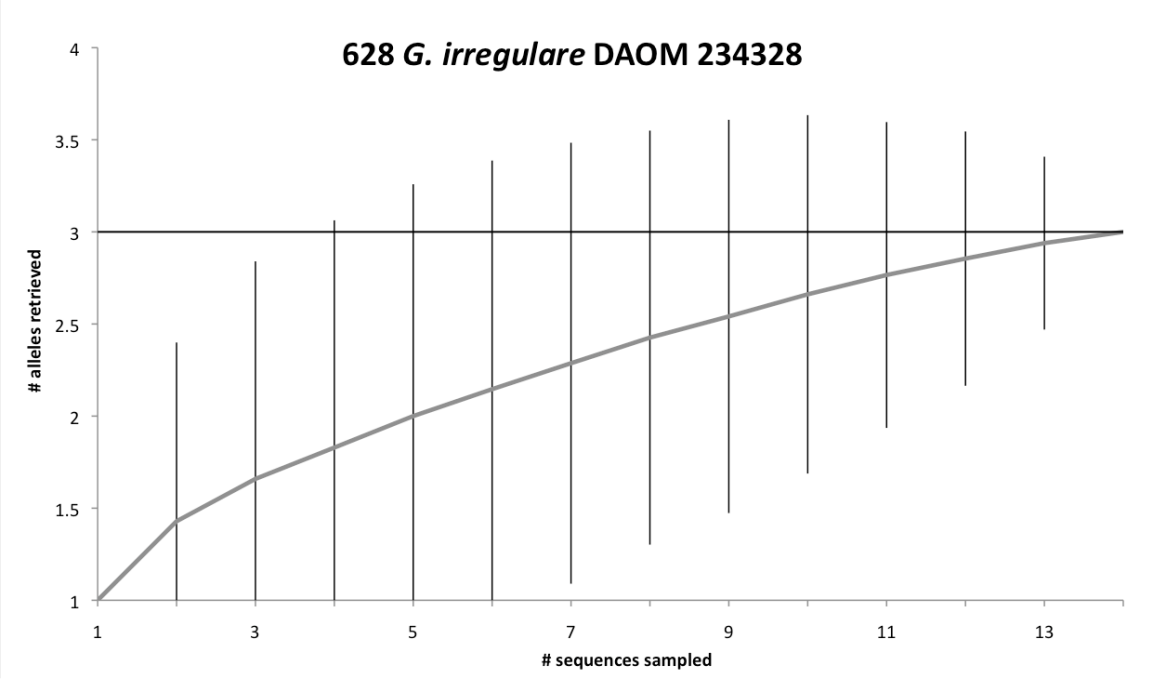
c)


2150 G. irregulare DAOM 197198
alleles retrieved
sequences sampled
3
2
1
1
50
100
150
200
250
300
350
400
450
500
550
600
650
700
750

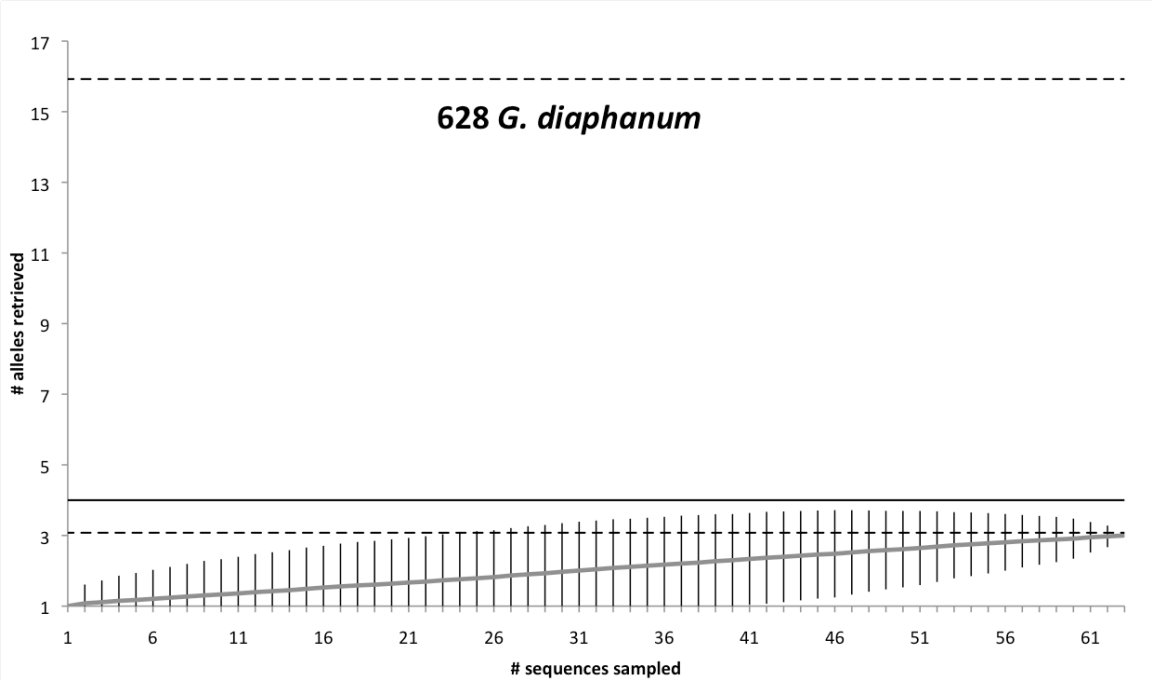
d)


628 G. irregulare DAOM 197198
alleles retrieved
sequences sampled
81
71
61
51
41
31
21
11
1
1
3
5
7
9
11
13

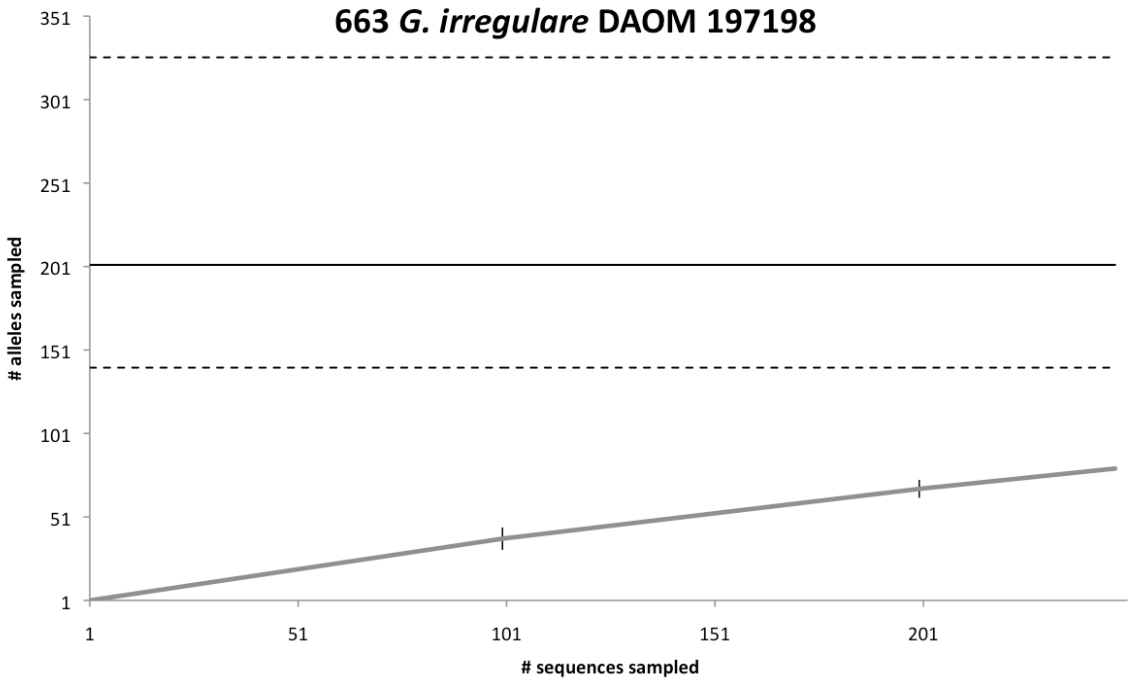
e)


628 G. irregulare DAOM 234328
alleles retrieved
sequences sampled
4
3.5
3
2.5
2
1.5
1
1
3
5
7
9
11
13

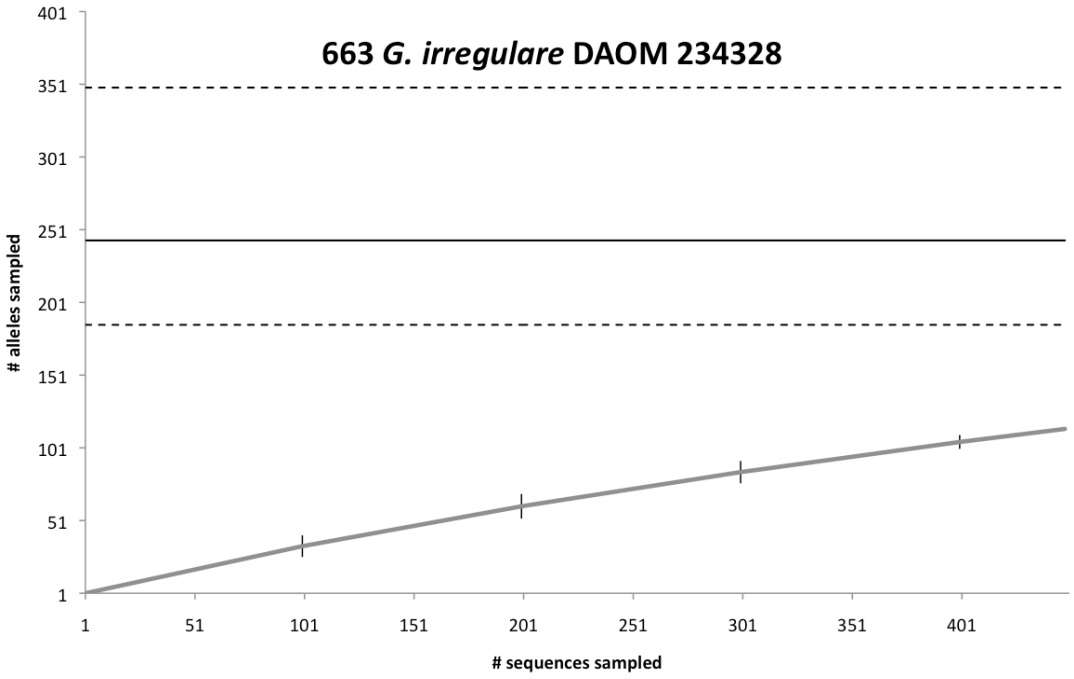
f)


628 G. diaphanum
alleles retrieved
sequences sampled
17
15
13
11
9
7
5
3
1
1
6
11
16
21
26
31
36
41
46
51
56
61

g)


663 G. irregulare DAOM 197198
alleles sampled
sequences sampled
351
301
251
201
151
101
51
1
1
51
101
151
201

h)


663 G. irregulare DAOM 234328
alleles sampled
sequences sampled
401
351
301
251
201
151
101
51
1
1
51
101
151
201
251
301
351
401

i)

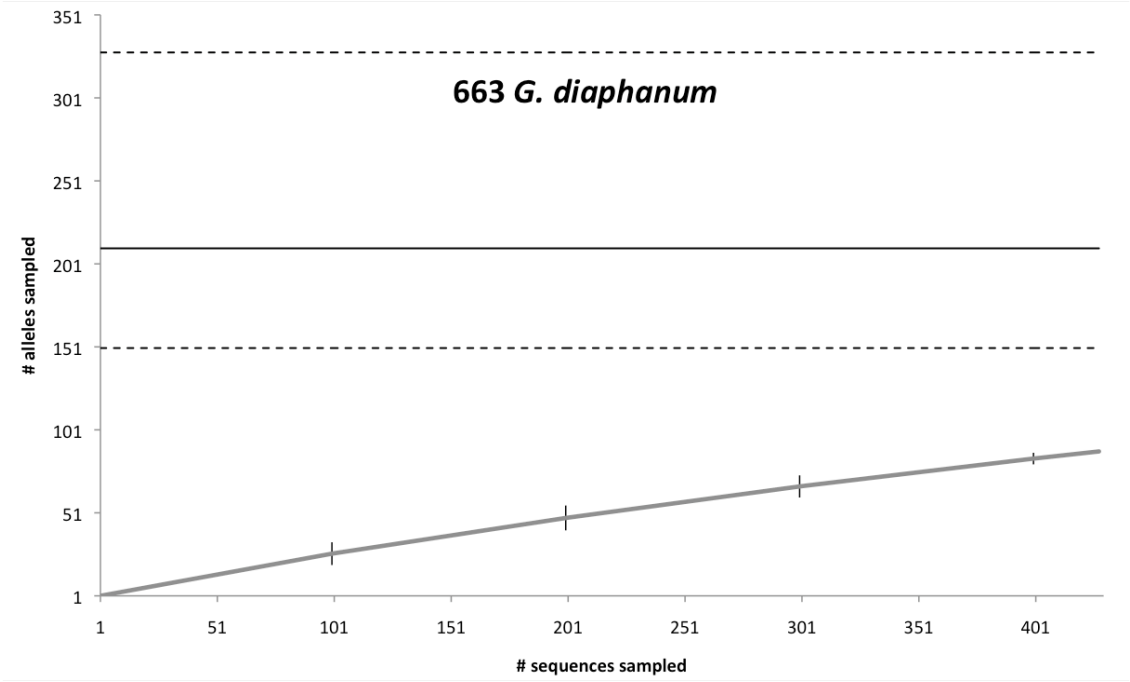
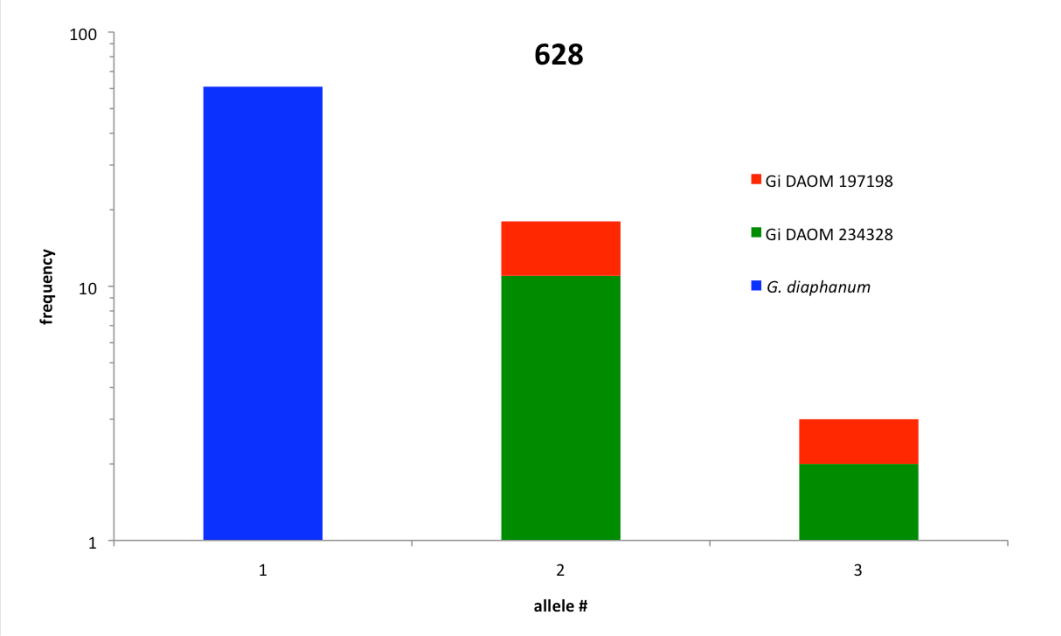


Figure 4-5 Rarefaction curves for single copy makers **a)** 1287 *G. irregulare* DAOM 234328 **b)** 1631 *G. irregulare* DAOM 234328 **c)** 2150 *G. irregulare* DAOM 197198 **d)** 628 *G. irregulare* DAOM 197198 **e)** 628 *G. irregulare* DAOM 234328 **f)** 628 *G. diaphanum* **g)** 663 *G. irregulare* DAOM 197198 **h)** 663 *G. irregulare* DAOM 234328 **i**) 663 *G. diaphanum*. The number of recovered alleles (y axis, grey line, 95% confidence intervals indicated by vertical lines) is compared to the maximum Chao1 value (Chao *et al.* 2005), which is the estimated minimum richness for each group (solid black line, 95% confidence intervals in dotted lines).

Distribution of alleles for single copy markers

a)



b)

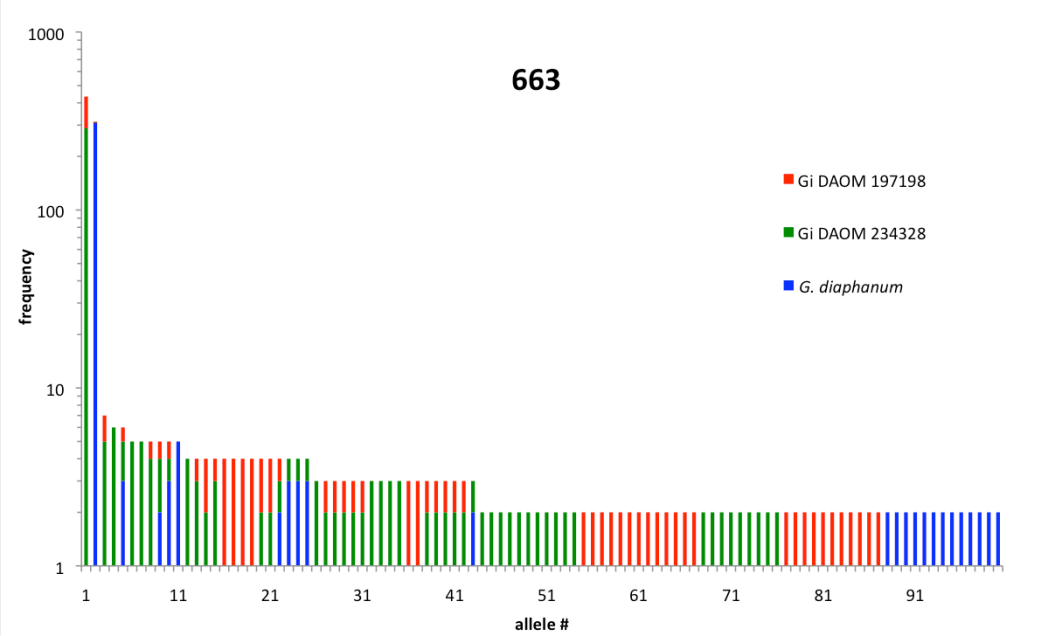


Figure 4-6 Allele distributions for single copy markers that have more than two alleles which occur at frequencies higher than one **a)** 628 and **b)** 663

5. Discussion

In this final chapter, I summarize the results of my thesis (section 5.1). Subsequently, I argue that the peculiar genome organization of AMF warrants special care when generating data and interpreting results in several fields of AMF biology, after discussing the pitfalls of studying genetic variation in multigenomic organisms (5.2). I will place my results in the larger context of AMF biology by sketching what we know on how genetic variation is generated, lost and maintained (5.3). This chapter finishes with a discussion of individuality and selection in AMF, and the special place AMF occupy in evolutionary theory (5.4).

5.1. Summary of results

The present thesis has confirmed the extensive intra-isolate polymorphism that was previously observed for LSU rRNA encoding genes (in *G. irregulare*) and the polymerase1-like locus *PLS* (in *G. etunicatum*). This thesis is the first to report that this polymorphism is transcribed. There are no indications that the transcribed sequences cannot be translated in functional proteins: very few frameshift mutations, stop codons or, in the case of rDNA genes, structural variation was observed. This thesis is also the first to report a bottleneck of genetic variation at sporulation for the *PLS* locus, which suggests an important role for anastomosis in the AMF life cycle. Furthermore, for the very first time, five polymorphic single copy markers have been characterized for *G. irregulare* and *G. diaphanum*. Extensive nucleotide variation between alleles from these markers, from the same isolate, show that in *Glomus* intra-isolate polymorphism is not limited to the LSU rDNA and *PLS* loci. Finally, a network-based clustering analysis based on four pyrosequencing runs from three different AMF isolates, representing two different

species (*G. irregulare* and *G. diaphanum*), shows that these three different isolates show genome wide patterns of diversity that are different from fungi with comparable genome sizes. We interpret these patterns as additional evidence in favor of genome differentiation between nuclei in the same isolate. Thus, the genome organization of at least three isolates from the *Glomus* genus, *Glomus irregulare* isolate DAOM 197198, *Glomus irregulare* isolate DAOM 234179 and *Glomus diaphanum* DAOM 229456, appears to be best described by genetically differentiated genomes co-occurring in the same cytoplasm.

5.1.1. The pitfalls of studying AMF

The peculiar genome organization of AMF warrants special care in the collection and interpretation of nucleotide sequence data. First, I will discuss the implications for the study of AMF genetics. I will then review the specific interpretation difficulties for other fields of AMF biology that use molecular biology techniques.

5.1.2.AMF genetics: a caveat

It is rarely questioned in molecular biology that, when we study DNA from a multi-cellular organism, or more precisely, from a multi-nuclear organism, the genomes from these multiple nuclei are identical. Even though this assumption will never entirely hold due to somatic mutations, in sexually reproducing organisms we can at least rely on the reduction of genome variation at every generation cycle, by the passage through a unicellular phase. Nevertheless, for clonally reproducing organisms the problem of intra-organismal genome heterogeneity (IGH) might be more significant (Pineda-Krch and Lehtila 2004). In the case of AMF, the question of inter-nuclear variation becomes insurmountable. If we want to study intra-isolate

variation in AMF, it is imperative to address the implicit assumptions of most standard laboratory and analysis techniques.

Firstly, there is the DNA extraction step. When DNA is isolated from spores or hyphae of a fungal isolate, we end up with a mixture of many different genomes. Unfortunately, it is quite hard to isolate individual nuclei and extract their DNA. AMF nuclei are very small and contain only between 0,18 and 1,08 picogram of DNA per nucleus (investigated for 11 AMF species; Bianciotto and Bonfante 1992; Hosny *et al.* 1998). There are solutions to this difficulty. Recent technological advances now allow for the amplification of a few femtograms of DNA in a bacterium into micrograms of high molecular weight DNA suitable for DNA library construction and Sanger sequencing (Raghunathan et al. 2005; Lasken 2007). Another possibility is to amplify DNA from a single spore, which is quite feasible for AMF (Pawlowska and Taylor 2004, chapter 3, this thesis). Considering that the number of nuclei in a spore is roughly known, at least this would give an idea about the maximum number of expected alleles in a sample, if we also know the copy number of the locus in question.

The next challenge is the polymerase chain reaction (PCR). AMF DNA extracted from spores and hyphae from a single isolate contains a mixed population of sequences that originate from different genomes. Due to the stochastic nature of the PCR reaction, and consequent sampling effects especially at low DNA concentrations (Gill *et al.* 2000; Butler and Hill 2010), it is quite possible that the genetic diversity of amplified alleles after a PCR reaction does not represent the original genetic diversity. Nevertheless, this problem is well documented for PCR reactions and is not a unique complication to the study of AMF. A second potential problem with PCR reactions in AMF could be that chimerical sequences might easily arise between homologous sequences. Again, this problem has long been documented for genomic (Meyerhans *et al.* 1990) and cDNA sequences (Brakenhoff *et al.* 1991) and several bioinformatics approaches

have been developed to detect chimerical sequences (Komatsoulis and Waterman 1997; Huber *et al.* 2004; Gonzalez *et al.* 2005; Rosendahl 2008). However, the problem of chimerical sequences generated by PCR reactions is more serious when the locus in question occurs in multiple copies (Wang and Wang 1996, 1997), as is the case in AMF. In addition, in AMF PCR-generated chimeras can also occur between alleles from single-copy markers, originating from different nuclei. Thus, the genetic diversity we find after a PCR reaction might not be the genetic diversity originally present in the sample. There is only a partial solution to this problem, which is choosing a high fidelity polymerase (Bradley and Hillis 1997).

The next technique under scrutiny is RT-PCR, which is employed to measure (relative) DNA concentrations, for example to estimate gene copy number. However, gene copy number estimates are actually averages over all nuclei within the same cytoplasm. It is not possible to investigate the distribution of copy number variation *between* nuclei within an AMF isolate. Therefore we cannot differentiate between a homogeneous population with two gene copies per locus and, say, a heterogeneous population in which 60% of the genomes contain the same locus in single copy, 20% contains two copies of that locus and 20% does not contain the gene at all. This is not a purely hypothetical distinction: gene copy number can vary between isolates, which has been shown using RT-PCR for the protein-coding genes *BiP* and for ribosomal genes (Corradi *et al.* 2007) and the ITS2 region (Kuhn *et al.* 2001) using Fluorescent In-Situ Hybridization (FISH) techniques). It is not inconceivable that gene copy number in AMF similarly varies between nuclei, especially if anastomosis plays an important role in AMF life history.

Following a similar logic, it is also very difficult to unambiguously determine genome ploidy for AMF. *Glomus intraradices* (later renamed to *G. irregulare* (Stockinger *et al.* 2009) and after that renamed to *Rhizophagus irregularis* by the same authors (Krüger *et al.* 2012)) was found to be haploid (Hijri and Sanders 2004). Nevertheless, two of the approaches used to determine ploidy levels in this study

(i.e. reassociation kinetics and genomic reconstruction) do not investigate these values on the level of a nucleus. That is, the values we obtain in these measurements are averages over multiple nuclei and give no information about the distribution of genetic variation between nuclei. Theoretically, it is possible that, say, 50% of the nuclei miss half their (haploid) genome, 25% of the nuclei are haploid and 25% of the nuclei in AMF are diploid. Still, average DNA length per nucleus was estimated to be 14.07±3.52 Mb using flow cytometry, which is a technique that measures DNA content *per nucleus*. The standard error is not sufficiently large for ploidy level to differ between nuclei, although sufficiently large for the smallest amount of DNA (10.55Mb) and the largest amount of DNA (17.59Mb) to differ by a factor of 2/3.

Not all massive parallel sequencing approaches are appropriate for studying intra-isolate variation in AMF. In chapters 3 and 4 of this thesis, for example, we found the error rates of 454 pyrosequencing techniques can approach the levels of genetic variation that are found in the *PLS* and single copy gene loci. This difficulty can be circumvented by stringent error checks. However, since it is hard to distinguish low frequency alleles and recombination from sequencing artifacts, the optimal approach would be to increase coverage to such an extend that all alleles are retrieved several times. Alternatively, we could choose to use a massive parallel sequencing approach that is less prone to amplification error, such as the single molecule technology of the RS sequencer (Pacific Biosciences).

Another problematic massive parallel sequencing approach is shotgun sequencing, because this technique relies on shearing the genome(s) in many small fragments to subsequently sequence and align them into scaffolds. Unfortunately, small fragments fail to align when there are many (low complexity) repetitive sequences, as has been previously reported in AMF (Zézé *et al.* 1996; Hijri and Sanders 2004, Yves Terrat, personal communication). Furthermore, many genes studied so far in AMF (both protein-coding and ribosomal) seem to occur in

multiple copies within AMF genomes as a paralogue or orthologue (Kuhn *et al.* 2001; Corradi *et al.* 2004a; Corradi *et al.* 2004b; Corradi and Sanders 2006; Corradi *et al.* 2007).

Finally, there are several challenges for AMF genome assembly. First, high genetic variation causes major problems in assigning reads to contigs and contigs to scaffolds. A second difficulty, which does not depend on AMF genome organization, is that there are hardly any genomic sequences for *Glomus* to construct the scaffold on: *Glomus* sequences are too divergent from other eukaryotes or even from other fungi to reliably use the latter as a scaffold for a genome sequencing project. A third problem is posed by contamination, since the obligate nature of the symbiosis between AMF and plants prevents the culturing of AMF without their host. The Glomus Genome Consortium has encountered all the above problems in their efforts to produce a draft sequence for the *Glomus irregulare* genome (M. Hijri, personal communication) and is now tackling the problem with a combined fosmid/EST/454/Shotgun/Illumina sequencing approach (Martin *et al.* 2008 and personal communication). After having sequenced roughly 345 Mb, the estimated sequence depth is somewhere between 2 and 3. As the title of the latest consortium paper states, it will be a 'long hard road to a completed *Glomus* genome'.

Apart from the technical difficulties in obtaining sequence data and gene or genome information from AMF, it is also important to be prudent in the interpretation of results. In the following section, I will mention a few potential difficulties in the application of AMF nucleotide sequence information to questions in ecology, taxonomy and phylogenetics.

5.1.3. Interpreting AMF sequence data in ecology

Most of what we know of AMF biology stems from isolates that are grown in a laboratory environment. Even though the bi-compartmentalized *in vitro* system that was proposed by St-Arnaud *et al.* (1995) has done much to facilitate work on these obligate symbionts, fieldwork on AMF is still much hindered by the difficulty of sampling interlacing networks of multigenomic fungi.

The first ecological studies of AMF were based on spore morphology. However, taxon determination based on spore morphology is not always consistent between studies due to the specialized nature of this knowledge, where only a few experts are qualified to assign spores to taxons. Subsequent studies were based on isozyme (Rosendahl and Sen 1992; Dodd *et al.* 1996) and rDNA sequences (Gollotte *et al.* 2004; Rosendahl and Stukenbrock 2004; Öpik *et al.* 2006 and references therein). It is hard to compare morphological data to more recent studies based on molecular data (Rosendahl 2008), since taxon definitions obtained by the two identification methods are not always concordant. Another problem with ecological studies based on rDNA sequences is that they cannot distinguish variation *within* a hyphal system from variation *between* hyphal systems. Considering that sequences within the same hyphal system can be more divergent than sequences from different hyphal systems in the same field (Croll *et al.* 2008b) or even from different continents (Rodriguez *et al.* 2004; Boon *et al.* 2010), this could be potentially problematic.

Further ecological studies were undertaken using co-dominant markers such as variable introns (Stukenbrock and Rosendahl 2005a, b; Croll *et al.* 2008b) and microsatellites (Croll *et al.* 2008a; Croll *et al.* 2008b). Finally, mitochondrial markers were also employed (Croll *et al.* 2008b). These markers do not distinguish variation within a hyphal system from variation between hyphal systems either. This is important, since in fact most hyphal systems within the same field readily anastomose, as experiments that were based on cultures from the same species

that were isolated from the same sampling sites have shown (Croll 2009). This raises the question of whether we can actually distinguish individuals (or OTUs) in the same field, or whether the potential of hyphal systems to fuse should classify them as a single organism. This question will be explored in more detail in section 5.4.

Another approach is to conduct ecological studies on AMF using massive parallel sequencing technologies. An advantage of this approach is that, in contrast to previously described techniques, it becomes possible to exhaustively sample all genetic variation in a sample. This approach resolves the question whether we are dealing with a representative subset of variation from the population and bypasses the problems that surround the definition of 'the individual' in AMF. Effectively, by careful experimental design, it becomes possible to address the question of population structure within a site. Pivato *et al.* (2007) found 12 OTUs based on the LSU rDNA, of which only two OTUs corresponded to previously published sequences from known species. An estimated 45-50 sequences were needed to retrieve all 12 OTUs from the soil sample. Opik *et al.* (2009) found a high number of taxa that differentially colonized generalist and specialist forest species. Taxa were based on SSU rDNA and defined by comparison to a reference database (Öpik *et al.* 2010). Sequences that diverged more than 97% from known taxa were designated 'molecular operational taxonomic unit' (MOTU). Opik *et al.* (2009) managed to exhaustively sample AMF OTUs on some plant species, which testifies to the great potential of their approach.

Irrespective of the technology used, in order to derive meaningful ecological conclusions from AMF nucleotide diversity it is imperative that definitions of the 'individual', OTU or species or are normalized between studies. Ecological studies in AMF are hindered by the difficulty of linking the molecular diversity found in the field to the functional diversity found in the laboratory. As other authors have

already remarked (Van der Heijden and Scheublin 2007), the question remains whether molecular diversity corresponds to functional diversity in AMF.

5.1.4. Interpreting AMF sequence data in taxonomy

Glomeromycota are difficult to classify taxonomically. Originally, spore morphology was used to describe the microscopic members of this phylum, as mentioned in the previous section. Unfortunately, taxonomy practices based on spore morphology require a lot of time and expertise. An more convenient solution offered itself with the onset of molecular biology techniques. These techniques opened the door to faster and more convenient taxonomic assignment and were expected to finally allow researchers to determine the evolutionary relations within AMF, on the assumption that nucleotide sequences of appropriate markers reflect the evolutionary history of a taxon.

The marker of choice was ribosomal DNA. This marker has been widely employed for taxon assignment, even though intra-species and even intra-isolate variation in rDNA sequences is widely documented for many taxa (Thornhill *et al.* 2007; Simon and Weiss 2008) and there is often no discernable upper limit for intraspecific variability (Nilsson *et al.* 2008). As was mentioned in the introduction, chapter 2 and the previous section, rDNA is also variable within AMF species and isolates and the use of ribosomal markers to identify AMF taxa is not straightforward. A single allele cannot be representative of a taxon, and the assignment of ribotypes to particular species should be interpreted as a mere approximation that facilitates comparison with other studies. Huge rDNA genetic variation within AMF (see, for example, chapter 2) makes it impossible to resolve taxa if variation is not sampled exhaustively. As of yet, including this thesis, rDNA variation has not been sampled exhaustively even within a single isolate.

Therefore, apart from the theoretical objection that a sequence is not a taxon, taxon assignments cannot be confidently based on rDNA sequence only.

The regular revision of AMF taxonomy reflects the lack of consistent criteria to designate AMF taxa, be it by molecular or morphological criteria (Morton and Benny 1990; Schussler *et al.* 2001a; Schussler *et al.* 2001b; Schwarzott *et al.* 2001; Gamper *et al.* 2009; Krüger *et al.* 2009; Stockinger *et al.* 2009; Sokolski *et al.* 2010; Krüger *et al.* 2012). I will conclude this section by stressing that the only way to confidently delineate taxa in AMF is to apply multiple internally consistent criteria, be they molecular, morphological or ecological. In this way, higher levels of taxonomy might be resolved, if they exist. Anastomosis between AMF hyphal networks might warrant a rethinking of our species definition. This subject will be explored in more depth in section 5.4.

5.1.5. Interpreting AMF sequence data in phylogenetics

For lack of morphological characteristics, taxonomy in AMF is often synonymous with phylogenetics. Thus, at the peril of repeating myself, AMF phylogenetics is plagued by much of the same difficulties as AMF ecology and taxonomy. Nonetheless, the huge amounts of variation found in AMF isolates and populations are less of a problem at higher taxonomic levels, since it is possible to employ less diverse markers for phylogenetic reconstruction. Still, due to their early divergence and lack of closely related groups, the phylogenetic position of the Glomeromycota within the fungal kingdom remains as yet unresolved (James *et al.* 2006; Liu *et al.* 2009; Ebersberger *et al.* 2011).

However, even *within* the phylum Glomeromycota, phylogenetic assignment remain elusive. The frequent revision of AMF taxonomy is a testimony to the difficulty of assigning AMF to particular groups, especially at the genus level (Krüger *et al.* 2012 and references therein). And this level is exactly the level at

which spore morphology becomes inconclusive and intra-isolate variation becomes an issue. The desirability to assign species names to molecular diversity sampled in the field needs to be discussed within the AMF community (Hibbett *et al.* 2011). In addition to this, it is important to realize that we do not have a species definition for AMF, and we do not have a species concept either. It is my opinion that we can only proceed with AMF biology if we have a clear idea of what is an AMF individual, a population or a species. This subject will be explored in more depth in section 5.4. First, it is time to closely investigate what is known on what the effects of genetically differentiated genomes on AMF biology.

5.2. Multiple genomes and AMF biology

How does the extraordinary amount of genetic variation in AMF come about? To answer that question, I will discuss the generation, maintenance and loss of genetic variation in AMF. Much recent work has elucidated important details of AMF life history, and in the following section I will discuss these finds together with my thesis results.

But first, I present a disclaimer. From now on, I will assume the general term 'AM fungi' or 'AMF' when describing the isolates for which genetically differentiated nuclei have been demonstrated, i.e. *Glomus irregulare* DAOM 234328 (chapter 4, this thesis), *G. irregulare* DAOM197198 (Kuhn *et al.* 2001, chapter 2 and 4, this thesis), *G. diaphanum* DAOM 229456 (chapter 4, this thesis) and *G. etunicatum* NPI (Hijri and Sanders 2005, chapter 2 and 3, this thesis). Even though it is not an unreasonable assumption to attribute this genome organization to other members of the phylum, these members have not been investigated. Therefore, unless mentioned otherwise, the term 'AM fungi' or 'AMF' is meant to solely refer to the fungal isolates that have been studied in this thesis or elsewhere and have been demonstrated to contain genetic differentiated between nuclei.

5.2.1. Generation of genetic variation in AMF

A major mechanism for generating genetic variation is nucleotide sequence mutation, i.e. substitution, insertion and deletion. Mutation rates are highly variable within the genome and across a wide range of taxa (Drake *et al.* 1998; Lynch 2007). Are there reasons to assume that mutation rates in AMF are higher than usual? This would depend directly on the stringency of the error checking mechanisms in AMF, and ultimately on AMF life history and on 'evolutionary accidents' (Drake *et al.* 1998). Unfortunately, very little is known about these mechanisms. Judging from the molecular evolution studies of LSU rDNA (chapter 2), *PLS* (chapter 3) and single copy markers (chapter 4) in this thesis, nucleotide polymorphism for these loci is sufficiently low to guarantee sequence conservation in the rDNA stem sections and exons. Also, in the cluster analyses of chapter 4, the AMF groups that were in excess in comparison to the simulated control genomes were those in the highest percentages of idenity (90-100 %, figure 4.1). These observations can point to either low AMF mutation rates, to efficient error checking mechanisms, or to strong balancing selection pressures on the abovementioned loci. In short, the only confident statement we can make is that we have not enough information to make any statements about AMF nucleotide mutation rates.

Of course, there are many other mechanisms that generate genetic variation. Much genetic variation is generated during meiosis. In yeast, for example, mutation rates are 6 to 20 times higher during meiosis than during mitosis (Magni and Von Borstel 1962). The meiotic machinery for AMF seems to be in place (Halary *et al.* 2011) and several studies have reported evidence of recombination (Vandenkoornhuyse *et al.* 2001; Gandolfi *et al.* 2003; Stukenbrock and Rosendahl 2005a; Croll and Sanders 2009; Den Bakker *et al.* 2010). The mechanism of recombination in itself does not directly create more variation in a

population, it merely shuffles the information into new allelic combinations. Indirectly, however, recombination can lead to large insertions or deletions in the genome. Again, little is known about the occurrence and importance of these mechanisms in AMF.

Finally, lateral gene transfer (LGT) can add genetic variation to a genome. Studies detecting LGT in AMF are not straightforward because of poor annotation of AMF sequence data and the ever-present threat of contamination from the plant symbiont (data not shown). There are many transposable elements (TE) in AMF (Gollotte *et al.* 2006 and Yves Terat, personal communication) which, apart from providing additional evidence for recombination (Wright and Finnegan 2001), confirms the possibility of LGT in AMF.

5.2.2. Loss of genetic variation in AMF

The major cause of a loss of genetic variation in AMF is through the bottleneck of genetic variation that occurs at sporulation (Marleau *et al.* 2011). The size of this bottleneck is very important to the rate of loss of genetic variation, as Bever *et al.* (2005; 2008) showed. Their model will be discussed in the next section. Here, I would like to focus on what we know of the size of the bottleneck.

The size of the genetic bottleneck in AMF at sporulation depends on the number of genetically different nuclei that are transferred to the spore. The number of nuclei in AMF spores are highly variable, not only between species and isolates (Bianciotto and Bonfante 1992; Hosny *et al.* 1998), but also within isolates (Marleau *et al.* 2011). Chapter 3 of this thesis is the first study to measure the size of the genetic bottleneck for a genetic marker, *PLS*. The loss of allelic variants is highly variable between spores, as would be expected if the number of nuclei varies between spores. However, all eight spores showed a reduction in allele variants in comparison to the total amount of variation in all 8 spores pooled

together (table 3-1). Therefore, because of genetic drift, we would expect a rapid extinction of genetic variation in AMF. What counters this loss of variation?

5.2.3. Maintenance of genetic variation in AMF

Assuming that AMF have no extraordinary mechanisms to generate large amounts of genetic variation, and that a genetic bottleneck at sporulation leads to a loss of variation, the high levels of genetic variation that are found in this group are largely due to efficient maintenance of genetic variation. Population genetics theory predicts that genetic variation is best maintained in populations with a large effective population size, which is the number of individuals (here: genomes or nuclei) that contribute to the total genetic variation in the next generation. If we describe AMF nuclei in a cytoplasm as a population of genomes reproducing clonally (i.e. undergoing mitosis), AMF hyphal systems do indeed contain large numbers of individuals (Bianciotto and Bonfante 1992; Hosny *et al.* 1998; Marleau *et al.* 2011). However, as chapter 3 of this thesis demonstrated, the amount of genetic variation in a single spore, which in chapter 3 was measured with a single marker, *PLS*, is a mere subset of the total genetic variation to be found in the 'mother' cytoplasm. In this scenario, genetic variation would quickly disappear because of genetic drift at sporulation. We need a mechanism that can restore the loss of variation at sporulation: enter anastomosis.

Anastomosis is the fusion of hyphae between the same or closely related hyphal networks. Anastomosis is widespread throughout the fungi and is essential to maintain connectivity in hyphal networks (Bryce 2001; Glass and Kaneko 2003; Glass *et al.* 2004). In AMF, much work has been done on hyphal fusion, and the varying degrees to which AMF isolates anastomose or not (Giovannetti *et al.* 1999; Giovannetti *et al.* 2001; De la Providencia *et al.* 2005). Decreasing anastomosis rates between more distantly related AMF strains indicates that there are self

recognition and non-self incompatibility mechanisms at play (Giovannetti *et al.* 2006), which might be a fertile avenue for taxon delineation in AMF (on which more in the section 5.4).

For the purpose of this discussion, the key feature of anastomosis is its potential for the exchange of genetic information. Recent studies have demonstrated that indeed, nuclei are exchanged during anastomosis (Croll 2009; Marleau *et al.* 2011) Furthermore, the genetic variation that is transferred during anastomosis can make a difference to the phenotype of the receiving AMF and the host plant it colonizes (Angelard *et al.* 2010; Angelard and Sanders 2011).

But how much anastomosis is needed to maintain the high levels of genetic variation that are observed in AMF? We can model the maintenance of genetic variation with and without anastomosis. This is exactly what Bever *et al.* did (Bever and Wang 2005; Bever *et al.* 2008). Assuming the presence of a 1000 nuclei in the AMF cytoplasm, which, in the light of our current understanding, is a very modest number (Bianciotto and Bonfante 1992; Hosny *et al.* 1998; Marleau *et al.* 2011), and no recombination, they modeled the role of anastomosis in the maintenance of genetic variation in AMF. With no anastomosis, the number of variants decreases rapidly, and is extinguished within the time span of 25 generations. With a fusion rate of 25%, almost all variation is conserved through 200 generations.

It is relevant to state that in this model, there is no need to invoke meiosis or recombination to maintain the observed levels of genetic variation, Nevertheless, as discussed in section 5.2.1, there are indications that these processes occur in AMF. By shuffling genetic information between nuclei, thus breaking linkage disequilibrium, recombination will lead to novel combinations of alleles and loci within AMF nuclei. This is important to counter Muller's ratchet (Muller 1964), which is a process that was already mentioned in the introduction. Muller's ratchet describes how, in the absence of recombination, deleterious mutations accumulate in the genome of asexual organisms. This 'mutational load' can be quite severe,

and is one of the major arguments why asexual lineages are supposed to be short-lived (Otto 2009).

Mutational load in relation to sporogenesis mode was studied in AMF by Jany and Pawlowska (2010). Following Kondrashov (1994), they describe three possible forms of sporogenesis. The false sporogenesis mode is the initial migration of a single founder nucleus in the spore, resulting in little mutation accumulation. All spore nuclei would be more or less identical. In the random sporogenesis mode, nuclei migrate at random from the parental mycelium into the spore. As a result, the spores contain mixtures of mutant and non-mutant nuclei. In the sectorial sporogenesis mode, nuclei represent a sampling from a mycelium sector populated by offspring from a single nucleus. These three modes were implemented in a model based on the work by Otto and Orive (1995). Jany and Pawlowska concluded that, to avoid the accumulation of deleterious mutations, AMF must engage in a mixture of random and sectorial sporogenesis. The balance between nuclear-level selection and individual-level selection (here: isolate-level selection) is what purges spores that have too many nuclei with deleterious alleles.

There is supporting evidence for just this occurring in a recent study by Marleau *et al.* (2011). The authors demonstrated that multiple nuclei migrate into the spore, and that mitosis in the spore is limited. Furthermore, they found that in *G. irregulare* spore size and the number of nuclei per spore is correlated, and that bigger spores (that contain higher numbers of nuclei) show higher germination rates. Interestingly, Burggraaf and Beringer (1989) also report that potential for growth in *Gigaspora margarita* is correlated to the number of nuclei present. These results support the idea that AMF must engage in a mixture of random and sectorial sporogenesis. Moreover, the observation that spores under a certain size do not germinate indicates the possibility of selection at the individual (here: spore) level.

It is interesting to place these results in a larger taxonomic context. The *Gigasporaceae*, which as their name suggests, produce huge spores with from 2,000 (Zézé *et al.* 1996) up to 20,000 nuclei (Burggraaf and Beringer 1989; Viera and Glenn 1990) show different anastomosis patterns than fungi from the *Glomeraceae*. Fungi from the *Gigasporaceae* differ in their capacity to anastomose within the same hyphal network, and never anastomose between different hyphal networks (De la Providencia *et al.* 2005; Voets *et al.* 2006). In contrast, fungi from the *Glomeraceae* interconnect more readily, even between different hyphal networks (Giovannetti *et al.* 2004; Marleau *et al.* 2011). Is this observation linked to larger nuclear population sizes in *Gigasporaceae* spores? This larger population size could mean that *Gigasporaceae* have a lesser need to replenish genetic variation or to counter mutational load, and thus anastomose less. Alternatively, the incapacity of *Gigasporaceae* to anastomose has lead to an increased mutational load that can only be countered by larger nuclear population size per spore. Either way, the difference in anastomosis capacity between *Gigasporaceae* and *Glomeraceae* can be translated in fascinating evolutionary hypotheses.

This overview of the dynamics of genetic variation in AMF is far form complete. As of now, there is still very little known about mutational load in AMF and the precise dynamics of generation, loss and maintenance of genetic variation. However, it should be clear that multiple levels of selection act on AMF nuclei, spores and hyphae. What is the level of the 'individual' in AMF, and on which level(s) does selection act? The next section will address these questions and put them in the perspective of some recent developments in evolutionary theory.

5.3. Multiple genomes and evolutionary theory

AMF defy many conventions in evolutionary biology and their complicated life history incites a unique assemblage of fascinating questions. I have identified

two that I would like to discuss in more detail: “What is an individual?’ and ‘What is the unit of evolution?.

5.3.1. What is an individual in AMF?

Defining individuality in evolutionary biology is a major challenge. At the same time, maybe not surprisingly for a discipline that traditionally focuses on population level processes, evolutionary biology does not seem very concerned with the ‘individual’. There are scant discussions on the subject, and few studies explicitly define what is meant by the term. Similarly, it is very difficult to find a non-ambiguous definition for ‘organism’. In this discussion, I use the terms interchangeably, even though Michael Ghiselin has correctly remarked that ‘although all organisms are individuals, not all individuals are organism’ (Ghiselin 2011).

The biologists that have undertaken an investigation into the meaning of individuality are without exception workers that study non-model organisms. Santelices (1999) works on macroalgae and lists three attributes that are classically used to characterize individuality; genetic uniqueness, autonomy and genetic homogeneity. For each of these attributes exceptions can be easily found. For example, genetic uniqueness is less relevant when discussing clonal organisms and autonomy is a difficult concept for symbionts or parasites. Santelices continued investigations into individuality try to discern differences in ecological responses between clonal, aclonal and chimeric macroalgae or seaweeds (Santelices 2004a). These different kinds of seaweed are often morphologically indistinguishable and have only recently been split in the abovementioned categories, defined by differences in genetic composition. Santelices’ findings confirm that different ‘kinds of individual’ (i.e. clonal, aclonal and chimeric seaweeds) indeed show differences in their interaction with the

environment. Based on his findings, he argues that it is important to distinguish between mechanisms by which intra-organismal genetic heterogeneity (IGH) is generated. A mosaic organism, which is genetically heterogeneous through somatic mutations, clonal reproduction and the ensuing lack of a bottleneck, is not confronted with the same life history decisions as a chimeric organism, which is the result of a fusion between two genetically distinct organisms (Santelices 2004b).

Santelices published this emphasis on the distinction between mosaic and chimeras as a comment on the review by Pineda-Krch and Lehtila (2004b), whose theoretical work elucidates the effect that the spatial arrangement of genetically homogenous and heterogeneous modular organisms has on organismal fitness (Pineda-Krch and Poore 2004). In their review, the authors argue that genetic homogeneity might be the exception instead of the rule in biology and discuss the advantages and disadvantages of genetic heterogeneity to an individual. Like Santelices, they differentiate between IGH in chimeras and IGH in mosaic organisms. They state that chimeras might be quite rare but that mosaic organisms can profit form numerous advantages such as synergistic effects on the phenotype, size advantages and plant-herbivore interactions.

Queller and Strassman, who work on social amoeba, also stress the importance of differentiating between mosaics and chimeras, stating that cooperation between clonal cells is more readily achieved than cooperation between aclonal cells (Strassmann and Queller 2004; Kuzdzal-Fick *et al.* 2011), although there are situations where chimerism delivers a net benefit to all parties involved (Foster *et al.* 2002). What is the difference between chimeras and mosaic organisms? A mosaic organism consists of cells that share the same genealogy, and thus have the same most recent common ancestor. Their coexistence is the results of a lack of separation between germline and soma, in other words, a lack of a genetic bottleneck between generations. A chimeric organism, on the other hand, consists of an aggregate of cells that do not share the same genealogy.

It is proposed that for chimeric organisms, which are composed of aclonal cell lineages, it is much harder to reduce genetic conflict within the aggregate (Michod and Roze 2001). The reduction of conflict lies at the base of current definitions of individuality (Buss 1987; Maynard Smith and Szathmáry 1995) and will be discussed in the next section, in the light of multilevel selection theory. For now, it suffices to state that an important difference between chimerism and mosaicism lies in the potential for genetic conflict. This conflict is costly to the organism, and it is thus not surprising that mechanisms have evolved to avoid this cost.

These mechanisms are best described by self-nonself recognition, and serve the organism to differentiate related cells, which represent minimum genetic conflict, from unrelated cells, which would lead to the formation of chimeras and might entail a fitness loss for the organism through competition between cells. The immune system is an example of a self-nonself recognition system, and the conceptual elegance of using the immune system as a proxy for individuality has led to the proposition that the immune system is an essential component of biological individuality (Pradeu 2010). However, it seems to me that Pradeu's perspective does not distinguish the proximate and ultimate causes of individuality. The immune system is an important effect of the organization of biological entities into groups like cells and organisms, not its cause.

Self and nonself recognition is well studied in fungi, where anastomosis and consequent fusion of genetically related and unrelated hyphal systems is an essential component of fungal (sexual) reproduction (Glass *et al.* 2004). Here, policing mechanisms exist that prevent intergenomic conflicts while allowing for gene transfer and subsequent introgression (Roper *et al.* 2011). Thus, fungi harvest the benefits of being chimerical without facing the associated risks. AMF are the culmination of this strategy.

Thus, the last part of our discussion on individuality brings us back to AMF. The co-existence of genetically divergent nuclei in AMF raises the question of what exactly constitutes an individual within this enigmatic group of fungi. Even though we might not be able to provide a universal answer to the question 'what is an individual', we might have discovered some concepts that could help us resolve the problem of 'the individual' in AMF. Which of the abovementioned components of individuality can help us in our search?

AMF do not show genetic homogeneity, but different isolates come close to the concept of individuality with respect to genetic uniqueness and autonomy, and thus fulfill two out of three attributes that were identified by Santelices (1999). The concept of an isolate thus forms a pragmatic solution to the issue of defining individuality in AMF. However, isolates are fungal strains that were started from a single spore, are maintained in a culture system, and might show decreased genetic variation due to a lack of anastomosis and the effects of genetic drift (Cárdenas-Flores *et al.* 2010, chapter 3, this thesis). Therefore, isolates might not be representative for AMF in the field. Attempts at identifying AMF in the field have mostly been focused on the use of genetic markers. Researchers have attempted to identify AMF using ribosomal DNA, introns in protein-coding genes, microsatellites, AFLP, RAPD and mitochondrial sequences (reviews in Sanders 2004; Rosendahl 2008; Sanders and Croll 2010). None of these markers, however, can summarize the enormous amount of genetic variation that is contained by AMF hyphae. In fact, if all nuclei within AMF hyphae are genetically different, one would have to sequence the entire organism to capture a unique 'genotype', which is why we proposed that AMF contain a pangenome in chapter 4.

The use of anastomosis as a criterion for the delineation of the individual in AMF seems promising. Anastomosis is vital to the maintenance of genetic variation in AMF (see chapter 3 and section 5.3 of this thesis) and a self-nonself recognition system based on relatedness seems to be in place (Giovannetti *et al.* 1999; De la

Providencia *et al.* 2005; Croll 2009; Marleau *et al.* 2011). However, this is a phenomenon that can only be studied successfully in a laboratory environment. So far, we have no single definition of what constitutes an AMF individual in the field.

If we accept that AMF identification can be approximated by specific genetic markers, progress has been made through pragmatism. By defining phylotypes (Winther and Friedman 2008) or Operational Taxonomic Units (Pivato *et al.* 2007; Öpik *et al.* 2009), researchers have managed to differentiate between different AMF, be they individuals, populations or species. One might want to raise concerns about reproducibility, consistent criteria for OTU delineation, and the risk of erroneously splitting and/or lumping groups because of insufficient sampling. Nevertheless, for the moment this is probably as good as it gets regarding field identification of AMF.

However, if we want to move past a definition of individuality based on a limited sampling of genetic variation, we tread on virgin territory. Some authors have attempted to expand on the difficulties of determining species, community and individual level processes in AMF (Rosendahl 2008) but offered no solution to the uncertainty. In my opinion, the only possible solution to the impasse of AMF identification in the laboratory or in the field is to give up attempts to integrate all aspects of individuality in a single definition for the AMF individual. For the moment, the microscopic nature of AMF condemns us to genetic approximations of its life history, while its complex genome organization makes it hard to apply trusted molecular methods to our questions. It is possible, however, to embrace the impossibility to define individuality in biology and AMF in particular. In this case, we merely study our data without projecting conclusions on a framework that was devised to describe different biological realities.

An example of this approach can be found in the proposition by van der Heijden and Scheublin (2007) who state that ‘it is difficult to link the identification of AMF communities in the field with the functional significance of these AMF

communities'. They state that it makes more sense to directly describe AMF communities in terms of functional traits, without attempting to describe the community by proxy of (neutral) genetic markers. Similarly, Gamper *et al.* (2010) propose 'a shift toward plant and fungal protein-encoding genes as more immediate indicators of mycorrhizal contributions to ecological processes'.

In conclusion, I would like to point out the conceptual similarity of this approach with the field of metagenomics, where (microbial) genetic material from the environment is studied directly without attempts to establish cultures or assign species, community or individual definitions to the dataset (Handelsman *et al.* 1998; Handelsman 2004). Another similar approach has been developed by students of Lateral Gene Transfer (LGT), which is a process that effaces the boundaries between species as defined by genomic monophyly. To avoid the difficulties of defining species in the face of LGT, proponents of this approach propose to employ network theory as a method that is independent of individual and species assignations. This approach might offer an alternative method to describe the relations between interdependent components of a system that is not readily described through a more traditional branching evolutionary process (Halary *et al.* 2009; Tamminen *et al.* 2012).

5.3.2. What is the unit of evolution in AMF?

The idea of individuality is inextricably linked to the idea of closure. Conflict on a particular level of organization needs to be resolved before another level of biological organization can evolve (Weismann 1893; Buss 1987; Maynard Smith and Szathmáry 1995). However, in order to investigate the meaning of the unit of evolution in AMF, it might be informative to evaluate the balance between different levels of selection in these fungi.

The difficulty of defining individuality in AMF partially flows from implicit assumptions on the target of selection in evolutionary theory. Without explicitly stating so, 'the individual' is often equated with 'the genome', since this is perceived to be the most fundamental unit of evolution. However, selection can act on different levels and thus the concept of individuality can equally fluid. Especially D.S. Wilson and Elliott Sober have advocated a pluralistic approach to the quest for the fundamental unit of evolution (see Wilson and Sober 1994 for some of their most accessible and concise arguments). They propose that selection can act on different levels of biological organization simultaneously, with the eventual outcome (in terms of evolution of the biological system in question) depending on the balance between selection pressures on different levels. In their support, even staunch advocates of 'the gene as fundamental unit of selection' such as G.C. Williams have admitted that some systems might best be explained by selection operating on other levels but the gene (Williams 1992).

In any case, AMF are the perfect example to illustrate this question. There are indications that a certain number of nuclei within AMF need to be present for successful spore germination (Marleau *et al.* 2011). This means that a fully functional AMF genome might be partitioned between several nuclei and would imply an interdependence of nuclei, possibly based on complementarity. AMF nuclei together form, as it were, a pangenome with the organism. The next level of selection could be the hyphal system itself that germinated from a single spore. However, this hyphal system is possibly not viable in the long term without the replenishment of genetic variation through anastomosis (chapter 3, this thesis). Thus, strictly speaking, the next functional level should be the extended hyphal system, consisting of several fused hyphal systems, each originating from a single spore. Nevertheless, a germinating spore is not viable when it does not form a symbiosis with a plant's root system. Thus, fungal-plant symbiosis is another level of organization. And so on, seemingly *ad infinitum*.

As Okasha (2008) argues, multilevel selection was not proposed to help decide which is 'the' level of selection or which is 'the' unit of evolution, but is more meant to provide tools to analyze the relative importance of each level with respect to the biological system under study. In the end, it is more productive if researchers take a pragmatic approach and define their units of evolution according to their research question. As such, even if it might not be possible to resolve the debate on what the fundamental unit of evolution in AMF really is, teasing the relative contributions of each level apart might be a fruitful pursuit for researchers that want to understand AMF life history.

References

Adams KL, Cronn R, Percifield R, Wendel JF (2003) Genes duplicated by polyploidy show unequal contributions to the transcriptome and organ-specific reciprocal silencing. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 4649-4654.

Anderson MJ (2001) A new method for non-parametric multivariate analysis of variance. *Austral Ecology* **26**, 32-46.

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

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

Arnheim N (1983) Concerted evolution of multigene types. In: *Evolution of Genes and Proteins* (eds. Nei M, Koehn RK), pp. 38–61. Sinauer, Sanderland.

Avio L, Pellegrino E, Bonari E, Giovannetti M (2006) Functional diversity of arbuscular mycorrhizal fungal isolates in relation to extraradical mycelial networks. *New Phytologist* **172**, 347-357.

Azcon-Aguilar C, Barea JM (1996) Arbuscular mycorrhizas and biological control of soil-borne plant pathogens - An overview of the mechanisms involved. *Mycorrhiza* **6**, 457-464.

Becard G, Pfeffer PE (1993) Status of nuclear division in arbuscular mycorrhizal fungi during in-vitro development. *Protoplasma* **174**, 62-68.

Bergemann J, Florian V, Kremser T, Klein A (1996) Two different macronuclear EF1α-encoding genes of the ciliate *Euplotes crassus* are very dissimilar in their sequences, copy numbers and transcriptional activities. *Gene* **168**, 109-112.

Bertani G (1951) Studies on lysogenesis. I. The mode of phage liberation by lysogenic *Escherichia coli*. *Journal of Bacteriology* **62**, 293-300.

Bever JD (2002) Negative feedback within a mutualism: host-specific growth of mycorrhizal fungi reduces plant benefit. *Proceedings of the Royal Society B: Biological Sciences* **269**, 2595-2601.

Bever JD, Kang H-J, Kaonongbua W, Wang M (2008) Genomic organization and mechanisms of inheritance in arbuscular mycorrhizal fungi: contrasting the evidence and implications of current theories. In: *Mycorrhiza* (ed. Varma A), pp. 135-148. Springer-Verlag, Berlin Heidelberg.

Bever JD, Morton JB, Antonovics J, Schultz PA (1996) Host-dependent sporulation and species diversity of arbuscular mycorrhizal fungi in a mown grassland. *Journal of Ecology* **84**, 71-82.

Bever JD, Wang M (2005) Arbuscular mycorrhizal fungi - Hyphal fusion and multigenomic structure. *Nature* **433**, E3-E4.

Bianciotto V, Bonfante P (1992) Quantification of the nuclear-DNA content of 2 arbuscular mycorrhizal fungi. *Mycological Research* **96**, 1071-1076.

Blaalid R, Carlsen TOR, Kumar S*, et al.* (2011) Changes in the root-associated fungal communities along a primary succession gradient analysed by 454 pyrosequencing. *Molecular Ecology* **21**, 1897–1908.

Bonfante P, Anca I-A (2009) Plants, mycorrhizal fungi, and bacteria: a network of interactions. *Annual Review of Microbiology* **63**, 363-383.

Boon E, Zimmerman E, Lang BF, Hijri M (2010) Intra-isolate genome variation in arbuscular mycorrhizal fungi persists in the transcriptome. *Journal of Evolutionary Biology* **23**, 1519-1527.

Bougnoux M-E, Morand S, d'Enfert C (2002) Usefulness of multilocus sequence typing for characterization of clinical isolates of *Candida albicans*. *Journal of Clinical Microbiology* **40**, 1290-1297.

Bradley RD, Hillis DM (1997) Recombinant DNA sequences generated by PCR amplification. *Molecular Biology and Evolution* **14**, 592-593.

Brakenhoff RH, Schoenmakers JGG, Lubsen NH (1991) Chimeric cDNA clones- a novel artefact. *Nucleic Acids Research* **19**, 1949-1949.

Brundrett MC (2002) Coevolution of roots and mycorrhizas of land plants. *New Phytologist* **154**, 275-304.

Bryce K (2001) *The Fifth Kingdom*, 3 edn. Mycologue Publications, Boston.

Burggraaf AJP, Beringer JE (1989) Absence of nuclear DNA synthesis in vesicular-arbuscular mycorrhizal fungi during in vitro development. *New Phytologist* **111**, 25–33.

Buri P (1956) Gene frequency in small populations of mutant drosophila. *Evolution* **10**, 367-402.

Buss LW (1982) Somatic cell parasitism and the evolution of somatic tissue compatibility. *Proceedings of the National Academy of Sciences of the United States of America* **79**, 5337-5341.

Buss LW (1987) *The evolution of individuality* Princeton University Press, Princeton, NJ.

Butler JM, Hill CR (2010) Scientific issues with analysis of low amounts of DNA.

Cárdenas-Flores A, Draye X, Bivort C, Cranenbrouck S, Declerck S (2010) Impact of multispores in vitro subcultivation of Glomus sp. MUCL 43194 (DAOM 197198) on vegetative compatibility and genetic diversity detected by AFLP. *Mycorrhiza* **20**, 415-425.

Chao A, Chazdon RL, Colwell RK, Shen T-J (2005) A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters* **8**, 148-159.

Clapp JP, Fitter AH, Young JPW (1999) Ribosomal small subunit sequence variation within spores of an arbuscular mycorrhizal fungus, *Scutellospora sp*. *Molecular Ecology* **8**, 915-921.

Clapp JP, Rodriguez A, Dodd JC (2001) Inter- and intra-isolate rRNA large subunit variation in *Glomus coronatum* spores. *New Phytologist* **149**, 539-554.

Corradi N, Croll D, Colard A*, et al.* (2007) Gene copy number polymorphisms in an arbuscular mycorrhizal fungal population. *Applied and Environmental Microbiology* **73**, 366-369.

Corradi N, Hijri M, Fumagalli L, Sanders IR (2004a) Arbuscular mycorrhizal fungi (Glomeromycota) harbour ancient fungal tubulin genes that resemble those of the chytrids (Chytridiomycota). *Fungal Genetics and Biology* **41**, 1037-1045.

Corradi N, Kuhn G, Sanders IR (2004b) Monophyly of beta-tubulin and H+-ATPase gene variants in *Glomus intraradices*: consequences for molecular evolutionary studies of AM fungal genes. *Fungal Genetics and Biology* **41**, 262-273.

Corradi N, Sanders IR (2006) Evolution of the P-type II ATPase gene family in the fungi and presence of structural genomic changes among isolates of *Glomus intraradices*. *Bmc Evolutionary Biology* **6**.

Croll D, Corradi N, Gamper HA, Sanders IR (2008a) Multilocus genotyping of arbuscular mycorrhizal fungi and marker suitability for population genetics. *New Phytologist* **180**, 564-568.

Croll D, Giovannetti, M., Koch, A.M., Sbrana, C., Ehinger, M., Lammers, P.J., Sanders, I.R. (2009) Nonself vegetative fusion and genetic exchange in the arbuscular mycorrhizal fungus *Glomus intraradices*. *New Phytologist* **181**, 924–937.

Croll D, Sanders I (2009) Recombination in *Glomus intraradices*, a supposed ancient asexual arbuscular mycorrhizal fungus. *Bmc Evolutionary Biology* **9**, 13.

Croll D, Wille L, Gamper HA*, et al.* (2008b) Genetic diversity and host plant preferences revealed by simple sequence repeat and mitochondrial markers in a population of the arbuscular mycorrhizal fungus *Glomus intraradices*. *New Phytologist* **178**, 672-687.

Daubin V, Moran NA (2004) Comment on "The Origins of Genome Complexity". *Science* **306**, 978.

De la Providencia IE, de Souza FA, Fernández F, Delmas NS, Declerck S (2005) Arbuscular mycorrhizal fungi reveal distinct patterns of anastomosis formation and hyphal healing mechanisms between different phylogenetic groups. *New Phytologist* **165**, 261-271.

Den Bakker HC, VanKuren NW, Morton JB, Pawlowska TE (2010) Clonality and recombination in the life history of an asexual arbuscular mycorrhizal fungus. *Molecular Biology and Evolution* **27**, 2474-2486.

Dodd JC, Rosendahl S, Giovannetti M*, et al.* (1996) Inter- and intraspecific variation within the morphologically similar arbuscular mycorrhizal fungi *Glomus mosseae* and *Glomus coronatum*. *New Phytologist* **133**, 113-122.

Drake JW, Charlesworth B, Charlesworth D, Crow JF (1998) Rates of spontaneous mutation. *Genetics* **148**, 1667-1686.

Dumbrell AJ, Ashton PD, Aziz N*, et al.* (2011) Distinct seasonal assemblages of arbuscular mycorrhizal fungi revealed by massively parallel pyrosequencing. *New Phytologist* **190**, 794–804.

Ebersberger I, de Matos Simoes R, Kupczok A*, et al.* (2011) A consistent phylogenetic backbone for the fungi. *Molecular Biology and Evolution* **29**, 1319-1334.

Elder JF, Turner BJ (1995) Concerted Evolution of repetitive DNA-sequences in eukaryotes. *Quarterly Review of Biology* **70**, 297-320.

Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. *Genetics* **131**, 479-491.

Falkenberg M, Larsson NG, Gustafsson CM (2007) DNA replication and transcription in mammalian mitochondria. *Annual Review of Biochemistry* **76**, 679-699.

Feil EJ, Maiden MCJ, Achtman M, Spratt BG (1999) The relative contributions of recombination and mutation to the divergence of clones of *Neisseria meningitidis*. *Molecular Biology and Evolution* **16**, 1496-1502.

Feil EJ, Spratt BG (2001) Recombination and the population structures of bacterial pathogens. *Annual Review of Microbiology* **55**, 561-590.

Folse HJ, Roughgarden J (2010) What is an individual organism? A multilevel selection perspective. *Quarterly Review of Biology* **85**, 447-472.

Foster KR, Fortunato A, Strassmann JE, Queller DC (2002) The costs and benefits of being a chimera. *Proceedings: Biological Sciences* **269**, 2357-2362.

Frey-Klett P, Garbaye J, Tarkka M (2007) The mycorrhiza helper bacteria revisited. *New Phytologist* **176**, 22-36.

Gamper HA, van der Heijden MGA, Kowalchuk GA (2010) Molecular trait indicators: moving beyond phylogeny in arbuscular mycorrhizal ecology. *New Phytologist* **185**, 67-82.

Gamper HA, Walker C, Schussler A (2009) *Diversispora celata sp* nov: molecular ecology and phylotaxonomy of an inconspicuous arbuscular mycorrhizal fungus. *New Phytologist* **182**, 495-506.

Gandolfi A, Sanders IR, Rossi V, Menozzi P (2003) Evidence of recombination in putative ancient asexuals. *Molecular Biology and Evolution* **20**, 754-761.

Ghiselin MT (2011) A consumer's guide to superorganisms. *Perspectives in Biology and Medicine* **54**, 152 -167

Gill P, Whitaker J, Flaxman C, Brown N, Buckleton J (2000) An investigation of the rigor of interpretation rules for STRs derived from less than 100 pg of DNA. *Forensic Science International* **112**, 17-40.

Gilles A, Meglecz E, Pech N*, et al.* (2011) Accuracy and quality assessment of 454 GS-FLX Titanium pyrosequencing. *Bmc Genomics* **12**, 245.

Giovannetti M, Avio L, Fortuna P*, et al.* (2006) At the root of the wood wide web: self recognition and nonself incompatibility in mycorrhizal networks. *Plant Signaling & Behavior* **1**, 1-5.

Giovannetti M, Azzolini D, Citernesi AS (1999) Anastomosis formation and nuclear and protoplasmic exchange in arbuscular mycorrhizal fungi. *Applied and Environmental Microbiology* **65**, 5571-5575.

Giovannetti M, Fortuna P, Citernesi AS, Morini S, Nuti MP (2001) The occurrence of anastomosis formation and nuclear exchange in intact arbuscular mycorrhizal networks. *New Phytologist* **151**, 717-724.

Giovannetti M, Gianinazzi-Pearson V (1994) Biodiversity in arbuscular mycorrhizal fungi. *Mycological Research* **98**, 705-715.

Giovannetti M, Sbrana C, Avio L, Strani P (2004) Patterns of below-ground plant interconnections established by means of arbuscular mycorrhizal networks. *New Phytologist* **164**, 175-181.
Glass NL, Kaneko I (2003) Fatal attraction: nonself recognition and heterokaryon incompatibility in filamentous fungi. *Eukaryotic Cell* **2**, 1-8.
Glass NL, Rasmussen C, Roca MG, Read ND (2004) Hyphal homing, fusion and mycelial interconnectedness. *Trends in Microbiology* **12**, 135-141.
Gollotte A, L'Haridon F, Chatagnier O*, et al.* (2006) Repetitive DNA sequences include retrotransposons in genomes of the Glomeromycota. *Genetica* **128**, 455-469.
Gollotte A, van Tuinen D, Atkinson D (2004) Diversity of arbuscular mycorrhizal fungi colonising roots of the grass species *Agrostis capillaris* and *Lolium perenne* in a field experiment *Mycorrhiza* **14**, 111-117.
Gonzalez JM, Zimmermann J, Saiz-Jimenez C (2005) Evaluating putative chimeric sequences from PCR-amplified products. *Bioinformatics* **21**, 333-337.
Guindon Sp, Lethiec F, Duroux P, Gascuel O PHYML Online, a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic Acids Research* **33**, W557-W559.
Haas BJ, Gevers D, Earl AM*, et al.* (2011) Chimeric 16S rRNA sequence formation and detection in Sanger and 454-pyrosequenced PCR amplicons. *Genome Research* **21**, 494-504.
Halary S, Leigh JW, Cheaib B, Lopez P, Bapteste E (2009) Network analyses structure genetic diversity in independent genetic worlds. *Proceedings of the National Academy of Sciences* **107**, 127-132.
Halary S, Malik S-B, Lildhar L*, et al.* (2011) Conserved meiotic machinery in *Glomus spp.*, a putatively ancient asexual fungal lineage. *Genome Biology and Evolution* **3**
Hall TA (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT *Nucleic Acids Symposium Series* **41**, 95-98.
Handelsman J (2004) Metagenomics: Application of genomics to uncultured microorganisms. *Microbiology and Molecular Biology Reviews* **68**, 669-685.
Handelsman J, Rondon MR, Brady SF, Clardy J, Goodman RM (1998) Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology* **5**, R245-R249.
Harrison MJ (1999) Molecular and cellular aspects of the arbuscular mycorrhizal symbiosis. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**, 361-389.
Hasegawa M, Kishino H, Yano T (1985) Dating the human-ape split by a molecular clock of mitochondrial DNA. *Journal of Molecular Evolution* **22**, 160-174.
Hedin MC, Maddison WP (2001) Phylogenetic utility and evidence for multiple copies of elongation factor-1α in the spider genus *Habronattus* (Araneae: Salticidae). *Molecular Biology and Evolution* **18**, 1512-1521.
Helgason T, Watson IJ, Young JPW (2003) Phylogeny of the Glomerales and diversisporales (Fungi: Glomeromycota) from actin and elongation factor 1-alpha sequences. *Fems Microbiology Letters* **229**, 127-132.
Hibbett DS, Ohman A, Glotzer D*, et al.* (2011) Progress in molecular and morphological taxon discovery in Fungi and options for formal classification of environmental sequences. *Fungal Biology Reviews* **25**, 38-47.
Hijri M, Hosny M, van Tuinen D, Dulieu H (1999) Intraspecific ITS polymorphism in *Scutellospora castanea* (Glomales, Zygomycota) is structured within multinucleate spores. *Fungal Genetics and Biology* **26**, 141-151.
Hijri M, Niculita H, Sanders IR (2007) Molecular characterization of chromosome termini of the arbuscular mycorrhizal fungus *Glomus intraradices* (Glomeromycota). *Fungal Genetics and Biology* **44**, 1380-1386.
Hijri M, Redecker D, Petetot J*, et al.* (2002) Identification and isolation of two ascomycete fungi from spores of the arbuscular mycorrhizal fungus *Scutellospora castanea*. *Applied and Environmental Microbiology* **68**, 4567-4573.

Hijri M, Sanders IR (2004) The arbuscular mycorrhizal fungus *Glomus intraradices* is haploid and has a small genome size in the lower limit of eukaryotes. *Fungal Genetics and Biology* **41**, 253-261.
Hijri M, Sanders IR (2005) Low gene copy number shows that arbuscular mycorrhizal fungi inherit genetically different nuclei. *Nature* **433**, 160-163.
Hildebrandt U, Kaldorf M, Bothe H (1999) The zinc violet and its colonization by arbuscular mycorrhizal fungi. *Journal of Plant Physiology* **154**, 709-717.
Hildebrandt U, Regvar M, Bothe H (2007) Arbuscular mycorrhiza and heavy metal tolerance. *Phytochemistry* **68**, 139-146.
Hosny M, deBarros JPP, Gianinazzi-Pearson V, Dulieu H (1997) Base composition of DNA from glomalean fungi: High amounts of methylated cytosine. *Fungal Genetics and Biology* **22**, 103-111.
Hosny M, Gianinazzi-Pearson V, Dulieu H (1998) Nuclear DNA content of 11 fungal species in Glomales. *Genome* **41**, 422-428.
Hosny M, Hijri M, Passerieux E, Dulieu H (1999) rDNA units are highly polymorphic in Scutellospora castanea (Glomales, Zygomycetes). *Gene* **226**, 61-71.
Huber T, Faulkner G, Hugenholtz P (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments. *Bioinformatics* **20**, 2317-2319.
Hudson RR (2000) A New Statistic for Detecting Genetic Differentiation. *Genetics* **155**, 2011-2014.
Hugenholtz P, Huber T (2003) Chimeric 16S rDNA sequences of diverse origin are accumulating in the public databases. *International Journal of Systematic and Evolutionary Microbiology* **53**, 289-293.
Huse S, Huber J, Morrison H, Sogin M, Welch D (2007) Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biology* **8**, R143.
Huse SM, Welch DM, Morrison HG, Sogin ML (2010) Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* **12**, 1889-1898.
Huson DH, Bryant D (2006) Application of phylogenetic networks in evolutionary studies. *Molecular Biology and Evolution* **23**, 254-267.
Huson DH, Mitra S, Ruscheweyh H-J, Weber N, Schuster SC (2011) Integrative analysis of environmental sequences using MEGAN4. *Genome Research* **21** 1552-1560.
Hwang SK, Kim JG (2000) Secondary structural and phylogenetic implications of nuclear large subunit ribosomal RNA in the ectomycorrhizal fungus *Tricholoma matsutake*. *Current Microbiology* **40**, 250-256.
Jackson AP, Gamble JA, Yeomans T*, et al.* (2009) Comparative genomics of the fungal pathogens *Candida dubliniensis* and *Candida albicans*. *Genome Research* **19**, 2231-2244.
James TY, Kauff F, Schoch CL*, et al.* (2006) Reconstructing the early evolution of Fungi using a six-gene phylogeny. *Nature* **443**, 818-822.
Jany JÄ, Pawlowska TE (2010) Multinucleate spores contribute to evolutionary longevity of asexual glomeromycota. *The American Naturalist* **175**, 424-435.
Ji GE, Orlowski M (1990) Primary and secondary structure of the 25S rRNA from the dimorphic fungus *Mucor racemosus*. *Current Genetics* **17**, 499-506.
Kaiser VB, Charlesworth B (2009) The effects of deleterious mutations on evolution in non-recombining genomes. *Trends in Genetics* **25**, 9-12.
Kent W (2002) BLAT - the BLAST-like alignment tool. *Genome Research* **12**, 656 - 664.
Kim HL, Shin E, Kim HM*, et al.* (2007) Heterogeneous rRNA molecules encoded by *Streptomyces coelicolor* M145 genome are all expressed and assembled into Ribosomes. *Journal of Microbiology and Biotechnology* **17**, 1708-1711.
Kimura M (1985) *The Neutral Theory of Molecular Evolution* Cambridge University Press, Cambridge.
Kivlin SN, Hawkes CV, Treseder KK (2011) Global diversity and distribution of arbuscular mycorrhizal fungi. *Soil Biology and Biochemistry* **43**, 2294-2303.

Koch AM, Croll D, Sanders IR (2006) Genetic variability in a population of arbuscular mycorrhizal fungi causes variation in plant growth. *Ecology Letters* **9**, 103-110.
Komatsoulis GA, Waterman MS (1997) A new computational method for detection of chimeric 16S rRNA artifacts generated by PCR amplification from mixed bacterial populations. *Applied and Environmental Microbiology* **63**, 2338-2346.
Kondrashov AS (1994) Mutation load under vegetative reproduction and cytoplasmic inheritance. *Genetics* **137**, 311-318.
Koonin EV (2005) Orthologs, paralogs, and evolutionary genomics. *Annual Review of Genetics* **39**, 309-338.
Krüger M, Krüger C, Walker C, Stockinger H, Schüßler A (2012) Phylogenetic reference data for systematics and phylotaxonomy of arbuscular mycorrhizal fungi from phylum to species level. *New Phytologist* **193**, 970-984.
Krüger M, Stockinger H, Krüger C, Schüßler A (2009) DNA-based species level detection of Glomeromycota: one PCR primer set for all arbuscular mycorrhizal fungi. *New Phytologist* **183**, 212-223.
Kuhn G (2003) *Organisation of genetic variation in multinucleate arbuscular mycorrhizal fungi* University of Lausanne, Lausanne.
Kuhn G, Hijri M, Sanders IR (2001) Evidence for the evolution of multiple genomes in arbuscular mycorrhizal fungi. *Nature* **414**, 745-748.
Kuo C-H, Moran NA, Ochman H (2009) The consequences of genetic drift for bacterial genome complexity. *Genome Research* **19**, 1450-1454.
Kuzdzal-Fick JJ, Fox SA, Strassmann JE, Queller DC (2011) High relatedness is necessary and sufficient to maintain multicellularity in *Dictyostelium*. *Science* **334**, 1548-1551.
Lake JA (2009) Evidence for an early prokaryotic endosymbiosis. *Nature* **460**, 967-971.
Lasken RS (2007) Single-cell genomic sequencing using Multiple Displacement Amplification. *Current Opinion in Microbiology* **10**, 510-516.
Lekberg Y, Schnoor T, Kjøller R*, et al.* (2012) 454-sequencing reveals stochastic local reassembly and high disturbance tolerance within arbuscular mycorrhizal fungal communities. *Journal of Ecology* **100**, 151-160.
Liu Y, Steenkamp E, Brinkmann H*, et al.* (2009) Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support. *Bmc Evolutionary Biology* **9**, 272.
Lloyd-MacGilp SA, Chambers SM, Dodd JC*, et al.* (1996) Diversity of the ribosomal internal transcribed spacers within and among isolates of Glomus mosseae and related mycorrhizal fungi. *New Phytologist* **133**, 103-111.
Lopez-Lopez A, Benlloch S, Bonfa M, Rodriguez-Valera F, Mira A (2007) Intragenomic 16S rDNA divergence in *Haloarcula marismortui* is an adaptation to different temperatures. *Journal of Molecular Evolution* **65**, 687-696.
Lozupone C, Hamady M, Knight R (2006) UniFrac - An online tool for comparing microbial community diversity in a phylogenetic context. *Bmc Bioinformatics* **7**, 371.
Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71**, 8228-8235.
Lundberg KS, Shoemaker DD, Adams MWW*, et al.* (1991) High-fidelity amplification using a thermostable DNA polymerase isolated from *Pyrococcus furiosus*. *Gene* **108**, 1-6.
Lynch M (2007) *The origins of genome architecture* Sinauer Associates Inc, Sunderland, MA.
Lynch M, Bobay L-M, Catania F, Gout J-F, Rho M (2011) The repatterning of eukaryotic genomes by random genetic drift. *Annual Review of Genomics and Human Genetics* **12**, 347-366.
Lynch M, Conery JS (2003) The origins of genome complexity. *Science* **302**, 1401-1404.
Madsen HO, Poulsen K, Dahl O, Clark BFC, Hjorth JP (1990) Retropseudogenes constitute the major part of the human elongation factor 1α gene family. *Nucleic Acids Research* **18**, 1513-1516.

Magni GE, Von Borstel RC (1962) Different rates of spontaneous mutation during mitosis and meiosis in yeast. *Genetics* **47**, 1097-1108.
Maiden MCJ, Bygraves JA, Feil E*, et al.* (1998) Multilocus sequence typing: A portable approach to the identification of clones within populations of pathogenic microorganisms. *Proceedings of the National Academy of Sciences* **95**, 3140-3145.
Malloch DW, Pirozynski KA, Raven PH (1980) Ecological and evolutionary significance of mycorrhizal symbioses in vascular plants (a Review). *Proceedings of the National Academy of Sciences of the United States of America-Biological Sciences* **77**, 2113-2118.
Margulies M, Egholm M, Altman WE*, et al.* (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376-380.
Marleau J, Dalpe Y, St-Arnaud M, Hijri M (2011) Spore development and nuclear inheritance in arbuscular mycorrhizal fungi. *Bmc Evolutionary Biology* **11**, 51.
Martin AP (2002) Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology* **68**, 3673-3682.
Martin F, Gianinazzi-Pearson V, Hijri M*, et al.* (2008) The long hard road to a completed *Glomus intraradices* genome. *New Phytologist* **180**, 747-750.
Maynard Smith J, Szathmáry E (1995) *The major transitions in evolution* W. H. Freeman, Oxford, UK.
Meissner R, Jacobson Y, Melamed S*, et al.* (1997) A new model system for tomato genetics. *Plant Journal* **12**, 1465-1472.
Meyerhans A, Vartanian JP, Wainhobson S (1990) DNA recombination during PCR. *Nucleic Acids Research* **18**, 1687-1691.
Michod RE, Roze D (2001) Cooperation and conflict in the evolution of multicellularity. *Heredity* **86**, 1-7.
Moora M, Berger S, Davison J*, et al.* (2011) Alien plants associate with widespread generalist arbuscular mycorrhizal fungal taxa: evidence from a continental-scale study using massively parallel 454 sequencing. *Journal of Biogeography* **38**, 1305-1317.
Morton JB, Benny GL (1990) Revised classification of Arbuscular Mycorrhizal Fungi (Zygomycetes) - a new Order, Glomales, 2 new Suborders, Glomineae and Gigasporineae, and 2 new Families, Acaulosporaceae and Gigasporaceae, with an emendation of Glomaceae. *Mycotaxon* **37**, 471-491.
Muller HJ (1964) The relation of recombination to mutational advantage. *Mutational Research* **1**, 2-9.
Munkvold L, Kjoller R, Vestberg M, Rosendahl S, Jakobsen I (2004) High functional diversity within species of arbuscular mycorrhizal fungi. *New Phytologist* **164**, 357-364.
Musters W, Venema J, Vanderlinden G*, et al.* (1989) A system for the analysis of yeast ribosomal DNA mutations. *Molecular and Cellular Biology* **9**, 551-559.
Nei M (1987) *Molecular Evolutionary Genetics* Columbia University Press, New York.
Nilsson RH, Kristiansson E, Ryberg M, Hallenberg N, Larsson K-H (2008) Intraspecific ITS variability in the kingdom fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evolutionary Bioinformatics* **4**, 193-201.
Nyren P (1987) Enzymatic method for continuous monitoring of DNA polymerase activity. *Analytical biochemistry* **167**, 235-238.
Ockenden I, Dorsch JA, Reid MM*, et al.* (2004) Characterization of the storage of phosphorus, inositol phosphate and cations in grain tissues of four barley (*Hordeum vulgare* L.) low phytic acid genotypes. *Plant Science* **167**, 1131-1142.
Ohta T (1992) The nearly neutral theory of molecular evolution. *Annual Review of Ecology and Systematics* **23**, 263-286.
Okasha S (2008) *Evolution and the Levels of Selection* Oxford University Press, Oxford.
Opdenakker G, Cabeza-Arvelaiz Y, Fiten P*, et al.* (1987) Human elongation factor 1α: a polymorphic and conserved multigene family with multiple chromosomal localizations. *Human Genetics* **75**, 339-344.

Öpik M, Metsis M, Daniell TJ, Zobel M, Moora M (2009) Large-scale parallel 454 sequencing reveals host ecological group specificity of arbuscular mycorrhizal fungi in a boreonemoral forest. *New Phytologist* **184**, 424-437.

Öpik M, Moora M, Liira J, Zobel M (2006) Composition of root-colonizing arbuscular mycorrhizal fungal communities in different ecosystems around the globe. *Journal of Ecology* **94**, 778-790.

Öpik M, Vanatoa A, Vanatoa E*, et al.* (2010) The online database MaarjAM reveals global and ecosystemic distribution patterns in arbuscular mycorrhizal fungi (Glomeromycota). *New Phytologist* **188**, 223–241.

Otto SP (2009) The evolutionary enigma of sex. *The American Naturalist* **174**, S1-S14.

Otto SP, Orive ME (1995) Evolutionary consequences of mutation and selection within an individual. *Genetics* **141**, 1173-1187.

Papke RT, Koenig JE, Rodriguez-Valera F, Doolittle WF (2004) Frequent recombination in a saltern population of *Halorubrum*. *Science* **306**, 1928-1929.

Pawlowska TE (2005) Genetic processes in arbuscular mycorrhizal fungi. *Fems Microbiology Letters* **251**, 185-192.

Pawlowska TE, Taylor JW (2004) Organization of genetic variation in individuals of arbuscular mycorrhizal fungi. *Nature* **427**, 733-737.

Pawlowska TE, Taylor JW (2005) Arbuscular mycorrhizal fungi - Hyphal fusion and multigenomic structure - Reply. *Nature* **433**, E4-E4.

Pepper JW, Herron MD (2008) Does biology need an organism concept? *Biological Reviews* **83**, 621-627.

Petit R, Vendramin, GG (2007) Plant phylogeography based on organelle genes: an introduction. In: *Phylogeography of Southern European Refugia: Evolutionary perspectives on the origins and conservation of European biodiversity* (ed. Weiss S, Ferrand, N.), pp. 23-97. Springer, Heidelberg.

Pineda-Krch M, Lehtila K (2004a) Challenging the genetically homogeneous individual. *Journal of Evolutionary Biology* **17**, 1192-1194.

Pineda-Krch M, Lehtila K (2004b) Costs and benefits of genetic heterogeneity within organisms. *Journal of Evolutionary Biology* **17**, 1167-1177.

Pineda-Krch M, Poore AGB (2004) Spatial interactions within modular organisms: genetic heterogeneity and organism fitness. *Theoretical Population Biology* **66**, 25-36.

Pivato B, Mazurier S, Lemanceau P*, et al.* (2007) Medicago species affect the community composition of arbuscular mycorrhizal fungi associated with roots. *New Phytologist* **176**, 197-210.

Posada D (2008) jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253-1256.

Pradeu T (2010) What is an organism? An immunological answer. *History of Philosophy of the Life Sciences* **32**, 247-267.

Pringle A, Moncalvo JM, Vilgalys R (2000) High levels of variation in ribosomal DNA sequences within and among spores of a natural population of the arbuscular mycorrhizal fungus *Acaulospora colossica*. *Mycologia* **92**, 259-268.

Queller DC, Strassmann JE (2009) Beyond society: the evolution of organismality. *Philosophical Transactions of the Royal Society B: Biological Sciences* **364**, 3143-3155.

Raghunathan A, Ferguson HR, Jr., Bornarth CJ*, et al.* (2005) Genomic DNA amplification from a single bacterium. *Applied and Environmental Microbiology* **71**, 3342-3347.

Rand DM (2001) The units of selection on mitochondrial DNA. *Annual Review of Ecology and Systematics* **32**, 415-448.

Rapp R, Udall J, Wendel J (2009) Genomic expression dominance in allopolyploids. *Bmc Biology* **7**, 18.

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

Remy W, Taylor TN, Hass H, Kerp H (1994) 4-hundred-million-year-old vesicular-arbuscular mycorrhizae. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 11841-11843.

Richter D, Ott F, Auch A, Schmid R, Huson D (2008) MetaSim - a sequencing simulator for genomics and metagenomics. *PLoS ONE* **3**, e3373.

Rodriguez A, Clapp JP, Dodd JC (2004) Ribosomal RNA gene sequence diversity in arbuscular mycorrhizal fungi (Glomeromycota). *Journal of Ecology* **92**, 986-989.

Rodriguez A, Dougall T, Dodd JC, Clapp JP (2001) The large subunit ribosomal RNA genes of *Entrophospora infrequens* comprise sequences related to two different glomalean families. *New Phytologist* **152**, 159-167.

Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, Nyren P (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Analytical biochemistry* **242**, 84-89.

Ronaghi M, Uhlen M, Nyren P (1998) A sequencing method based on real-time pyrophosphate. *Science* **281**, 363-365.

Roper M, Ellison C, Taylor JW, Glass NL (2011) Nuclear and genome dynamics in multinucleate ascomycete fungi. *Current Biology* **21**, R786-R793.

Rosendahl S (2008) Communitites, populations and individuals of arbuscular mycorrhizal fungi. *New Phytologist* **178**, 253-266.

Rosendahl S, Sen R (1992) Isozyme analysis of mycorrhizal fungi and their mycorrhiza. In: *Methods in Microbiology* (eds. Norris JR, Read DJ, Varma AK), pp. 169-194.

Rosendahl S, Stukenbrock EH (2004) Community structure of arbuscular mycorrhizal fungi in undisturbed vegetation revealed by analyses of LSU rDNA sequences. *Molecular Ecology* **13**, 3179-3186.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R (2003) DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* **19**, 2496-2497.

Sanders IR (2002) Ecology and evolution of multigenomic arbuscular mycorrhizal fungi. *The American Naturalist* **160**, S128-S141.

Sanders IR (2004) Intraspecific genetic variation in arbuscular mycorrhizal fungi and its consequences for molecular biology, ecology, and development of inoculum. *Canadian Journal of Botany-Revue Canadienne De Botanique* **82**, 1057-1062.

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

Sanders IR, Croll D (2010) Arbuscular Mycorrhiza: the challenge to understand the genetics of the fungal partner. *Annual Review of Genetics, Vol 44* **44**, 271-292.

Santelices B (1999) How many kinds of individual are there? *Trends in Ecology & Evolution* **14**, 152-155.

Santelices B (2004a) A comparison of ecological responses among aclonal (unitary), clonal and coalescing macroalgae. *Journal of Experimental Marine Biology and Ecology* **300**, 31-64.

Santelices B (2004b) Mosaicism and chimerism as components of intraorganismal genetic heterogeneity. *Journal of Evolutionary Biology* **17**, 1187-1188.

Schatz G (2007) The magic garden. *Annual Review of Biochemistry* **76**, 673-678.

Schloss PD (2008) Evaluating different approaches that test whether microbial communities have the same structure. *Isme Journal* **2**, 265-275.

Schloss PD, Gevers D, Westcott SL (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE* **6**, e27310.

Schloss PD, Westcott SL, Ryabin T*, et al.* (2009) Introducing Mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75**, 7537-7541.

Schnare MN, Damberger SH, Gray MW, Gutell RR (1996) Comprehensive comparison of structural characteristics in eukaryotic cytoplasmic large subunit (23 S-like) ribosomal RNA. *Journal of Molecular Biology* **256**, 701-719.

Schrider DR, Hahn MW (2010) Gene copy-number polymorphism in nature. *Proceedings of the Royal Society B: Biological Sciences* **277**, 3213-3221.

Schussler A, Gehrig H, Schwarzott D, Walker C (2001a) Analysis of partial Glomales SSU rRNA gene sequences: implications for primer design and phylogeny. *Mycological Research* **105**, 5-15.

Schussler A, Schwarzott D, Walker C (2001b) A new fungal phylum, the Glomeromycota: phylogeny and evolution. *Mycological Research* **105**, 1413-1421.

Schwarzott D, Walker C, Schussler A (2001) *Glomus*, the largest genus of the arbuscular mycorrhizal fungi (Glomales), is nonmonophyletic. *Molecular Phylogenetics and Evolution* **21**, 190-197.

Seddas PMA, Arnould C, Tollot M, Arias CM, Gianinazzi-Pearson V (2008) Spatial monitoring of gene activity in extraradical and intraradical developmental stages of arbuscular mycorrhizal fungi by direct fluorescent in situ RT-PCR. *Fungal Genetics and Biology* **45**, 1155-1165.

Sędzielewska KA, Fuchs J, Temsch EM*, et al.* (2011) Estimation of the *Glomus intraradices* nuclear DNA content. *New Phytologist* **192**, 794–797.

Shade A, Handelsman J (2011) Beyond the Venn diagram: the hunt for a core microbiome. *Environmental Microbiology* **14**, 4-12.

Simon L, Bousquet J, Levesque RC, Lalonde M (1993) Origin and diversification of endomycorrhizal fungi and coincidence with vascular land plants. *Nature* **363**, 67-69.

Simon UK, Weiss M (2008) Intragenomic variation of fungal ribosomal genes is higher than previously thought. *Molecular Biolology and Evolution* **25**, 2251-2254.

Smith GP (1976) Evolution of repeated DNA sequences by unequal crossover. *Science* **141**, 528-534.

Smith S, Dickson S (1997) *VA mycorrhizas: basic research and techniques* Cooperative Research Centre for Soil and Land Management, Glen Osmond.

Smith SE, Read DJ (2008) *Mycorrhizal Symbiosis*, Third Edition edn. Academic Press, London.

Sokolski S, Dalpe Y, Piche Y (2010) Phosphate transporter genes as reliable gene marker for the identification and discrimination of arbuscular mycorrhizal fungi in the genus *Glomus*. *Applied and Environmental Microbiology* **77**, 1888-1891.

St-Arnaud M, Hamel C, Vimard B, Caron M, Fortin JA (1995) Altered growth of *Fusarium oxysporum* f. *sp. chrysanthemi* in an in vitro dual culture system with the vesicular arbuscular mycorrhizal fungus *Glomus intraradices* growing on *Daucus carota* transformed roots *Mycorrhiza* **5**, 431-438.

St-Arnaud M, Vujanovic V (2007) Effects of the arbuscular mycorrhizal symbiosis on plant diseases and pests. In: *Mycorrhizae in Crop Production* (eds. Hamel C, Plenchette C), pp. 67-122. Haworth Press, Binghampton, New York.

St.-Arnaud M, Hamel C, Vimard B, Caron M, Fortin JA (1996) Enhanced hyphal growth and spore production of the arbuscular mycorrhizal fungus Glomus intraradices in an in vitro system in the absence of host roots. *Mycological Research* **100**, 328–332.

Steinke D, Vences M, Salzburger W, Meyer A (2005) TaxI: a software tool for DNA barcoding using distance methods. *Philosophical Transactions of the Royal Society B-Biological Sciences* **360**, 1975-1980.

Stewart CN, Excoffier L (1996) Assessing population genetic structure and variability with RAPD data: Application to *Vaccinium macrocarpon* (American Cranberry). *Journal of Evolutionary Biology* **9**, 153-171.

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

Strassmann JE, Queller DC (2004) Genetic conflicts and intercellular heterogeneity. *Journal of Evolutionary Biology* **17**, 1189-1191.
Stukenbrock EH, Rosendahl S (2005a) Clonal diversity and population genetic structure of arbuscular mycorrhizal fungi (*Glomus* spp.) studied by multilocus genotyping of single spores. *Molecular Ecology* **14**, 743-752.
Stukenbrock EH, Rosendahl S (2005b) Development and amplification of multiple co-dominant genetic markers from single spores of arbuscular mycorrhizal fungi by nested multiplex PCR. *Fungal Genetics and Biology* **42**, 73-80.
Summers K, da Silva J, Farwell MA (2002) Intragenomic conflict and cancer. *Medical Hypotheses* **59**, 170-179.
Swofford DL, Olsen GJ, Waddell PJ, Hillis DM (1996) Phylogenetic inference. In: *Molecular Systematics* (eds. Hillis DM, Moritz C, Mable BK), pp. 407–514. Sinauer Associates, Inc., Sunderland, MA.
Tamminen M, Virta M, Fani R, Fondi M (2012) Large-scale analysis of plasmid relationships through gene-sharing networks. *Molecular Biology and Evolution* **29**, 1225-1240.
Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular evolutionary genetics analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599.
Thornhill DJ, Lajeunesse TC, Santos SR (2007) Measuring rDNA diversity in eukaryotic microbial systems: how intragenomic variation, pseudogenes, and PCR artifacts confound biodiversity estimates. *Molecular Ecology* **16**, 5326-5340.
Tisserant E, Kohler A, Dozolme-Seddas P*, et al.* (2012) The transcriptome of the arbuscular mycorrhizal fungus *Glomus intraradices* (DAOM 197198) reveals functional tradeoffs in an obligate symbiont. *New Phytologist* **193**, 755-769.
Van der Heijden MGA, Klironomos JN, Ursic M*, et al.* (1998) Mycorrhizal fungal diversity determines plant biodiversity, ecosystem variability and productivity. *Nature* **396**, 69-72.
Van Der Heijden MGA, Sanders IR (2002) *Mycorrhizal Ecology*, 1 edn. Springer-Verlag, Heidelberg, Germany.
Van der Heijden MGA, Scheublin TR (2007) Functional traits in mycorrhizal ecology: their use for predicting the impact of arbuscular mycorrhizal fungal communities on plant growth and ecosystem functioning. *New Phytologist* **174**, 244-250.
Van der Heijden MGA, Wiemken A, Sanders IR (2003) Different arbuscular mycorrhizal fungi alter coexistence and resource distribution between co-occurring plant. *New Phytologist* **157**, 569-578.
Vandenkoornhuyse P, Leyval C, Bonnin I (2001) High genetic diversity in arbuscular mycorrhizal fungi: evidence for recombination events. *Heredity* **87**, 243-253.
Viera A, Glenn M (1990) DNA content of vesicular-arbuscular mycorrhizal fungal spores. *Mycologia* **82**, 263-267.
Villesen P (2007) FaBox: an online toolbox for fasta sequences. *Molecular Ecology Notes* **7**, 965-968.
Voets L, de la Providencia IE, Declerck S (2006) Glomeraceae and Gigasporaceae differ in their ability to form hyphal networks. *New Phytologist* **172**, 185-188.
Vogelsang KM, Reynolds HL, Bever JD (2006) Mycorrhizal fungal identity and richness determine the diversity and productivity of a tallgrass prairie system. *New Phytologist* **172**, 554-562.
Wang B, Qiu YL (2006) Phylogenetic distribution and evolution of mycorrhizas in land plants. *Mycorrhiza* **16**, 299-363.
Wang B, Yeun LH, Xue J-Y*, et al.* (2010) Presence of three mycorrhizal genes in the common ancestor of land plants suggests a key role of mycorrhizas in the colonization of land by plants. *New Phytologist* **186**, 514-525.
Wang GCY, Wang Y (1996) The frequency of chimeric molecules as a consequence of PCR co-amplification of 16S rRNA genes from different bacterial species. *Microbiology* **142**, 1107-1114.

Wang GCY, Wang Y (1997) Frequency of formation of chimeric molecules is a consequence of PCR coamplification of 16S rRNA genes from mixed bacterial genomes. *Applied and Environmental Microbiology* **63**, 4645-4650.

Weismann A (1893) *The Germ-Plasm: a Theory of Heredity* Charles Scribner's Sons, New York.

Whitney KD, Boussau B, Baack EJ, Garland T, Jr. (2011) Drift and genome complexity revisited. *PLoS Genetics* **7**, e1002092.

Whitney KD, Garland T, Jr. (2010) Did genetic drift drive increases in genome complexity? *PLoS Genetics* **6**, e1001080.

Williams GC (1992) *Natural Selection: Domains, Levels, and Challenges.* Oxford University Press, New York.

Wilson DS, Sober E (1994) Reintroducing group selection to the human behavioral sciences. *Behavioral and Brain Sciences* **17**, 585-608.

Winther JL, Friedman WE (2008) Arbuscular mycorrhizal associations in Lycopodiaceae. *New Phytologist* **177**, 790-801.

Wright S (1931) Evolution in Mendelian populations. *Genetics* **16**, 97–159.

Wright S (1938) Size of population and breeding structure in relation to evolution. *Science* **87**, 430-431.

Wright S, Finnegan D (2001) Genome evolution: Sex and the transposable element. *Current Biology* **11**, R296-R299.

Zézé A, Hosny M, Gianinazzi-Pearson V, Dulieu H (1996) Characterization of a highly repeated DNA sequence (SC1) from the arbuscular mycorrhizal fungus *Scutellospora castanea* and its detection *in planta*. In: *Applied and Environmental Microbiology*, pp. 2443-2448.

Zhang JZ (2003) Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18**, 292-298.

Zimmerman E (2009) *Extent of intra-isolate genetic polymorphism in Glomus etunicatum using a molecular genetic approach* M.Sc. thesis, University of Montreal.

Zuker M (2003) Mfold web server for nucleic acid folding and hybridization prediction. *Nucleic Acids Research* **31**, 3406-3415.

Zylstra P, Rothenfluh HS, Weiller GF, Blanden RV, Steele EJ (1998) PCR amplification of murine immunoglobulin germline V genes: Strategies for minimization of recombination artefacts. *Immunology and Cell Biology* **76**, 395-405.