Resolving the haplotypes of arbuscular mycorrhizal fungi highlights the role of two nuclear populations in host interactions

Jana Sperschneider [1*], Gokalp Yildirir [2*], Mathu Malar C [2], Ariane Mayrand Nicol [2], Essam Sorwar [2], Eric CH Chen [3], Elizabeth K. Brauer [2, 4], Whynn Bosnich [4] and Nicolas Corradi [2 #]

[1] Black Mountain Science and Innovation Park, CSIRO Agriculture and Food, Canberra, Australia

[2] Department of Biology, University of Ottawa, ON, Ottawa, K1N 6N5, Canada

[3] The University of Tokyo, Department of Integrated Biosciences, Graduate School of Frontier Sciences

[4] Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, Ottawa, ON, Canada; Department of Biology, University of Ottawa, Ottawa, ON, Canada

* Contributed equally

Corresponding Author

Abstract

Arbuscular mycorrhizal fungi (AMF) are prominent root symbionts with a multinucleate cytoplasm that can carry thousands of nuclei deriving from two parental strains and varying in relative abundance in a large syncytium. Here, we set out to improve our understanding of such remarkable genetics by resolving the nuclear genomes of all publicly available AMF heterokaryons using PacBio HiFi and Hi-C sequencing. We find that all AMF heterokaryons carry two sets of homologous chromosomes, where genes associated with plant colonization reside in gene-sparse, repeat-rich compartments. The co-existing nuclear genomes are phylogenetically related but differ significantly in content and epigenetics, resulting in nucleus-specific regulatory programs during mycorrhizal interactions. AMF heterokaryons carry signatures of past genetic exchange indicative of sexual reproduction, followed by clonal haplotype evolution. This work uncovers the contribution and origin of nuclear genomes present in AMF heterokaryons and opens avenues for improvement and environmental application of these strains.

Introduction

Arbuscular mycorrhizal fungi are prominent root symbionts of the subphylum Glomeromycotina [1] that improve access to nutrients and resistance against pathogens in most land plants [2]. These keystone mutualists increase ecosystem productivity in global terrestrial ecosystems [3] and many industries specialize in their production as bio-stimulants for agricultural practices, forestry, and plant nurseries. AMF can contain over 20,000 nuclei within individual spores and, as opposed to other multinucleate fungi, there is no stage in their life cycle where cells contain one or two nuclei [4]. Fossil and phylogenomic evidence indicates this plant-fungus symbiosis dates back more than 400 million years [5,6], but despite this longevity no sexual apparatus has been observed in these organisms. Because of this AMF were long pigeonholed in a selected group of organisms called "ancient asexuals" [7], and this absence of sexual reproduction was proposed to be offset by high degree of sequence divergence among their co-existing nuclei [8]. The view that AMF have a purely clonal lifestyle and non-mendelian genetics has, however, been challenged by discoveries of gene orthologues that function in meiosis [9], and more recently by evidence that some strains of the model species *Rhizophagus irregularis*, referred to as AMF heterokaryons (or AMF dikaryons) carry thousands of nuclei deriving from two parental strains [10]. This genetic state is analogous to the heterokaryotic (sexual) stage of the Dikarya (ascomycete and basidiomycete fungi) where two or more parental nuclei co-exist in one cell for multiple generations [11], and indeed AMF heterokaryons share remarkable genomic features, ecological traits and nuclear dynamics with heterokaryotic fungal relatives [4].

For example, the parental nuclei of fungal heterokaryons carry diverging mating-type loci (MAT-loci) that orchestrate sexual interactions between compatible strains, and loci with remarkable similarity are also present in nuclei with distinct genotypes in AMF heterokaryons [10,12]. The co-existence of two nuclear types in fungal heterokaryons improves fitness and adaptability to environmental change, and this ability correlates with change in the relative ratio of co-existing nuclei within the mycelium [13,14]. Consistent with this, AMF heterokaryons outperform homokaryotic relatives in many life-history traits [15], including hyphal growth speed and density, and their two haplotypes also vary in relative abundance depending on environmental factors and host identity [16,17].

In the dikaryotic basidiomycete *Agaricus bisporus* and the heterokaryons of *Neurospora tetrasperma*, two parental nuclei express different genes throughout the life cycle [18,19], and in other species allele-specific expression is tightly linked to the ratio of the two nuclear types [20]. These findings show that parental genomes play distinct molecular roles in the biology and nuclear dynamics of fungal heterokaryons, but whether similar mechanisms also exist in AMF heterokaryons is unknown because sequence information on parental nuclei is unavailable. As a result, their relative genetic and epigenetic contribution to the biology of these strains and their mycorrhizal partners could thus far not be investigated.

To address these questions, we resolved the parental nuclear sequences of all publicly available AMF heterokaryons using PacBio High Fidelity (HiFi) reads and Hi-C sequencing and analyzed their gene regulation across multiple hosts and conditions. Untangling each of the haploid genomes also allowed determination of the evolutionary origin of these strains, and examination of whether AMF heterokaryons carry two dominant parental genomes that physically separate across thousands of nuclei, as proposed by Ropars and colleagues [10].

Results

Nuclear genome phasing and annotation of AMF heterokaryons

HiFi PacBio long-read and Hi-C sequencing data was obtained for all publicly available AMF heterokaryons (A4, DAOM-664343; A5, DAOM-664344; G1, DAOM-970895; SL1, DAOM-240409). This data was assembled with hifiasm in Hi-C integration mode [21], which returned two haplotypes for each strain. Phasing of the assembled haplotypes was assessed with NuclearPhaser [22] and after minor phase switching correction the assembly was scaffolded into 32 chromosome-scale scaffolds, guided by Hi-C data (**Table 1**, **Figure 1, Supplemental Figures 1-4)**..

Table 1: General genome assembly statistics for the AMF heterokaryons.

	A4		**A5**		**G1**		**SL1**	
Statistics	**Hap 1**	**Hap2**	**Hap1**	**Hap2**	**Hap1**	**Hap2**	**Hap1**	**Hap2**
Assembly size	159.1 Mb	158.7 Mb	147.0 Mb	146.0 Mb	143.5 Mb	145.3 Mb	150.2 Mb	148.7 Mb
# of scaffolds	32	32	32	32	32	32	32	32
Mating type locus identity	MAT-2	MAT-1	MAT-6	MAT-3	MAT-5	MAT-1	MAT-5	MAT-1
Average HiFi read coverage	85.8	78.8	111.3	108.4	124.1	76.1	83.6	80.6
GC content	27.8%	27.8%	27.8%	27.8%	27.5%	27.5%	27.36%	27.38%
Complete BUSCOs	97.4%	96.4%	96.4%	96.4%	97.0%	97.0%	94.9%	92.7%
Duplicated BUSCOs	1.7%	1.9%	1.6%	1.5%	1.9%	2.1%	1.7%	1.9%
# Genes	22,978	22,998	20,563	20,424	20,532	20,765	22,089	21,901
% Repetitive sequence	54.5%	54.5%	52.1%	52%	51.6%	51.5%	52.7%	52.8%
Aligned bases between haplotypes	98.4%		98.4%		98.7%		96.2%	
Average identity of 1-to-1 alignments	99.5%		99.5%		99.6%		98.8%	
Total SNPs	261,818		263,443		233,618		562,789	

On average, only 54 short contigs totaling 2.7 Mb (L50: 61 Kb) remained un-scaffolded per strain. Consistent with phasing of fungal dikaryons based on Hi-C data [22–24], the haplotypes of AMF heterokaryons exhibit a strong dikaryotic phasing signal, with ~90% of inter-chromosome (*trans*) Hi-C contacts contained within nuclear haplotypes (**Supplemental Figures 1-4)**. In our assemblies, one part of the chromosome 28 previously identified in AMF homokaryons [25] maps to chromosome 2 and the other to chromosome 17, but inspection of Hi-C contact maps and HiFi read coverage did not provide conclusive evidence that these chromosomes should be broken. The identification of AMF centromeres will ultimately be necessary to determine if *Rhizophagus irregularis* carries 32 or 33 chromosomes [26], and for ease of comparative analysis we numbered the chromosomes according to the homologous chromosomes of AMF homokaryons [25]

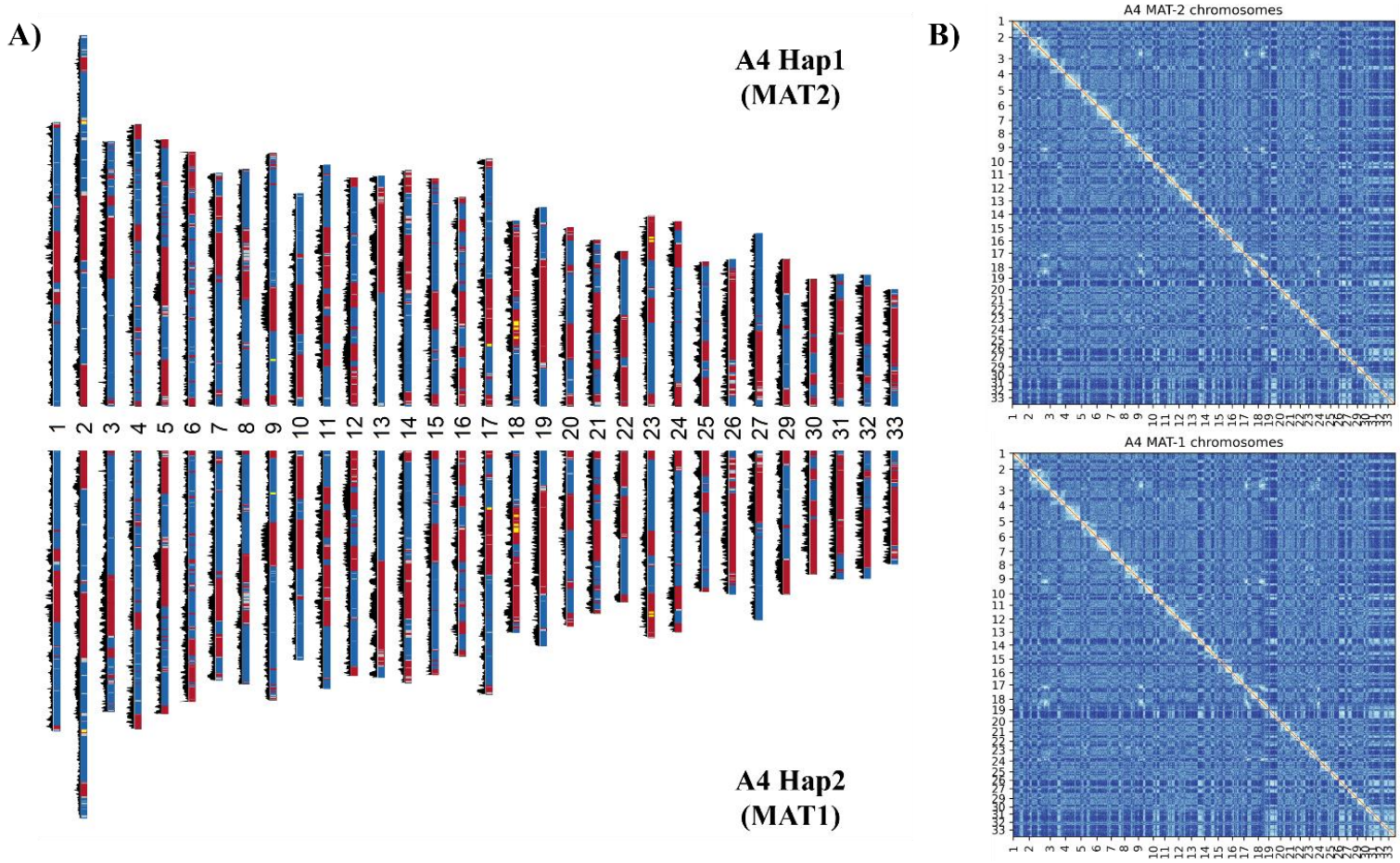

Figure 1: The phased chromosomes of the AMF heterokaryons. (A) The chromosomes of the A4 haplotypes. Gene density is shown in black (20 Kb windows). A and B compartments are shown in red and blue, respectively. rRNA operons are shown in yellow. All haplotypes carry 32 chromosomes and A/B compartments. (B) The Hi-C contact map of the A4 haplotypes shows a compartmentalization of the chromosomes into A and B compartments. A 'checkered' pattern, resulting from the interactions of several chromosomes, indicates the genome arrangement of the euchromatin or heterochromatin compartments of *R. irregularis*. Data from other isolates is available in Supplemental Figures 2-4.

Phased haplotypes have genome sizes and number of genes and repeats in expected range for the species [10] , and their complete BUSCO gene repertoire averages 96% (**Table 1**). Unlike other fungi, AMF do not carry thousands of homogenous ribosomal operons (rRNA), but harbor up to 11 divergent rRNA copies on each

haplotype [25] (**Figure 1**, **Supplemental Table 1**). Consistent with molecular analyses of single spores [17], the MAT-2 and MAT-5 haplotypes have higher coverage (relative abundance) in the isolates A4 and G1, respectively (~9.2% and 63% higher) but are closer to balance in A5 and SL1 in extraradical mycelium after propagation with *Daucus carota cv* P68 (Carrot; 3.5% average difference) (**Table 1**). The Hi-C contact maps confirm that chromosomes of AMF heterokaryons separate into two dominant euchromatin (A) and heterochromatin (B) compartments known to dictate the genome biology and evolution of these symbionts [25]. The A-compartment has higher gene density, while the B-compartment has higher repeat density and is enriched in secreted proteins and candidate effectors, including mycFold effectors [27] (**Figure 1**, **Supplemental Figures 2-4**, **Supplemental Table 2, Supplemental Table 3**).

Phylogenetic and population analyses of AMF heterokaryons

The phased haplotypes were first used to assess the evolutionary origin of AMF heterokaryons. A maximum likelihood phylogeny was generated using available genome and RAD-seq data from 65 strains catalogued as *R. irregularis* [25,28] (**Figure 2**), which revealed at least 9 fully supported phylogenetic clades for this species.

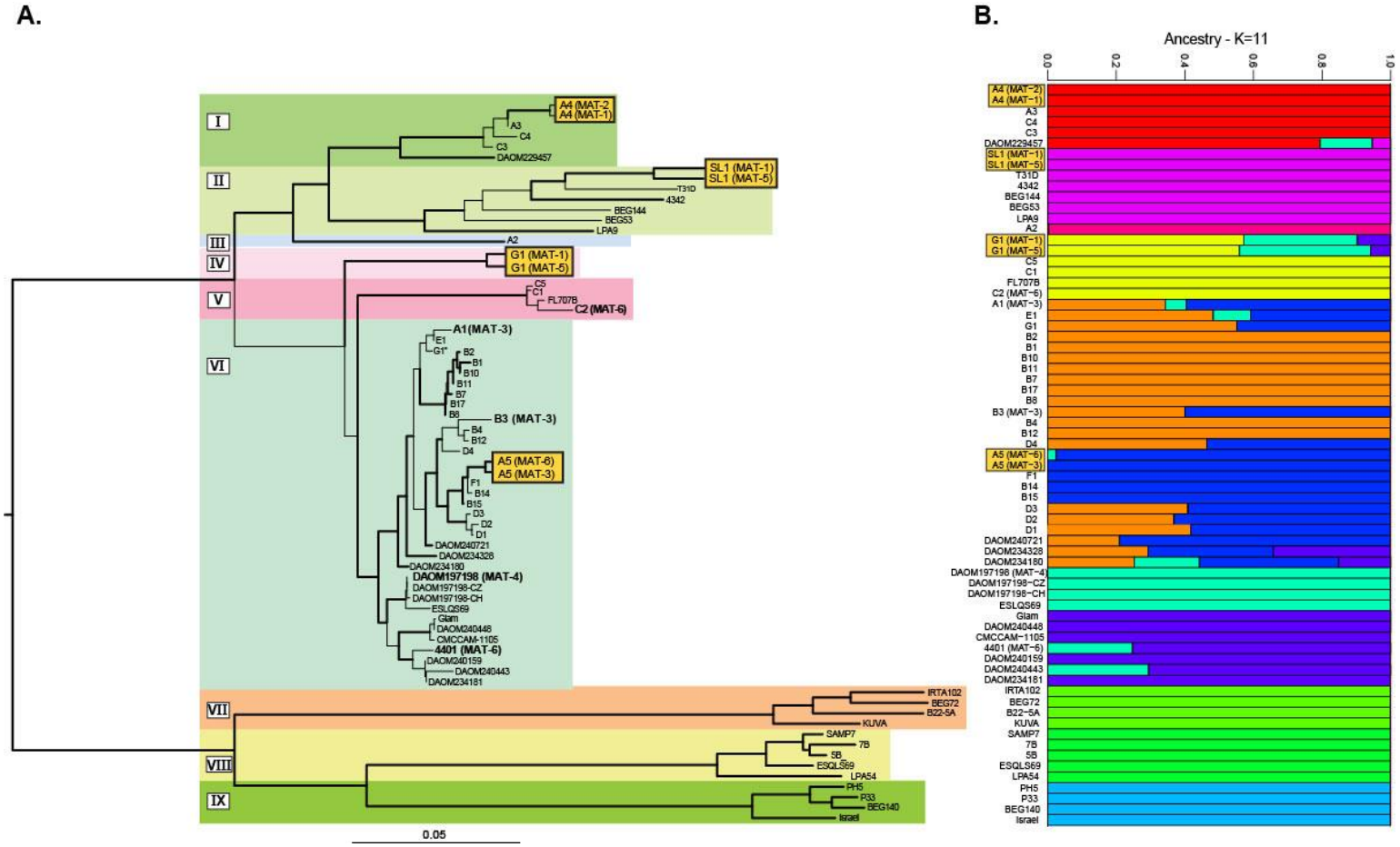

Figure 2: A. Phylogenetic tree constructed with 65 *R. irregularis* strains (incl. 8 haplotypes). Haplotypes from AMF heterokaryons are shown in yellow squares. Data obtained using chromosome-level genome assemblies are shown in bold. Based on relative branch lengths, the phylogeny resolves 9 clades which are highlighted in color. Thicker branches indicate bootstrap support > 95. The tree was made using IQTREE algorithm, in GTR-FO mode with 1000 bootstrap replicates. Scale bar represents 0.04 substitutions per site. The G1 strain located in clade VI and noted with an asterisk is homokaryotic and does not represent the heterokaryotic isolate G1 (DAOM-970895) from clade IV. B. Population structure of *R. irregularis* at K = 11. The structure has been inferred using NGSadmix. Each row represents a strain, listed as in the ML tree (A). Colored bars represent their coefficients of membership in the various gene pools based on SNP data.

Consistent with past analyses, *R. irregularis* strains did not cluster in the phylogeny according to geographical location or mating-type identity [28,29]. All AMF heterokaryons cluster into separate clades, but their co-existing haplotypes always group together and away from other sequenced strains, meaning that their parental homokaryotic strains could not be identified with available datasets. Long branch lengths indicative of long-term clonal haplotype divergence separate some co-existing haplotypes, especially G1 and SL1. Inspection of aligned homologous sequences used to produce the phylogeny did not reveal shifts in SNP tracts shared between of co-existing haplotypes and their closely related strains that could indicate the presence of recombination following the formation of AMF heterokaryons. However, evidence of admixture indicative of past genetic exchanges (e.g., recombination, introgression) exists between multiple strains. In AMF heterokaryons, evidence of admixture is found in both haplotypes of G1, and a small portion of the MAT-6 haplotype in A5.

Comparative analyses of co-existing haplotypes

Large-scale chromosome reassortments occur among co-existing haplotypes, and at smaller genome scale variability in gene order and content is common. While homologous chromosomes can experience inversions and translocations (**Figure 3a, Supplemental Figure 5**), at the putative AMF mating type locus only five genes (a hypothetical protein, a choline transporter, the homeodomains HD-1 and HD2, and a phosphoglycerate mutase) are colinear among all haplotypes on chromosome 11 (**Figure 3b**).

AMF heterokaryon haplotypes also differ in relative gene function and number of repeats. For instance, in the strain A4 we observed an enrichment of the tetratricopeptide repeat domain Sel1 on MAT-2, whereas an enrichment of protein kinase domains is found on MAT-1 **(Supplemental Figure 6, Supplemental Figure 7)**. All strains also carry genes specific to each haplotype. These include candidate effectors and secreted proteins potentially involved in the dialogue between the mycorrhizal partners, as well as rRNA operons and repeats. In the strain G1, the MAT-5 haplotype has one more rRNA operon, but 233 less genes than the MAT-1 haplotype (**Table 1, Supplemental Table 1**). Moreover, co-existing haplotypes can differ in chromosomal compartmentalization (**Supplemental Figure 8).**

Despite these differences, haplotypes are more similar in content and chromosome size than different *R. irregularis* strains are to one another [25]. For instance, co-existing haplotypes differ by an average of 145 genes and 578 kilobases (Kb) of repeats in AMF heterokaryons compared to 2,646 genes and 16.4 megabases (Mb) of repeats between the AMF homokaryons, and their homologous chromosomes differ in size by just over 100 Kb compared to over 550 Kb for AMF homokaryons [25].

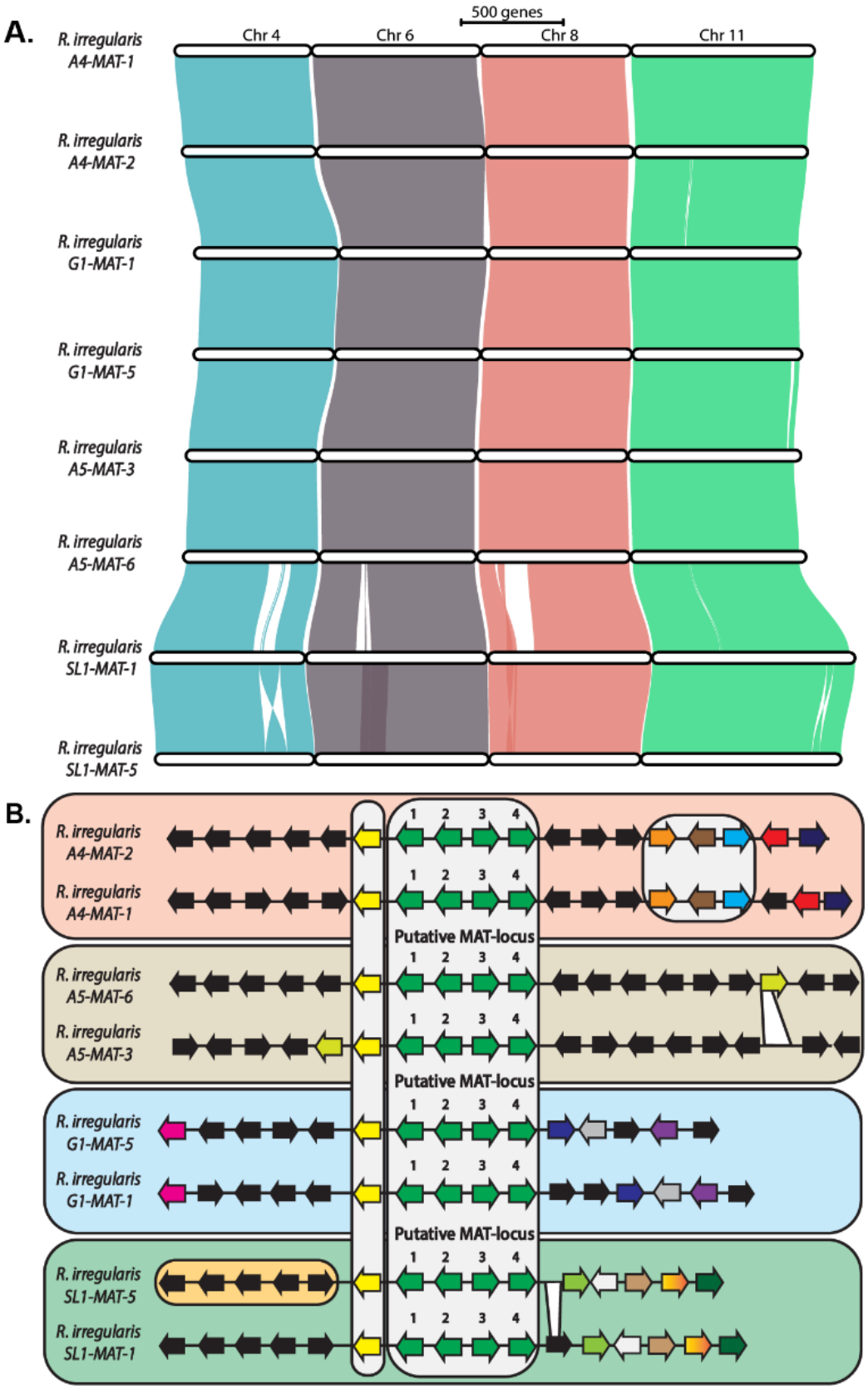

Figure 3. A. Gene synteny and reassortments in representative chromosomes (4, 6, 8 and 11) of AMF heterokaryons. Similar patterns affect all chromosomes. B. Gene order is relaxed in regions surrounding the putative MAT-locus of all AMF heterokaryons. The four genes thought to compose the AMF MAT-locus (located on chromosome 11) are shown as green arrows and highlighted by numbers as follows; 1. Choline transporter, 2. HD1, 3. HD2, 4. phosphoglycerate mutase). Arrows indicate gene transcriptional direction and color highlights sequence orthology among haplotypes. Black arrows indicate genes showing no evidence of orthology among co-existing haplotypes. SL1-MAT-5 orthologs within orange ellipse are located 20Kb upstream in the MAT-1 haplotype.

Haplotype-specific gene expression and gene regulation

We hypothesized that the observed genetic and epigenetic differences affect the transcriptome of AMF heterokaryons. To assess this, we obtained RNA-sequencing data for the strains A4, A5 and G1 from three conditions: extraradical mycelium, colonized roots of *D. carota* cv. P68 (carrot) and colonized roots of *C. intybus* (chicory). We first assessed the localization of differentially expressed genes on the chromosomes, i.e. the A/B compartments [25], and detected up-regulation for 17% of all genes in the A compartment during *in planta* colonization (vs extraradical mycelium), compared to 11% of all genes in the B compartment (**Figure 4a**). However, for genes encoding secreted proteins, we detected up-regulation during *in planta* colonization (vs extraradical mycelium) in equal proportions in the A and B compartments (26% of genes in the B compartment, 27% of genes in the A compartment). Furthermore, in all strains over ~40% of the genes encoding secreted proteins and all but one mycFOLD effector upregulated *in-planta* are in the B compartment. We also detected a small number of genes that are up-regulated during *in planta* colonization in one host compared to the other and again, these host-specific induced genes are preferentially located in the repeat-rich B-compartment but also preferentially on one of the haplotypes (**Figure 4a**). Taken together, this supports the hypothesis that the gene-sparse, repeat-rich B-compartment plays an important role during *in planta* colonization and in the regulation of the secretome between carrot and chicory.

To determine the relative contribution of each haplotype, we used approaches that conservatively measure the expression of orthogroups with exactly one gene copy on each haplotype and where the orthologs can be differentiated by at least one SNP in their coding sequences [18,19] (A4: 8,064 genes; A5: 7,152 genes; G1: 6,816 genes). Average expression and regulation of these 1-1 orthologs can differ significantly among co-existing haplotypes across hosts and conditions (**Figure 4b**. Overall, total and average gene expression are consistent with the relative haplotype abundance measured using digital droplet PCR (ddPCR) on the same colonized roots (**Supplemental Figure 9**), that is, the most abundant haplotype has higher total expression. Specifically, in the strains A5, A4 and G1, 1-1 orthologs from the MAT-3, MAT-2 and MAT-5 haplotypes have higher average gene expression across conditions (not significant with chicory for A5 and A4)

With few exceptions, haplotype-specific gene regulation follows this trend based on host identity (**Figure 4b**). For example, in A4, 80.3% of the 1-1 orthologs that are up-regulated in chicory compared to carrot are located on the MAT-1 haplotype which has slightly lower abundance in roots (**Supplemental Figure 9**). In contrast, 68.9% of the 1-1 orthologs that are up-regulated in *D. carota* compared to *C. intybus* are located on the more abundant MAT-2 haplotype. In A5, 63.4% of the 1-1 orthologs that are up-regulated in chicory compared to *D. carota* are located on the more abundant MAT-6 haplotype, and again the inverse pattern is seen for those up-regulated in carrot compared to chicory (74.6% on more abundant MAT-3 haplotype). Remarkably, in G1 78.6% of 1-1 orthologs on the MAT-1 haplotype are up-regulated with *D. carota* compared to chicory despite this strain having its nuclei belonging overwhelmingly to opposite mating-type (**Figure 4b**).

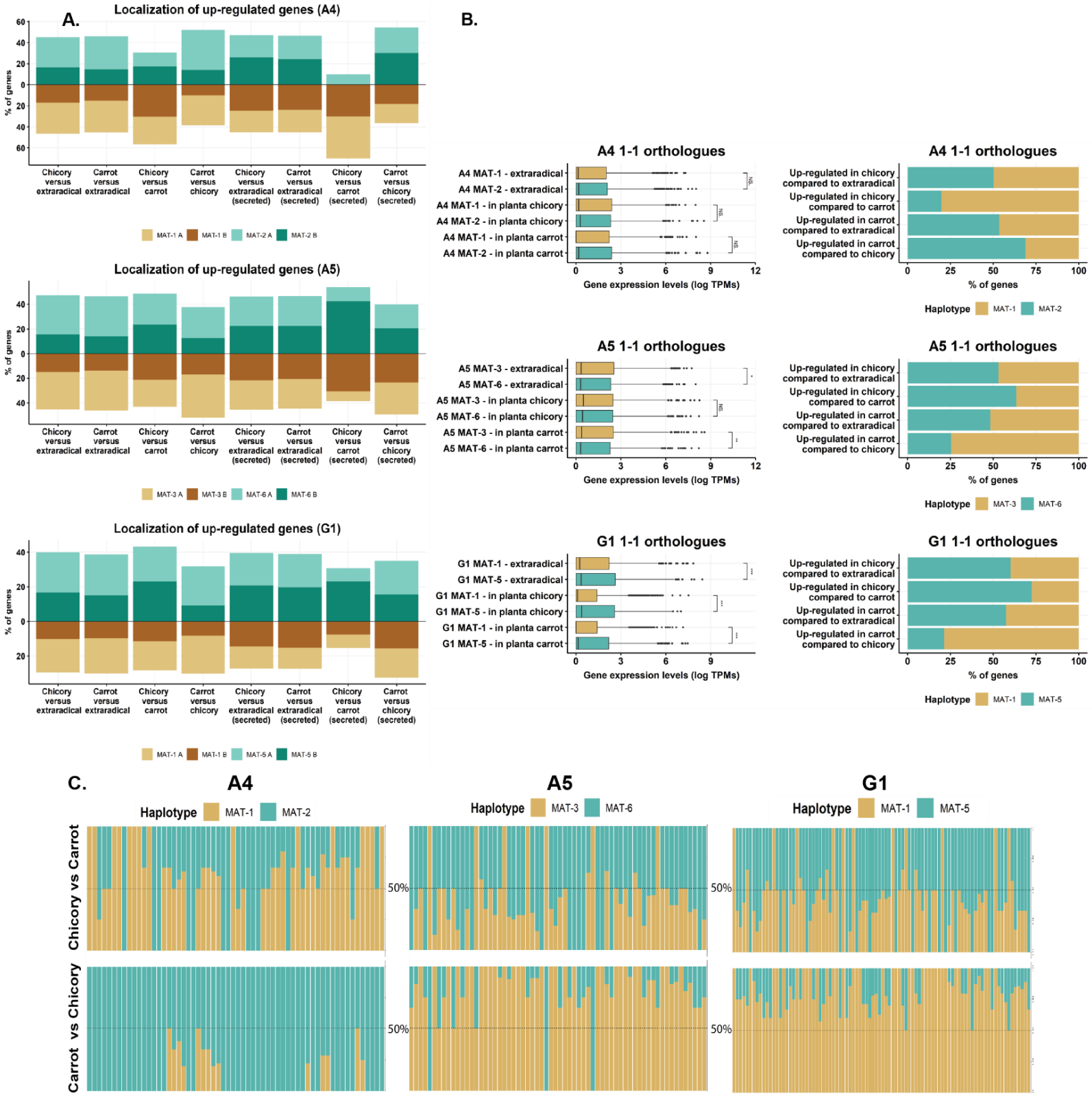

Figure 4: Haplotype specific RNA-seq analyses. A: Chromosome localization of gene upregulation. When all differentially regulated genes are considered, upregulation of secreted genes is higher in planta in the B-compartment. B. Haplotype-specific expression (left) and regulation of 1-1 orthologues (right). Left: Box plot edges show the first and third quartile in box plots, with the median shown as the middle lane. The data range is indicated by whiskers and dots represent outliers. Asterisks indicate a t-test statistical significance at $P < 0.05$. Right: % of genes upregulated in each haplotype across the three conditions used in this study (in planta-carrot, in planta-chicory and extraradical mycelium). C. GoTerm enrichment comparisons of 1-1 orthologues between haplotype specific genes upregulated in carrot versus chicory (upper box) and chicory vs carrot (bottom box) in A4, A5 and G1. Each bar represent percentage of upregulated GoTerm for the same GoTerm in each specific haplotype. Only GoTerms shared among haplotype-specific upregulated genes are shown to highlight the respective functional contribution of each haplotype across hosts. Shared GoTerms are listed on Supplemental Figure 10. GoTerm enrichments unique to each haplotype also exist (see Supplemental Figure 10).

The observed variability leads to distinct putative functional contributions for co-existing haplotypes across growth conditions (**Figure 4c**). For example, in the strain A4, the genes upregulated in the MAT-1 haplotype in *C. intybus* compared to *D. carota* are enriched for GO-terms involved in sporulation, secretion or stimuli response, but not in *D. carota* as those are primarily upregulated in the MAT-2 haplotype (**Figure 4c)**. The same patterns are seen in A5, where genes with putative functions in sexual reproduction, sporulation or signal transductions are upregulated primarily by genes located in the MAT-3 haplotype in *D. carota* and the MAT-6 haplotype in *C. intybus*. In contrast, comparisons against extraradical mycelium show a generally more balanced haplotype contribution to each GO-Term, and each haplotype also regulate unique functions during root colonization and in the extra radical mycelium (**Supplemental Figure 10**).

Discussion

The nuclear organization of AMF has been a long source of debate [30]. The recent identification of homokaryotic and heterokaryotic strains reconciled previous views in this area [10], but these findings were built on highly fragmented genome and nuclear datasets that may skew data interpretation and conclusions. Here, we show that combining PacBio HiFi long reads with Hi-C sequencing phases two complete sets of chromosomes in all AMF heterokaryons. These haplotypes physically separate among thousands of individual nuclei, which confirms current models that explain AMF genetics [10,30]. The absence of haplotype degeneration contrasts with predictions that long-term clonal evolution in AMF should lead to extensive nuclear divergence through accumulation of deleterious mutations [8,31,32]. What maintains haplotype integrity in the absence of formal reproduction in AMF needs further investigation, but this feature may be achieved via mutation suppression mechanisms recently reported in dikaryotic fungi [33].

The fragmentary and collapsed nature of genomic data from AMF heterokaryons also hampered the identification of their evolutionary origins. By untangling co-existing haplotypes, we were able to confirm that none of these strains are closely related to homokaryons with sequenced genomes [10], but emerged instead from phylogenetically close (and yet to be sequenced) homokaryotic relatives with distinct MAT-loci. This information now provides a clear roadmap that can be used to identify and cross suitable candidates and produce new strains tailored to plant growth.

Haplotype sequence data from dozens of *R. irregularis* strains supports evidence of past genetic exchange between individual strains of this species. These presumably arose via meiosis-like mechanisms, and similar events, possibly followed by clonal haplotype evolution, are also found in the parental nuclei of some AMF heterokaryons. In the absence of data from parental strains, and with no signatures of recombination involving co-existing haplotypes, only assumptions can be made for the emergence of AMF heterokaryons. As extensive divergence in gene order and structure exists among haplotypes, one possibility is that their haplotypes have yet to undergo meiotic recombination and emerged instead through somatic nuclear exchange;

a process that generates heterokaryotic progenies that seldom recombine via sexual reproduction in basidiomycetes [11,34], resulting in haplotypes evolving clonally over time. Alternatively, if sequence divergence allows recombination amongst haplotypes, then past mating between two reasonably diverged nuclei may have produced these heterokaryons, followed by some meiosis-like events and then clonal haplotype evolution for some time.

Our RNA-seq analyses clarified the relative contribution of co-existing haplotypes in AMF heterokaryons. In distant fungal heterokaryons, the co-existence and regulation of two parental genotypes improves fitness and adaptability to change [35–37]. The variability in content, expression, and epigenetics among haplotypes is thus likely to provide similar benefits to AMF heterokaryons. In support of this, changes in host identity or growth condition alter nuclear dynamics and expression in all strains, which can result in upregulation being primarily controlled by the least abundant haplotype. As such, our work indicates that co-existing haplotypes act as distinct and active regulatory units in AMF heterokaryons, and this may explain their ability to outgrow homokaryotic relatives across a wider host range [15,38]. Future studies should now investigate relative haplotype expression with additional varying parameter (temperature, pH, salinity, hosts) to obtain a more global view of the molecular interactions that occur between haplotypes and the surrounding environment.

Overall, this study resolved long-standing questions on the evolutionary genetics and origin of AMF heterokaryons, and now opens opportunities to investigate the nuclear biology and dynamics of these peculiar strains using microscopy and haplotype-specific probes. While evidence of past genetic exchanges found in many *R. irregularis* strains and haplotypes supports the view that AMF can recombine with close relatives, the identification and analysis of sequence data from parental homokaryotic strains will be required to determine whether AMF heterokaryons originate because of sexual or clonal processes (or both). Findings that two nuclear haplotypes govern the genome biology of these strains further underpins similarities between the genetics of AMF and relatives in the Dikarya and, within this context, it is noteworthy that recent phylogenomic analyses are now proposing these fungal lineages as being sister-clades [39,40].

Material and Methods

Culturing, DNA extraction and Hi-Fi and Hi-C sequencing

Rhizophagus irregularis strains A4, DAOM-664343; A5, DAOM-664344; G1, DAOM-970895; SL1, DAOM-240409) were cultured using either *Daucus carota* (*cv 68*) or *Cichorium intybus* root organ cultures (ROCs) as host [17,41]. The strains A4 and A5 were originally collected from a field in Tänikon [42]. The SL1 and G1 strains were originally collected by staff at Agriculture and Agri-Food Canada (Ottawa, Canada) in Sainte-Foy (Québec, Canada). All strains propagated in two-compartment ROCs, allowing us to produce mycelium and spores without obvious contaminants. High-molecular-weight DNA was extracted for all strains using a protocol proposed by Schwessinger & McDonald [43] , and over 10 micrograms of DNA was subjected to high

fidelity (HiFi) PacBio sequencing at Discovery Life Science (Huntsville, AL). PacBio sequencing produced an average 24,647,320,191 base pairs (bp) of HiFi reads with read quality of Q33 and median read length of 13,138.

To obtain high-quality Hi-C data, approximately 200 mg of fresh mycelium from each strain was crosslinked and shipped to Phase Genomics (Seattle, WA, USA) to obtain strain-specific Hi-C data. Hi-C data were produced using the Proximo Hi-C Kit (Microbe) at Phase Genomics with DPNII restriction enzymes. Before generating high-throughput Hi-C Illumina data, the quality of these libraries was assessed by mapping low-coverage paired-end data onto available *R. irregularis* assemblies. In all cases, the values obtained for our samples far exceeded the quality thresholds set by Phase Genomics.

RNA extraction and sequencing

RNA was extracted from *Rhizophagus irregularis* strains A4, DAOM-664343; A5, DAOM-664344; G1, DAOM-970895) after cultivation in symbiosis with either *Daucus carota* or *Cichorium intybus* in root organ cultures (ROCs). The SL1 colonization was too poor, and RNA was thus not extracted for this strain. In contrast, strong colonization (>80%) was measured for all other strains and ROC plates from which RNA was extracted (**Supplemental Figure 11**). Total RNA was extracted from colonized roots or extraradical tissue using the RNeasy plant mini kit (Qiagen) following the manufacturer recommendations before quantification with the Nanodrop fluorospectrometer. RNA quality was determined by using electrophoresis on an Agilent Bioanalyzer, with samples showing an average RIN or 8.1. Four RNA biological replicates per condition (material from distinct four ROC plates) were sent for single strand, 100bp paired-end sequencing at Discovery Life Science (Huntsville, AL).

RNA sequencing data were cleaned with fastp 0.22.0 and default settings [44]. Salmon v.1.8.0 [45] was used to align the clean RNA-seq reads to the transcripts and to estimate transcript abundances in each sample (salmon index –keepDuplicates and salmon quant –validateMappings). The entire genome was used as a decoy sequence during the read mappings. We used Tximport and DESeq2 [46] to assess gene differential expression ($P_{adj} < 0.1$).

Chromosome assembly and annotation

The HiFi reads of the strains A4, A5 and G1 were assembled with hifiasm 0.16.1 in Hi-C integration mode and with default parameters [47]. For G1 --hom-cov was set to 181. The resulting contigs were scanned for possible contaminations and mitochondrial sequences by running a Diamond blastx (diamond v0.9.14.115, Database format version = 1) and flagged contigs were removed from the assembly [48]. Phasing of the assembled haplotypes was confirmed with the NuclearPhaser pipeline (MAPQ=10, 100 Kb Hi-C contact map resolution) [22]. For scaffolding, the Hi-C reads were first mapped to each haplotype using BWA-MEM 0.7.17 [49] and further processed according to the Arima Genomics pipeline (https://github.com/ArimaGenomics/mapping_pipeline/blob/master/01_mapping_arima.sh).

Scaffolding was performed using SALSA 2.2 [50]. Hi-C contact maps were produced with HiC-Pro 2.11.1 [51] and visually examined for correct scaffolding.

De novo repeats were predicted with RepeatModeler 2.0.0 and the option -LTRStruct [52] . These were merged with the RepeatMasker repeat library and RepeatMasker 4.1.0 was run with this combined repeat database (http://www.repeatmasker.org).

Gene annotation was achieved by using Funannotate (v.1.7.4) (https://zenodo.org/record/2604804#.YPm336iSnIU) on repeat-masked genome assemblies. Genome annotations were performed using RNA-seq reads and protein models available for strains A4 and A5 (JGI, [29]), and using RNA-seq for strain G1 and RNA-seq reads and protein models available for strains A4 were used for SL1. The quality of annotations was evaluated using the Benchmarking Universal Single-Copy Orthologs software v5.2.2 [53]. Conserved protein domains were predicted using Pfam v.27 [54]. SignalP 4.1 (-t euk -u 0.34 -U 0.34) [55] and Tmhmm 2.0 [56] were used to predict secreted proteins. A protein was called secreted if it was predicted to have a signal peptide but no transmembrane domains. Effector candidates were predicted with EffectorP 3.0 [57].

To assess differences in repeat content, transposable element locations were extracted from the RepeatMasker output file and simple repeats, unknown and low-complexity repeats, satellites, tRNA, sRNA and rRNAs were fremoved. PfamScan [54] was used with default parameters to identify the protein domain annotations in all strains. To create the Pfam heatmaps, the domain numbers were counted for each *R. irregularis* strain or A/B compartment, and *t*-tests were used to highlight the protein domain abundances for each category.

We used dnadiff from the MUMmer suite for structural variation calculations between strains and compartments[58]. D-Genies dot plots were also used to analyze overall synteny between haplotypes and pinpoint structural variation sites between them [59]. Orthology analyses were made by OrthoMCL [60] using the following parameters: 50% identity and 50% coverage using protein sequences of all five assemblies. All karyoplots were produced using KaryoploteR [61].

GO annotations were performed using EggNOG (v5.0.2), with the funog database [62]. Proteins from all genes and 1-to-1 orthologous groups upregulated across conditions were selected and their GO annotations were analyzed. GO classifiers were categorised using WEGO 2.0 [63], and the results visualized using R.

Identification of chromosome compartments and topologically associated domains

We called A/B compartments with Hicexplorer 3.6[64] using the command hicPCA . Inter- and intrachromosomal Hi-C contact maps were produced with HiC-Pro 2.11.1 (MAPQ = 10) and Hicexplorer 3.6 and these Hi-C contact maps were manually inspected to assign chromosomal regions to A/B compartments. Specifically, the

regions along each chromosomes carrying the same PCA1 values – that is, positive or negative – were assigned to the same compartment. Following this assignment, compartments were manually inspected to investigate their proximity with other chromosomes. Those in close physical proximity with the smallest chromosomes were assigned to compartment A, while those carrying the counterpart PCA1 signal were assigned to compartment B. This process was manually repeated for all 32 chromosomes individually to separate the chromosomal regions into separate A/B compartments. Bedtools was used to assess overlap between genomic features such as genes/repeats and compartments [65]. The strength of the PCA1 values for the strains G1 and SL1 allowed the identification of compartment for only a subset of the chromosomes.

Phylogenetic analyses

The haplotype genomes from AMF heterokaryons and ddRAD-seq data were used included to perform a phylogenetic of R. irregularis populations using the Phame pipeline [66]. The ddRAD-seq data from 61 *R. irregularis* isolates [28] (all available replicates included) were quality assessed using FASTQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), and Illumina adapters were trimmed using TagCleaner [67]. Reads containing uncalled bases (N) and shorter reads (<50bp) were removed. The demultiplexing of the dataset was achieved with the *process_radtags* from the Stacks pipeline v2.60 [68].

To construct the phylogenetic tree, the 8 haplotype genome assemblies as well as five chromosome-level homokaryotic assemblies [25] and the ddRAD-seq reads that were processed with PhaME (version 1.0.2) using these parameters: data = 4, reads = 2, code = 0, cutoff = 0.0001, reference = 1, reffile = DAOM-197198 genome assembly [66]. IQTREE algorithm was used with GTR+FO mode and1000 bootstrap replicates on the 105,979 bp alignment file to construct the phylogenetic tree[62]. The phylogenetic tree was then visualized with FigTree (http://tree.bio.ed.ac.uk/software/figtree/). Admixture (v1.3.0) was used for cross-validation of the *K* values and the *K* value with the lowest coefficient of variant was selected (*K* = 11) [69]. Nucleotide alignment file created by Phame was then used by NGSadmix with -K 11 parameter was used to create the admixture plots [70].

Nuclear DNA content estimation using flow cytometry

Mycelium of G1 (DAOM-970895) and SL1 (DAOM-240409) strains was recovered from Phytagel medium used in ROC using a needle after incubating slants of gel on citrate buffer (3 mM sodium citrate and 0.5 mM citric acid) for 1 h under gentle agitation. Mycelium of each strain (alone or along a DNA standard) was chopped with a razor blade in the presence of 1 ml of nuclei extraction buffer following the protocol for filamentous fungi optimized by Pires et al. (2016) . DNA standards used were the fungus *Cenoccocum geophilum* (1C = 0.208 pg DNA)[58] and the plant *Rhamnus alaternus* (2C = 0.680 pg DNA) . The nuclear suspension obtained was then separated from plant debris using a 30 µm nylon filter. After filtration, 50 µg/mL of propidium iodide (PI; Sigma-Aldrich) were added to stain DNA and 50 µg/mL of RNase (Sigma-Aldrich) ware added to prevent staining of double stranded RNA. The samples were maintaining at room temperature

and analyzed using a CyFlow Space flow cytometer (Sysmex, Germany) equipped with a 30 mW green solid-state laser emitting at 532 nm for optimal PI excitation. Flow cytometer quality assessment, data acquisition parameters (including gating strategies to improve the quality of the histogram) and data analysis were performed as described by Pires et al. (2016) . At least five replicates were analyzed for each species. The haploid genome size in pg (1C) was estimated calculated using the fluorescence of the sample and the standard, as described by Talhinhas et al. (2017) . The approximate genome size in Mbp was also calculated for each strain using the factor 1 pg = 978 Mbp .

PCR amplification and digital droplet PCR (ddPCR)

PCR amplification and digital droplet PCR (ddPCR) were done on each sample in order to identify the nuclear ratio of the MAT locis for each strain. The PCR mixture for each sample contained 12 µL of 2X Supermix for Probes (No dUTP) (Bio-Rad), 1.2 µL of each primers/probe mixture (PrimeTime® Std qPCR Assay (500 rxn), 7.2 µL of autoclaved dH_2O and 1 µL of approximately 50 ng/µL of extracted DNA. The following replicates were used for ddPCR: A5/Chicory, 23 replicates; A5/P68, 24 replicates; A4/Chicory, 24 replicates; A4/P68, 24 replicates, G1/Chicory, 22 replicates, G1/P68, 18 replicates.

The MAT-1 (FAM) primers 5'-CATCAACAAGTCAACGATTTAT-3' and 5'-GTGGATACATGACATGGTGT-3' and probe 5'-CAGAAACATTTAATAATAATAATACACGTT-3'; MAT-2 (HEX) primers 5'-TACACAACAAGTCAACGATG-3' and 5'-CATGATGCTCAATATTAAGTG-3' and probe 5'-GTAATGAAATTATAGAAGGAAATATTAG-3'; MAT-3 (HEX) primers 5'-CGTCAAAGAATCACGACACTC-3' and 5'-CATTATTCACAATTGCGTTCGG-3' and probe 5'-CATATAAGAAACAAAAATGCCTTAATTCAAG-3'; MAT-5 (HEX) primers 5'-CAGATTTAGACAAAGATATTCG-3' and 5'-TCCTTATTACATATTTCTACA-3' and probe 5'-CTTGTATATCAACTGTAACATATCAG-3'; MAT-6 (FAM) primers 5'-CCGTCAAAGACTCACAACAC-3' and 5'-CATTATTCACAATTGCGTTCGG-3' and probe 5'-CACATCCGTATATATGTGCAGATAAACAA-3' were used for A4 (MAT-1 and MAT-2), A5 (MAT-3 and MAT-6) and G1(Mat-1 and MAT-5) accordingly [1]. Five to twelve PCR replicates were used depending on availability for each condition. Two positive controls (mycelium extracted DNA) and two no-template controls (containing the two respective MAT primers/probe) were used per strain. Samples were emulsified using the Bio-Rad QX200™ Droplet Generator with 72µL of droplet Generation Oil for Probes per sample. Samples were transferred to a 96-well plate for PCR and the C1000 Touch™ Thermal Cycler (Bio-Rad) was used at the following cycling conditions: 10 minutes at 95˚C, 1 minute at 94˚C, 2 minutes at 52.5˚C, followed by 44 cycles of 1 minute at 94˚C and 2 minutes at 52.5˚C, then 10 minutes at 98˚C and an infinite hold at 4˚C. At every step, a ramp rate of 1˚C/second was used [1]. Following PCR amplification, ddPCR and absolute DNA quantification analysis were performed using the Bio-Rad QX200™ Droplet Reader. The droplet reader identifies each droplet as positive or negative in regard to the fluorescence level of the dye probe, in this case FAM or HEX. Dual-target detection for each sample via

multiplexing was used (detection of the 2 respective MAT locis within the same well) to measure the ratio of the MAT-loci within each sample. The software QuantaSoft Analysis Pro (1.0.596) fits the fraction of positive droplets into a Poisson algorithm to determine the initial absolute concentration of DNA. The positive and no-template controls were used as guides to manually set the threshold using the Threshold Line Mode Tool at the 2D amplitude mode of QuantaSoft Analysis Pro (1.0.596). Wells containing less than 10,000 droplets were excluded from the analysis.

Acknowledgments

We thank Pierre-Marc Delaux, Timothy James, Allyson MacLean and Vasilis Kokkoris for their helpful comments on an earlier version of this manuscript. Our research is funded by the Discovery program of the Natural Sciences and Engineering Research Council (RGPIN2020-05643), a Discovery Accelerator Supplements Program (RGPAS-2020-00033). N.C. is a University of Ottawa Research Chair in Microbial Genomics. J.S. was supported by an Australian Research Council (ARC) Discovery Early Career Researcher Award (DE190100066).

Data accessibility

All genome and RNA-seq data and reads newly obtained are available in GenBank under the BioProject PRJNA922099.

References

1. Spatafora, J. W. *et al.* A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia* **108**, 1028–1046 (2016).
2. Martin, F. M., Uroz, S. & Barker, D. G. Ancestral alliances: Plant mutualistic symbioses with fungi and bacteria. *Science* **356**, eaad4501 (2017).
3. van der Heijden, M. G. A. *et al.* Mycorrhizal fungal diversity determines plant biodiversity , ecosystem variability and productivity. *Nature* **396**, 69–72 (1998).
4. Kokkoris, V., Stefani, F., Dalpé, Y., Dettman, J. & Corradi, N. Nuclear Dynamics in the Arbuscular Mycorrhizal Fungi. *Trends Plant Sci.* **25**, 765–778 (2020).
5. Redecker, D., Kodner, R. & Graham, L. E. Glomalean fungi from the Ordovician. *Science* **289**, 1920–1921 (2000).
6. Rich, M. K. *et al.* Lipid exchanges drove the evolution of mutualism during plant terrestrialization. *Science* **372**, 864–868 (2021).
7. Kaiser, V. B. & Charlesworth, B. The effects of deleterious mutations on evolution in non-recombining genomes. *Trends in Genetics* vol. 25 9–12 Preprint at https://doi.org/10.1016/j.tig.2008.10.009 (2009).

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

9. Halary, S. *et al.* Conserved Meiotic Machinery in Glomus spp., a Putatively Ancient Asexual Fungal Lineage. *Genome Biol. Evol.* **3**, 950–958 (2011).

10. Ropars, J. *et al.* Evidence for the sexual origin of heterokaryosis in arbuscular mycorrhizal fungi. *Nature Microbiology* **1**, (2016).

11. Anderson, J. B. & Kohn, L. M. Dikaryons, diploids, and evolution. *Sex in Fungi: molecular determination and evolutionary implications. ASM Press, Washington, DC* 333–348 (2007).

12. Chen, E. C. H. *et al.* Single nucleus sequencing reveals evidence of inter-nucleus recombination in arbuscular mycorrhizal fungi. *Elife* **7**, 1–17 (2018).

13. Jinks, J. L. Heterokaryosis: a system of adaptation in wild fungi. *Proceedings of the Royal Society of London. Series B - Biological Sciences* **140**, 83–99 (1952).

14. Davis, R. H. ADAPTATION IN PANTOTHENATE-REQUIRING NEUROSPORA II. NUCLEAR COMPETITION DURING ADAPTATION. *Am. J. Bot.* **47**, 648–654 (1960).

15. Serghi, E. U. *et al.* Homo- and Dikaryons of the Arbuscular Mycorrhizal Fungus Rhizophagus irregularis Differ in Life History Strategy. *Front. Plant Sci.* **12**, 715377 (2021).

16. Cornell, C. *et al.* The arbuscular mycorrhizal fungus Rhizophagus irregularis harmonizes nuclear dynamics in the presence of distinct abiotic factors. *Fungal Genet. Biol.* **158**, 103639 (2022).

17. Kokkoris, V. *et al.* Host identity influences nuclear dynamics in arbuscular mycorrhizal fungi. *Curr. Biol.* **31**, 1531-1538.e6 (2021).

18. Gehrmann, T. *et al.* Nucleus-specific expression in the multinuclear mushroom-forming fungus Agaricus bisporus reveals different nuclear regulatory programs. *Proceedings of the National Academy of Sciences* (2018) doi:10.1073/pnas.1721381115.

19. Meunier, C., Darolti, I., Reimegård, J., Mank, J. E. & Johannesson, H. Nuclear-specific gene expression in heterokaryons of the filamentous ascomycete Neurospora tetrasperma. *Proc. Biol. Sci.* **289**, 20220971 (2022).

20. James, T. Y., Stenlid, J., Olson, Å. & Johannesson, H. Evolutionary significance of imbalanced nuclear ratios within heterokaryons of the basidiomycete fungus Heterobasidion parviporum. *Evolution* **62**, 2279–2296 (2008).

21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).

22. Duan, H. *et al.* Physical separation of haplotypes in dikaryons allows benchmarking of phasing accuracy in Nanopore and HiFi assemblies with Hi-C data. *Genome Biol.* **23**, 84 (2022).

23. Sperschneider, J. *et al.* Extensive somatic nuclear exchanges shape global populations of the wheat leaf rust pathogen*Puccinia triticina*. *bioRxiv* (2022) doi:10.1101/2022.11.28.518271.

24. Liang, J. *et al.* Haplotype-phased and chromosome-level genome assembly of Puccinia polysora, a giga-scale fungal pathogen causing southern corn rust. *Mol. Ecol. Resour.* (2022) doi:10.1111/1755-0998.13739.

25. Yildirir, G., Sperschneider, J. & Malar C, M. Long reads and Hi-C sequencing illuminate the two-compartment genome of the model arbuscular mycorrhizal symbiont Rhizophagus irregularis. *New* (2022).

26. Manley, B. F. *et al.* A highly contiguous genome assembly reveals sources of genomic novelty in the symbiotic fungus Rhizophagus irregularis. *bioRxiv* 2022.10.19.511543 (2022) doi:10.1101/2022.10.19.511543.

27. Teulet, A. *et al.* A pathogen effector FOLD diversified in symbiotic fungi. *bioRxiv* (2022) doi:10.1101/2022.12.16.520752.

28. Savary, R. *et al.* A population genomics approach shows widespread geographical distribution of cryptic genomic forms of the symbiotic fungus Rhizophagus irregularis. *ISME J.* 1–14 (2017).

29. Chen, E. C. H. *et al.* High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont Rhizophagus irregularis. *New Phytol.* **220**, 1161–1171 (2018).

30. Corradi, N. & Brachmann, A. Fungal Mating in the Most Widespread Plant Symbionts? *Trends Plant Sci.* (2016).

31. Gandolfi, A., Sanders, I. R., Rossi, V. & Menozzi, P. Evidence of recombination in putative ancient asexuals. *Mol. Biol. Evol.* **20**, 754–761 (2003).

32. Judson, O. P. & Normark, B. B. Ancient asexual scandals. *Trends in Ecology and Evolution* vol. 11 41–46 Preprint at https://doi.org/10.1016/0169-5347(96)81040-8 (1996).

33. Hiltunen, M., Grudzinska-Sterno, M., Wallerman, O., Ryberg, M. & Johannesson, H. Maintenance of high genome integrity over vegetative growth in the fairy-ring mushroom Marasmius oreades. *Curr. Biol.* **29**, 2758-2765.e6 (2019).

34. Li, F. *et al.* Emergence of the Ug99 lineage of the wheat stem rust pathogen through somatic hybridisation. *Nat. Commun.* (2019) doi:10.1038/s41467-019-12927-7.

35. Stukenbrock, E. H. Evolution, selection and isolation: a genomic view of speciation in fungal plant pathogens. *New Phytol.* **199**, 895–907 (2013).

36. Croll, D. & McDonald, B. A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathog.* **8**, (2012).

37. Plissonneau, C., Hartmann, F. E. & Croll, D. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biol.* **16**, (2018).

38. Ehinger, M., Koch, A. M. & Sanders, I. R. Changes in arbuscular mycorrhizal fungal phenotypes and genotypes in response to plant species identity and phosphorus concentration. *New Phytol.* **184**, 412–423 (2009).

39. Strassert, J. F. H. & Monaghan, M. T. Phylogenomic insights into the early diversification of fungi. *Curr. Biol.* **32**, 3628-3635.e3 (2022).

40. Galindo, L. J., López-García, P., Torruella, G., Karpov, S. & Moreira, D. Phylogenomics of a new fungal phylum reveals multiple waves of reductive evolution across Holomycota. *Nat. Commun.* **12**, 4973 (2021).

41. Corradi, N., Hijri, M., Fumagalli, L. & Sanders, I. R. Arbuscular mycorrhizal fungi (Glomeromycota) harbour ancient fungal tubulin genes that resemble those of the chytrids (Chytridiomycota). *Fungal Genet. Biol.* **41**, 1037–1045 (2004).

42. Jansa, J., Mozafar, A., Banke, S., McDonald, B. A. & Frossard, E. Intra- and intersporal diversity of ITS rDNA sequences in Glomus intraradices assessed by cloning and sequencing, and by SSCP analysis. *Mycol. Res.* **106**, 670–681 (2002).

43. Schwessinger and Megan McDonald, B. High quality DNA from Fungi for long read sequencing e.g. PacBio, Nanopore MinION v3. (2017) doi:10.17504/protocols.io.jxxcppn.

44. Chen, S., Zhou, Y., Chen, Y. & Gu, J. fastp: an ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).

45. Patro, R., Duggal, G., Love, M. I., Irizarry, R. A. & Kingsford, C. Salmon provides fast and bias-aware quantification of transcript expression. *Nat. Methods* **14**, 417–419 (2017).

46. Love, M. I., Huber, W. & Anders, S. *DESeq2*. (2014).

47. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* **40**, 1332–1335 (2022).

48. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nature Methods* Preprint at https://doi.org/10.1038/nmeth.3176 (2014).

49. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

50. Ghurye, J., Pop, M., Koren, S., Bickhart, D. & Chin, C.-S. Scaffolding of long read assemblies using long range contact information. *BMC Genomics* **18**, (2017).

51. Servant, N. *et al.* HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.* **16**, 259 (2015).

52. Flynn, J. M. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci. U. S. A.* **117**, 9451–9457 (2020).

53. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* btv351 (2015).

54. Finn, R. D. *et al.* The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res.* **44**, D279–D285 (2016).

55. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat. Methods* **8**, 785–786 (2011).

56. Chen, Y., Yu, P., Luo, J. & Jiang, Y. Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. *Mamm. Genome* **14**, 859–865 (2003).

57. Sperschneider, J. & Dodds, P. N. EffectorP 3.0: Prediction of apoplastic and cytoplasmic effectors in fungi and Oomycetes. *Mol. Plant. Microbe. Interact.* **35**, 146–156 (2022).

58. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* (2018) doi:10.1371/journal.pcbi.1005944.

59. Cabanettes, F. & Klopp, C. D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958 (2018).

60. Li, L., Stoeckert, C. J. & Roos, D. S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res.* **13**, 2178–2189 (2003).

61. Gel, B. & Serra, E. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics* **33**, 3088–3090 (2017).

62. Huerta-Cepas, J. *et al.* EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* (2019) doi:10.1093/nar/gky1085.

63. Ye, J. *et al.* WEGO: a web tool for plotting GO annotations. *Nucleic Acids Res.* **34**, W293–W297 (2006).

64. Wolff, J. *et al.* Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Res.* **48**, W177–W184 (2020).

65. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).

66. Shakya, M. *et al.* Standardized phylogenetic and molecular evolutionary analysis applied to species across the microbial tree of life. *Sci. Rep.* **10**, 1723 (2020).

67. Schmieder, R., Lim, Y. W., Rohwer, F. & Edwards, R. TagCleaner: Identification and removal of tag sequences from genomic and metagenomic datasets. *BMC Bioinformatics* **11**, 341 (2010).

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

69. Alexander, D. H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19**, 1655–1664 (2009).

70. Skotte, L., Korneliussen, T. S. & Albrechtsen, A. Estimating individual admixture proportions from next generation sequencing data. *Genetics* **195**, 693–702 (2013).