# From single nuclei to whole genome assemblies

## Author information

Merce Montoliu-Nerin, Department of Ecology and Genetics, Evolutionary Biology, Uppsala University, Sweden

Marisol Sánchez-García, Department of Ecology and Genetics, Evolutionary Biology, Uppsala University, Sweden

Claudia Bergin, Department of Cell and Molecular biology, Uppsala University, and Microbial Single Cell Genomics Facility, Science for Life Laboratory, Sweden

Manfred Grabherr, Department of Medical Biochemistry and Microbiology, Uppsala University, Sweden

Barbara Ellis, Department of Ecology and Genetics, Evolutionary Biology, Uppsala University, Sweden

Verena Esther Kutschera, Science for Life Laboratory Solna and Department of Biochemistry and Biophysics, Stockholm university, Sweden

Marcin Kierczak, Department of Cell and Molecular biology, Uppsala University, Sweden

Hanna Johannesson, Department of Organismal Biology, Systematics biology, Uppsala University, Sweden

*Anna Rosling, Department of Ecology and Genetics, Evolutionary Biology, Uppsala University, Sweden

## Summary

A large proportion of Earth's biodiversity constitutes organisms that cannot be cultured, have cryptic life-cycles and/or live submerged within their substrates[1–4]. Genomic data are key to unravel both their identity and function[5]. The development

of metagenomic methods[6,7] and the advent of single cell sequencing[8–10] have revolutionized the study of life and function of cryptic organisms by upending the need for large and pure biological material, and allowing generation of genomic data from complex or limited environmental samples. Genome assemblies from metagenomic data have so far been restricted to organisms with small genomes, such as bacteria[11], archaea[12] and certain eukaryotes[13]. On the other hand, single cell technologies have allowed the targeting of unicellular organisms, attaining a better resolution than metagenomics[8,9,14–16], moreover, it has allowed the genomic study of cells from complex organisms one cell at a time[17,18]. However, single cell genomics are not easily applied to multicellular organisms formed by consortia of diverse taxa, and the generation of specific workflows for sequencing and data analysis is needed to expand genomic research to the entire tree of life, including sponges[19], lichens[3,20], intracellular parasites[21,22], and plant endophytes[23,24]. Among the most important plant endophytes are the obligate mutualistic symbionts, arbuscular mycorrhizal (AM) fungi, that pose an additional challenge with their multinucleate coenocytic mycelia[25]. Here, the development of a novel single nuclei sequencing and assembly workflow is reported. This workflow allows, for the first time, the generation of reference genome assemblies from large scale, unbiased sorted, and sequenced AM fungal nuclei circumventing tedious, and often impossible, culturing efforts. This method opens infinite possibilities for studies of evolution and adaptation in these important plant symbionts and demonstrates that reference genomes can be generated from complex non-model organisms by isolating only a handful of their nuclei.

## Main text

AM fungi is a group of diverse obligate symbionts that have colonized root cells and formed mycelial networks in soil since plants first colonized land[25–27]. Their entire life-cycle is completed underground and they propagate with multinuclear asexual spores[27] (Figure 1). Genomic research on AM fungi has been hampered by technical challenges involving isolation and culturing, and accordingly, few species have been successfully sequenced. To date, the reference genomes of only few species that can be grown in axenic culture, *i.e., Rhizophagus irregularis*, *R. clarus*, *R. diaphanus*, *R. cerebriforme*, *Gigaspora rosea* and *Diversispora epigaea,* have been published[28–33].

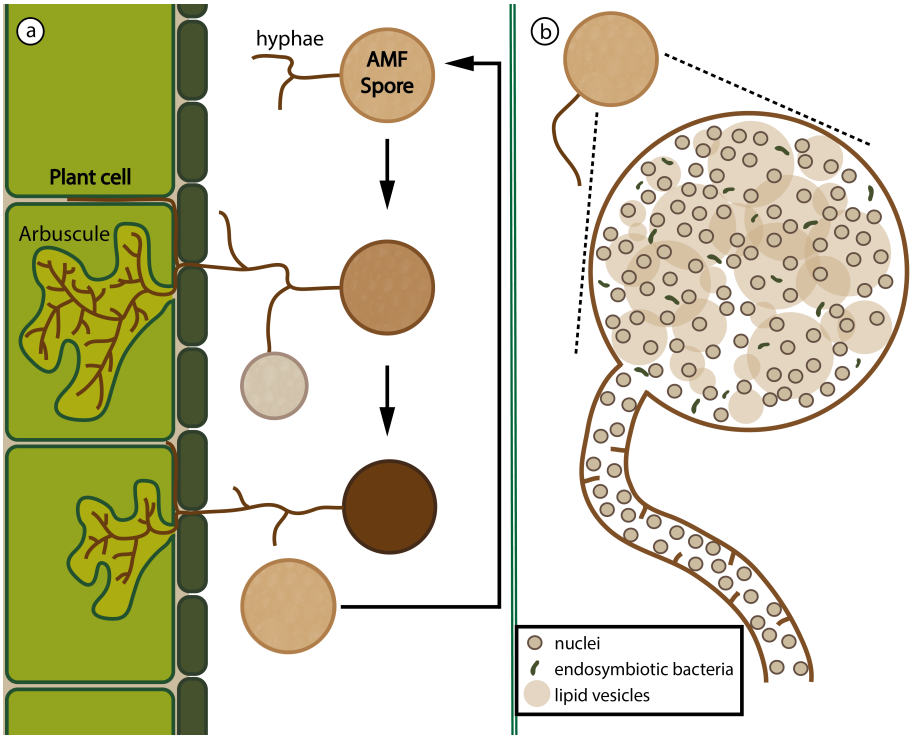


Figure 1. a) Schematic representation of the life-cycle in AM fungi. A spore detects a plant root in the vicinity and grows hyphae towards it. The hyphae penetrate the plant cell wall and form the characteristically branching haustoria with the shape of arbuscules, based on which the group of fungi is named. The arbuscules are used to exchange nutrients with the plant. New spores are produced in other hyphal

terminations, bud off upon maturity and remain in dormant state until the cycle starts again, while the first spore dies and the fungi retracts from the plant cell. b) Schematic representation of a spore, containing nuclei, lipid vesicles and endosymbiotic bacteria. The hyphae have very reduced compartmentalization with incomplete septa and nuclei appear to move freely.

A method was developed in which genomic fungal DNA can be obtained, free of plant and microbial DNA, directly from individual nuclei of multinucleate spores. In brief, spores from a trap culture fungal strain of *Claroideoglomus clarodeium/C. luteum* (SA101) were obtained from the INVAM pot culture collection. An initial trial to sort AM nuclei was carried out using pools of spores in order to assess the optimal settings. Cleaned spores were crushed vigorously, and the solution was stained and analyzed by Fluorescence-Activated Cell Sorting (FACS), recording level of fluorescence as a measure of DNA content, and light scattering as proxy for size and particle granularity (Figure 2 a-h). A distinct cloud of particles was observed above the background in the scatter plot (Figure 2h, inside the blue box) which, after microscopy and PCR verification tests with fungal and bacterial specific primers, was confirmed to consist of intact biological structures containing mostly fungal DNA (Figure S1-S3, Table S1). Hence, we concluded that these particles were fungal nuclei and restricted future sorting to this window. Thereafter, individual nuclei from a single spore of the same strain were sorted into wells of a 96-well plate (Figure S4, Table S2) and whole genome amplified (WGA) using multiple displacement amplification (MDA; Figure 2 i-j). The amplified DNA was scored for pure fungal origin by parallel amplification of rDNA barcode regions for both fungi and bacteria (Figure 2k, Figure S5). Twenty-four amplified nuclei samples, confirmed to contain only fungi (Figure S4, Table S3, S4), were sequenced with Illumina HiSeq X (Figure

2l). Further, the MinION Nanopore-based sequencing device (Oxford Nanopore Technologies, ONT, UK) was used to obtain long read sequences for amplified DNA from multiple (5-100) nuclei separated from a pool of 30 spores of the same strain (Figure 2 i-k, m).

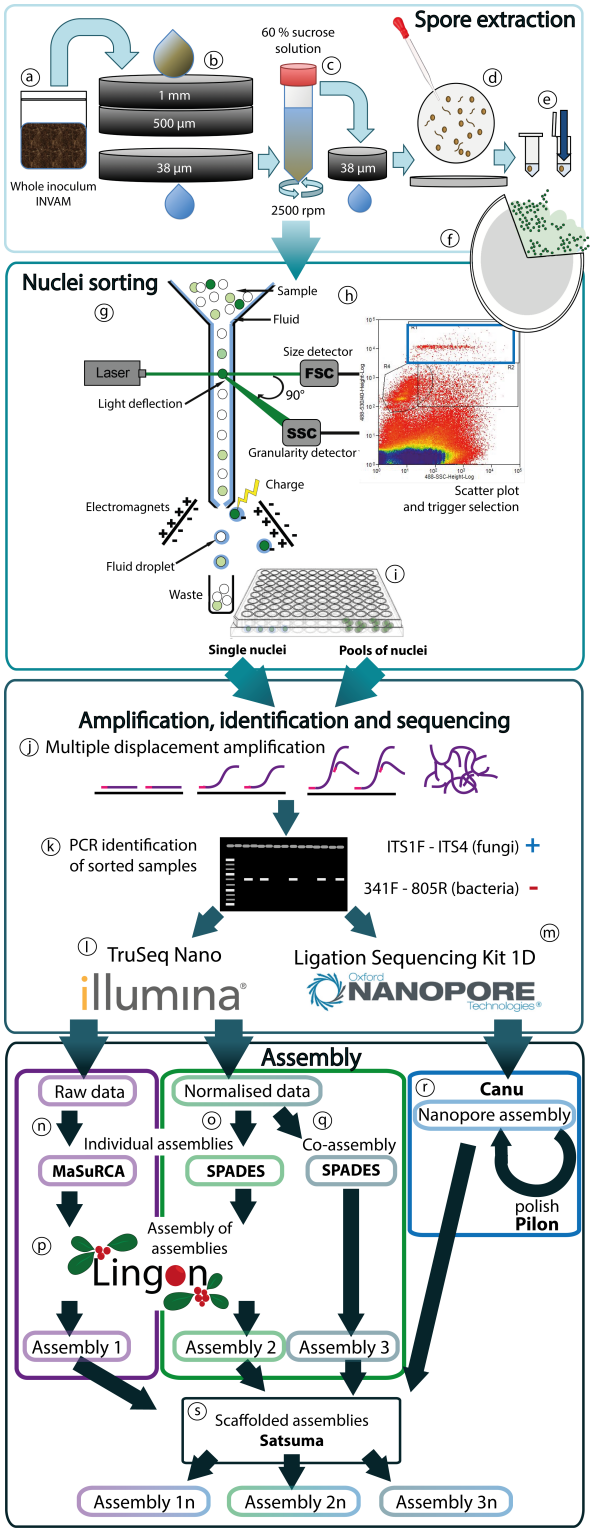

Figure 2. From a soil sample to AM fungal genome assemblies. a) Whole inoculum from the culture collection INVAM is blended with water and (b) poured into a set of sieves, the material stuck in the 38 µm sieve is placed into a (c) tube that contains a solution of 60% sucrose, then centrifuged for 1 min. The supernatant is again run through a 38 µm sieve and washed with water. d) The sieve content is placed in a Petri dish for the spores to be manually picked using a glass pipette. e) After cleaning the spores with dd$H_2O$, these are placed one-by-one into tubes and crushed with a pestle. f) The DNA from a broken spore is stained with SYBR Green, giving a strong fluorescent signal for the nuclei, and lighter for the background, organelles and microbes. g) The stained spore content is loaded on the FACS, in which the sample moves inside a constant flow of buffer and crosses a laser beam. An excitation laser of 488-nm and 530/40 band pass filter was used for the SYBR Green fluorescence detection. In addition scattered light, forward scatter (FSC) and side scatter (SSC) were used as proxy for size and granularity to identify the nuclei. h) The signals can be interpreted in a scatter-plot, and particles of a selected cloud (e.g., R1, blue-box) can be sorted individually or pooled (i) into individual wells of a 96-well plate by directing them with a charge. j) The content of each well is whole genome amplified using MDA. k) The amplified products are tested for fungi and bacteria by PCR screening with specific rDNA primers for fungi and bacteria. The products confirmed to be from fungal nuclei are sequenced with l) Illumina HiSeqX, for single nuclei; and m) Oxford Nanopore, for pools of nuclei. To produce assembly 1, Illumina reads are assembled separately for individual nuclei using MaSuRCA[34] (n). To produce assembly 2, reads are normalized for individual nuclei and assembled with SPADES[35] (o). For assembly 3 reads from all nuclei are combined before normalization and then assembled with SPADES[35] (q). Individual nuclei assemblies from method 1 and 2 are

assembled together using Lingon[36] (p). Nanopore data is assembled with Canu[37] (r), polished with Pilon[38] using the Illumina raw-reads and used to scaffold the three generated assemblies using Chromosemble, of Satsuma[39] (s).

Three customized assembly workflows were developed in order to evaluate assembly quality in the light of coverage bias introduced by WGA, which is the biggest challenge when assembling sequence data from amplified single nuclei. The MDA method, however, has an advantage over PCR-based methods in that it produces longer fragments of DNA with a lower error rate, and that the coverage bias is random[40,41].

For the first two assembly workflows individual nuclei assemblies were generated and subsequently combined to generate a consensus assembly using the workflow manager Lingon[36] (Figure 2p), which consists of a motif-distance based long sequence overlap finder that merge sequences based on mutual maximal overlaps. In the first assembly workflow, raw Illumina reads were assembled using MaSuRCA[34] (Figure 2n) resulting in 24 assemblies, ranging in size from 14 to 69 Mbp (Tables S5). In order to overcome MDA generated differences in coverage across the genome the second assembly workflow normalized raw reads to maximum 100X before assembly using SPADES[35] (Figure 2o), generating 24 assemblies ranging in size from 11 to 50 Mbp (Table S5). A third assembly was created using SPADES[35] after combining raw reads from 24 nuclei followed by normalization to 100X (Figure 2q). One full assembly with 24 nuclei was generated from each workflow and subsequently scaffolded with a Nanopore assembly built with Canu[37] (Figure 2r-s). To test for the effect of increasing number of assembled nuclei in the three methods, random combinations with different number of nuclei were assembled with the three assembly

146 workflows. Multiple replicate assemblies were performed for different random
147 combinations of two to twelve nuclei and one random combination for 13-23 nuclei.
148 BUSCO[42], assembly size and N50 was used to compare these to full and single nuclei
149 assemblies.
150
151 **Results**
152 The different assembly workflows resulted in assemblies that vary in sizes,
153 fragmentation and completeness (Table 1). Based on BUSCO analyses, workflow 3
154 generates the most complete assembly, with 89% for assembly 3n, compared to 2n at
155 80%, and 1n at 78% (Table 1). Of the core single copy genes identified by BUSCO,
156 few were fragmented or duplicated in assembly 3n indicating that the set of 14,600
157 predicted genes is likely to be complete and a close representation of the genetic
158 diversity in this strain (Table 1). This number is lower than the number of genes
159 found in other sequenced AM fungi such as *R. irregularis*[28] and *R. clarus*[31], also
160 lower than those predicted in assemblies 1n and 2n (Table 1). Interestingly, assembly
161 3n is considerably smaller (70.8 Mb) than the other full assemblies (92.4 Mb and
162 130.4 Mb for assembly 1n and 2n respectively), and markedly smaller than the
163 average estimated genome size of 119 Mb based on SGA-PreQC[43]. The smaller size
164 of 3n can be attributed to repeat sequences (20.6 Mb) that are captured, to a lesser
165 extent, compared to the other assembly workflows (41.3 - 58.6 Mb). Specifically,
166 normalization is expected to disproportionally reduce high coverage genomic
167 sequences such as repeat elements, and collapsing those regions when assembling.
168 Note that this effect of normalization is eluded in assembly workflow 2, in which
169 nuclei are normalized and assembled individually; repetitive regions will collapse but
170 in different parts of the genome, ending up represented in the final assembly when

combined. In contrast, workflow 1 is based on non-normalized reads. Due to uneven coverage, this workflow assembles less of the genome (an average of 55% of the raw reads align to the individual nuclei assemblies, as opposed to 96% of the reads mapping to the normalized individual nuclei assemblies (Table S5)) but generate contigs well supported by high coverage. Combining these incomplete assemblies from single nuclei using Lingon, generates an accurate assembly 1 comparable to assembly 3 with a better representation of repeats (Table 1).

Combinations of increasing number (1-24) of random nuclei were produced for all the assembly workflows in order to evaluate the number of nuclei needed to produce a good final assembly. As shown in figure 3, single nuclei assemblies are most complete when using normalized workflows (2 and 3), with an average of 40% BUSCO estimated completeness. Interestingly, there is an increasing number of gene duplications among the complete genes as more single nuclei assemblies are combined for method 2 compared to method 1 (Figure 3a-b). Higher amount of gene duplications was confirmed by locating known single copy genes in all assemblies (Table S6). The duplications in workflow 2 are likely generated because read normalization allows for assembly of regions with low coverage that are prone to errors and prevents contigs from being properly assembled by the workflow manager Lingon.

## Discussion

Methodological challenges in assembling genomes from amplified single nuclei or cells can be elevated by careful analysis of generated assemblies. Combining and normalizing reads (workflow 3) from only 6 individually sequenced nuclei can already generate a high coverage genome assembly. From this assembly, good quality

196 data of single copy genes are obtained, ideally suited for phylogenomic studies.
197 Assembly workflow 1 on the other hand is better suited to characterize repeat
198 elements in the genome since these are better represented in assemblies of non-
199 normalized data. With this method a high-quality genome can be assembled using
200 seven individually amplified and sequenced nuclei (Figure 3). Comparative genetic
201 analysis of single nuclei is best done using assemblies from workflow 2. However,
202 single nuclei assemblies based on normalized reads should not be assembled into
203 consensus assemblies since variable quality of contigs make them prone to
204 duplication.
205 To conclude, sequence data from single cell sequencing presents itself as challenging,
206 but as shown here, with the right combination of methods adapted to the data, *de novo*
207 genome references can be generated, opening the door for an expansion in genomic
208 and phylogenomic research in organisms like AM fungi, that have, for too long,
209 evaded large scale genome sequencing efforts due too methodological limitations
210 stemming from their complicated biology. Useful genomic information can be
211 generated from a handful of single nuclei greatly improving our ability to study
212 multicellular eukaryotes with complex life stages. The assembly method of choice
213 will ultimately depend on the research questions asked and the kind of data needed or
214 available.
215
216 Table 1. Comparative assessment of the 3 assembly workflows.

| Assembly | | Size (Mb) | # Contigs | N50 | Largest contig (Kb) | GC (%) | BUSCO (%)[a] | # Genes (Mb) | Repeats (Mb) |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Raw reads | 90.16 | 11077 | 12714 | 94.39 | 27.01 | C: 77<br>F: 10 | 18068 (49.42) | 40.39 |
| 1n | + Nanopore | 92.38 | 3899 | 37258 | 176.652 | 27.91 | C: 78<br>F: 9 | 16680 (69.54) | 41.32 |
| 2 | Normalized to 100x | 124.96 | 21934 | 16055 | 155.09 | 28.07 | C: 79<br>F: 8 | 24930 (69.79) | 57.77 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 2n | + Nanopore | 130.41 | 4632 | 60974 | 338.42 | 28.07 | C: 80 F: 7 | 22618 (105.48) | 58.57 |
| 3 | Combined, normalized to 100x | 68.31 | 11259 | 15947 | 199.90 | 28.08 | C: 88 F: 4 | 15882 (43.73) | 21.71 |
| 3n | + Nanopore | 70.81 | 3883 | 33135 | 220.22 | 28.08 | C: 89 F: 3 | 14662 (55.44) | 20.64 |

[a]BUSCO results in % of complete genes (C) or fragmented (F).

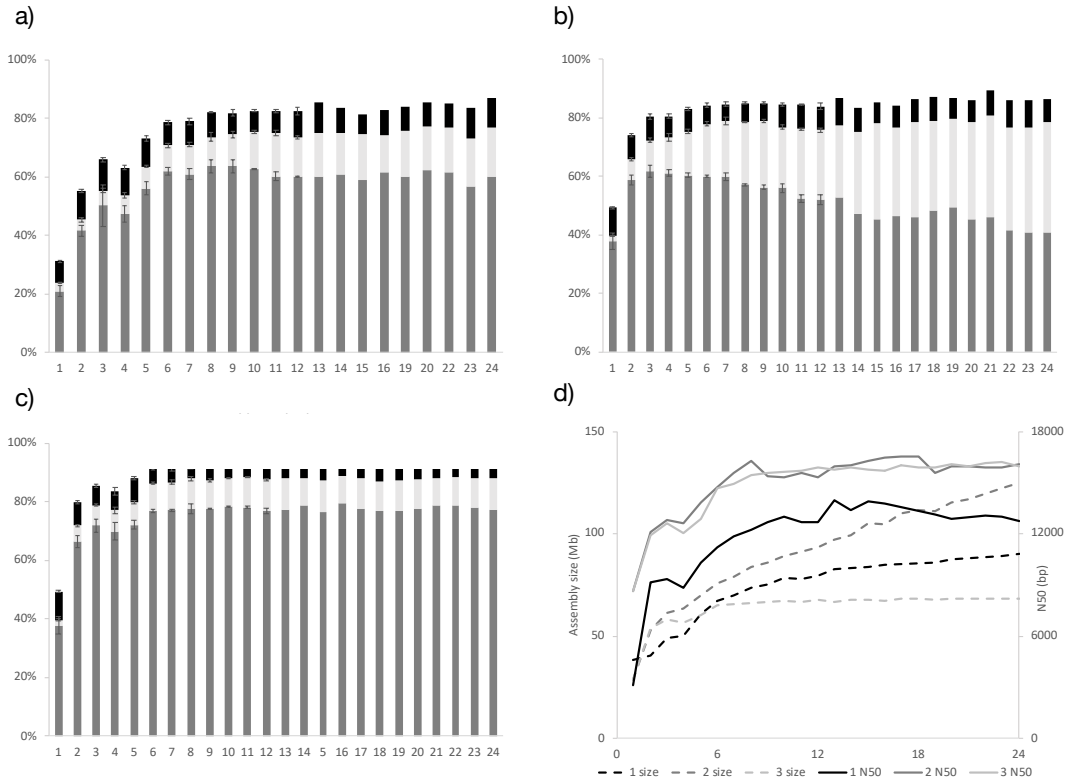


Figure 3

Summary statistics for different number of assembled nuclei (1-24) using three different assembly workflows. BUSCO estimates of completeness for a) workflow 1: raw reads of individual nuclei assembled using Masurca, consensus assembly using Lingon b) workflow 2: normalised reads of individual nuclei assembled using SPADES, consensus assembly using Lingon and c) workflow 3: reads from individual nuclei are pooled and normalised before assembling with SPADES. Percentage of single copy core genes detected as single copy (S: grey), duplicated (D: light grey) or fragmented (F: black). Average of 3-6 replicate assemblies up to 12 nuclei with error

bars indicating SEM. In d) assembly size (dashed lines) and N50 (solid lines) for the there methods 1 (black), 2 (grey) and 3 (light grey).

## Methods

### Fungal strain and spore extraction

*C. claroideium/C. luteum* (SA101) was obtained as whole inoculum from the International culture collection of (vesicular) arbuscular mycorrhizal fungi (INVAM) at West Virginia University, Morgantown, WV, USA. Soil (10-30 ml) was blended with 3 to 4 pulses using a blender half-filled with water (500 ml). The mix was filtered through a set of sieves (1 mm/500 µm/38 µm x 200 mm diameter, (VWR, Sweden). The content of the last sieve was transferred into a falcon tube containing 20 ml of 60% sucrose solution, and centrifuged for 1 minute at 2500-3000 rpm. The supernatant was poured into a small sieve (50 mm diameter) of 38 µm and the sucrose was washed with water. The contents were poured onto a petri dish for better visualization under the stereomicroscope. Spores were transferred individually or in groups to an Eppendorf tube using modified glass pipettes with reduced tip diameter and subsequently cleaned by adding and removing ddH2O five times. The step-by-step protocol can be found in the OSF Repository for the project[44].

### Nuclei extraction and sorting

After spore extraction from soil, individual spores were placed in 30 µl ddH2O in 1.5 ml Eppendorf tubes. One tube with 15 spores was used to establish the sorting window. An amount of 50 µl 1x PBS was added to each tube before crushing the spores using a sterile pestle. DNA was stained by adding 1 µl of 200x SYBR Green I Nucleic Acid stain (Invitrogen™, Thermo Fisher Scientific, MA, USA) and the sample was incubated for 20-50 min in the dark. More 1x PBS was added to increase

the volume to 100-200 µl before putting the sample on the FACS. The sorting was performed with a MoFlo™ Astrios EQ sorter (Beckman Coulter, USA) using a 488 nm laser for excitation, 70 µm nozzle, sheath pressure of 60 psi, and 0.1 µm filtered 1x PBS as sheath fluid. The sorter was triggered on forward scatter (FSC) at a threshold of 0.03% and sort regions were set on SYBR Green I fluorescence (488-530/40) over side scatter (SSC). The samples were sorted in enrich mode with a drop envelope of 1 at 700 to 1200 events per second. Thus, if a particle fitting within the sorting window passes by the laser together with another particle, these would be discarded. Particles from region R1, assumed to be nuclei (Figure S4), were sorted individually into 96 well plates containing 1 µl 1x PBS/well, groups of 5 particles were collected for positive control, and empty wells were kept as negative control (Table S2).

**Whole Genome Amplification**

Sorted nuclei were lysed and neutralized followed by whole genome amplification using Phi29 and MDA as described by Rinke et al., 2014[45]. In short, the cells were incubated in an alkaline solution (buffer DLB and DTT, Qiagen, Germany) for 5 min at room temperature, followed by 10 min on ice. Lysis reactions were neutralized by adding 1 µL neutralization buffer (stop solution, Qiagen, Germany). Both, the alkaline lysis solution as well as the neutralization buffer were UV treated with 2 Joule in a Biolinker. MDA was performed using the RepliPHI™Phi29 Reagent set (RH031110, Epicenter, WI USA) at 30°C for 16 h in 15 µl reaction volumes with a final concentration of 1x reaction buffer, 0.4 mM dNTPs, 10 mM DTT, 5% DMSO, 50 µM hexamers with 3'- phosphorothioate modifications (IDT Integrated DNA Technologies, Iowa USA), 40 U Phi 29 enzyme; 0.5 µM SYTO13® (InvitrogenTM, Thermo Fisher Scientific, MA, USA) and water. All reagents except SYTO13 were

UV decontaminated with 3 Joule in a UV crosslinker as described in Rinke et al., 2014[45] 12 µl of MDA mix were then added to each well.

The whole genome amplification was monitored in real time by detection of SYTO13 fluorescence every 15 minutes for 16 h using a Chromo4 real-time PCR instrument (Bio-Rad, USA) or a FLUOstar®Omega plate reader (BMG Labtech, Germany). The single amplified genome DNA was stored at -20°C for short-term, and transferred to -80°C for long-term storage.

**Selecting single amplified nuclei for sequencing**

MDA products were diluted to approximately 5 ng/µl (40x) and screened for the presence of fungal and bacterial ribosomal genes using PCR. Reaction mixtures were made as described above, using the fungal-specific primers ITS9[46] and ITS4. The PCR protocol had an initial denaturing step of 10 min at 95°C, followed by 35 cycles of 30 s at 95°C, 30 s at 58°C, and 50 s at 72°C for the fungi PCR. For the bacteria-specific 341F/805R[47] primer pairs a different reaction mixture was used containing 10x Standard Taq Reaction buffer (Qiagen), 2 mM MgCl2, 0.2 mM deoxynucleoside triphosphates (dNTPs), a 0.2 µM concentration of each primer, and 1 U Taq DNA polymerase (Qiagen). A positive control of DNA extracted from commercially available *Agaricus bisporus* provided by Dr. Ylva Strid, UU, was included, and ddH2O as negative control. The bacterial PCR protocol consisted on an initial step of 5 min at 95°C, followed by 30 cycles of 30 s at 95°C, 30 s at 58°C, and 50 s at 72°C for the bacteria PCR before a final elongation step of 7 min at 72°C. Bacteria PCR included a positive control of DNA extracted from Legionella provided by Tiscar Graells, Universitat Autónoma de Barcelona (UAB), Spain, and ddH2O as negative control. The reaction was performed with a 2720 Thermocycler of Applied Biosystems (USA). The presence of amplification products was verified by gel

303 electrophoresis by separation on a 2% agarose gel run for 35 min at 110V (fungi) and
304 70V (bacteria) including a Thermo Scientific GeneRuler 100 bp DNA Ladder. (Figure
305 S5), and the samples were identified as fungi positive, bacteria positive, fungi +
306 bacteria positive or failed/empty (Table S3). From the samples that scored positive for
307 presence of fungi, 24 undiluted samples were selected for sequencing and the DNA
308 amount was measured using Qubit (Brand, country) after addition of 30 µl ddH2O
309 (Table S4).

310 **Sequencing of single amplified nuclei**

311 From the 24 selected samples, around 800 ng of DNA was transferred to sequencing
312 plates. Library preparation and sequencing was performed by the SNP&SEQ
313 Technology Platform in Uppsala at the National Genomics Infrastructure (NGI)
314 Sweden and Science for Life Laboratory. For each sample, an individual library was
315 prepared using the TruSeq Nano DNA Library Prep Kit. The sequencing was
316 performed by doing a cluster generation and 150 cycles paired-end sequencing of the
317 24 libraries in 1 lane using the HiSeq X system and v2.5 sequencing chemistry
318 (Illumina Inc., USA). Read data were delivered to us as fastq.

319 **Spore sorting for Nanopore sequencing**

320 Spores were picked in groups of 30 with the help of a P10 and P100 pipette, then
321 washed 5x in nuclease-free water and transferred to Eppendorf tubes in 30 uL
322 nuclease-free water. For the FACS sorting, spores were crushed, then 30 µl 1x PBS
323 was added to the tube along with 1 µl of 200x SYBR Green for staining the DNA (20-
324 50 mins). Sample volume was increased to 200 µl with 1x PBS before loading on the
325 FACS. Pools of 5 and 100 nuclei were sorted into either individual 1.5 ml Eppendorf
326 tubes or into multi-well plates. The above-described WGA protocol was run, and the
327 presence of fungal DNA in the samples was verified by PCR on diluted samples of

amplified pooled nuclei before selecting fungi positive samples for library preparation. PCR reaction mixtures contained 10x Standard Taq Reaction buffer (Qiagen), 2 mM MgCl2, 0.2 mM deoxynucleoside triphosphates (dNTPs), a 0.2 µM concentration of each primer, and 1 U Taq DNA polymerase (Qiagen). The fungal-specific ITS1F/ITS4 and bacteria-specific 341F/805R primer pairs were used for each sample in two independent PCR reactions. The PCR protocol included an initial denaturing step of 5 min at 95°C, followed by either 35 cycles of 30 s at 95°C, 30 s at 55°C, and 50 s at 72°C for the fungi PCR or by 30 cycles of 30 s at 95°C, 30 s at 58°C, and 50 s at 72°C for the bacteria PCR before a final elongation step of 7 min at 72°C. The reaction was performed with a 2720 Thermocycler of Applied Biosystems (USA). Amplification products were visualized and documented by gel electrophoresis as described above.

Libraries were prepared by following the "Premium Whole Genome Amplification" protocol (version WAL_9030_v108_revJ_26Jan2017, Oxford Nanopore Technologies [ONT], Oxford, United Kingdom) in combination with the Ligation Sequencing Kit 1D (SQK-LSK108, ONT) with the following modifications: (a) an alternative WGA method was used (Qiagen Single Cell Kit instead of the Midi Kit); (b) samples were diluted to a 50 µl volume following WGA and quantified with a Qubit fluorometer (brand, country). Amounts of 1 - 2.5 µg DNA were then used for preparing individual libraries, starting with the first bead cleaning step explained in the whole genome amplification section. At the end of this step, samples were eluted in 19 µl nuclease-free water instead of 100 µl. 1 µl of the eluted sample was used for DNA quantification (Qubit fluorometer) while another 1 µl was used to measure DNA quality with Nanodrop (ND 2000); (c) no size selection and intentional shearing was performed to achieve read length as long as possible; (d) 17 µl amplified DNA

was added to the T7 endonuclease treatment; (e) an extended end-prep reaction was performed by incubating the samples for 30-30 mins at both 20°C and 65°C; (f) adapter ligation was allowed for 25-30 mins instead of 10; (g) elution buffer in the final step was incubated for 15 minutes instead of 10; (h) the loaded library contained no additional water but 14.5 µl DNA library instead of 12 µl. Additionally, flicking was used to mix reactions instead of pipetting to prevent DNA fragmentation. Further, eluates were removed and retained in a stepwise fashion (i.e. in multiple aliquots) after every cleaning step to assure that no beads were brought forward with the DNA into the next library preparation step. In general, by extending clean-up-, ligation- and elution steps the quality of the library and thus pore occupancy during sequencing could be improved.

A total of 3 libraries on 3 separate ONT MinION R9.4 flow cells (FLO-MIN106) were sequenced using live base-calling and the standard 48 h sequencing protocol (NC_48Hr_sequencing_FLO-MIN106_LSK-108_plus_Basecaller). One library was run on a fresh flow cell with ~1400 single pores available for sequencing in the beginning of the run. This 48 h run provided 1,686,715 reads. As for the other two libraries, previously used and washed flow cells were re-used with only a fraction of sequencing pores being functional (402 vs. 256 pores), thus the acquired data were much lower (100,000 and 106,000 reads respectively).

**Computational analysis, assembly and annotation**

The quality of the Illumina reads was assessed with FastQC[48]. Genome size estimation was done for each paired raw-reads from individual nuclei with SGA-PreQC[49]. Contamination was assessed with Kraken[50] in some of the raw-reads. CG content was computed using the NBIS-UtilityCode[51] toolbox.

Assembly workflow 1: Individual assemblies for each of the 24 nuclei was done using MaSuRCA[34] using default options. The resulting assemblies were iteratively merged using Lingon[36], which computed overlaps based on the spacing of sequence motifs (CATG, CTAG, GTAC, GATC, TATA, ATAT, and GC), and merged contigs based on pairwise maximal extensions. Each motif was iterated over ten times. Three versions of the assembly were generated when contigs smaller than <500, <1000 and <2000 were removed from the individual assemblies prior to Lingon.

Assembly workflow 2: Each set of reads was normalized using bbnorm of BBMap[52] v. 38.08 with a target average depth of 100x. Normalized data were assembled individually into 24 assemblies using SPADES[35], and a consensus assembly was generated with Lingon[36], with the same sequence motifs as for assembly 1.

Assembly workflow 3: The 24 datasets were combined and normalized with bbnorm of BBMap[52] v. 38.08 with a target average depth of 100x, and posteriorly assembled using SPADES[35].

Nanopore assembly: Nanopore reads were assembled using Canu[37] v.1.7-86da76b, this specific beta version made it possible to assemble a difficult dataset like ours, with highly uneven coverage across the genome. An assembly was created using default settings together with the known information (genomeSize=117m -Nanopore-raw). The resulting individual assembly was polished with three rounds of Pilon[38] v.1.22 using the raw Illumina reads from the 24 nuclei mapped with Bowtie2[53].

The contigs of the final assemblies from single nuclei were scaffolded with the Nanopore assembly using Chromosemble from the Satsuma package[39].

**Comparative assembly analysis**

A quantitative assessment of the assemblies was done with Quast[54] v.4.5.4 and a contamination check with Kraken[50] v1.0. In addition, a BUSCO[42] analysis was done to assess completeness of the genome. The BUSCO lineage set used was fungi_odb9 and the species set was rhizopus_oryzae. (Figure 3, Figure S

Raw-reads were mapped to the individual assemblies of method 1 and 2 (Table S5) with Bowtie2[53] v. 2.3.3.1 using the default settings.

Two genes, known to be single copy genes in fungal genomes, as elongation factor 1-alpha (EF1-alpha) and the largest subunit of RNA polymerase II (RPB1), were searched for in the genome assemblies to test for possible duplications generated by the assembly methods. Sequences belonging to *C. claroideum* were used to find the sequences with BLASTn[55] (Table S6). Genebank sequences: EF1-alpha GQ205008.1, RPB1 HG316018.1.

**Genome annotation**

Repeats and transposable elements (TEs) were *de novo* predicted in every assembly using RepeatModeler[56] v1.0.8. The repeat library from RepeatModeler was used to mask the genome assembly using RepeatMasker[57] v4.0.7. The classification reports can be found in the OSF Repository[44].

Protein coding genes were *de novo* predicted from the repeat-masked scaffolded genome assembly with GeneMark-ES[58] v4.33. GeneMark-ES uses unsupervised self-training and an algorithm that is optimized for fungal gene organization. To guide the gene predictions, we aligned UniProt/Swiss-Prot[59] protein sequences (downloaded 8 May 2018) to the repeat-masked genome assembly using MAKER[60] v3.01.1-beta and provided the genomic locations of the protein alignments to GeneMark-ES. The previously published transcriptomic data from *C. claroideum*[61] was not used to due to the low mapping success of the reads to the assembly (25%), which could be related

426 to the low BUSCO statistics shown in the study[61], and that could have negatively
427 affected the annotation quality.
428 Protein and gene names were assigned to the gene predictions using a BLASTx[55]
429 v2.6.0 search of predicted mRNAs against the UniProt/Swiss-Prot[59] database with
430 default e-value parameters (1x10-5). The ANNotation Information Extractor, Annie[62],
431 was used to extract BLAST matches and to reconcile them with the gene predictions.
432
433 Sequences, assemblies and annotation can be found in the BioProject: PRJNA528883
434
435 **References**

436 1. James, T. Y. & Berbee, M. L. No jacket required - new fungal lineage defies
437 dress code: Recently described zoosporic fungi lack a cell wall during trophic
438 phase. *BioEssays* (2012). doi:10.1002/bies.201100110
439 2. Rosling, A. *et al.* Archaeorhizomycetes: Unearthing an ancient class of
440 ubiquitous soil fungi. *Science* (2011). doi:10.1126/science.1206958
441 3. Spribille, T. *et al.* Basidiomycete yeasts in the cortex of ascomycete
442 macrolichens. *Science* (2016). doi:10.1126/science.aaf8287
443 4. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity.
444 *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1521291113
445 5. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* 1, (2016).
446 doi:10.1038/nmicrobiol.2016.48
447 6. Tyson, G. W. *et al.* Community structure and metabolism through
448 reconstruction of microbial genomes from the environment. *Nature* (2004).
449 doi:10.1038/nature02340
450 7. Saw, J. H. *et al.* Exploring microbial dark matter to resolve the deep archaeal

ancestry of eukaryotes. *Philos. Trans. R. Soc. B Biol. Sci.* (2015). doi:10.1098/rstb.2014.0328

8. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* (2005). doi:10.1128/AEM.71.6.3342-3347.2005

9. Yoon, H. S. *et al.* Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* (2011). doi:10.1126/science.1203163

10. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the science. *Nature Reviews Genetics* (2016). doi:10.1038/nrg.2015.16

11. Hug, L. A. *et al.* Critical biogeochemical functions in the subsurface are associated with bacteria from new phyla and little studied lineages. *Environ. Microbiol.* (2016). doi:10.1111/1462-2920.12930

12. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* 541**,** 353–358 (2017). doi:10.1038/nature21031

13. West, P. T., Probst, A. J., Grigoriev, I. V., Thomas, B. C. & Banfield, J. F. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* (2018). doi:10.1101/gr.228429.117

14. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* (2013). doi:10.1038/nature12352

15. Woyke, T., Doud, D. F. R. & Schulz, F. The trajectory of microbial single-cell sequencing. *Nature Methods* (2017). doi:10.1038/nmeth.4469

16. Ahrendt, S. R. *et al.* Leveraging single-cell genomics to expand the fungal tree of life. *Nat. Microbiol.* (2018). doi:10.1038/s41564-018-0261-0

17. Rantalainen, M. Application of single-cell sequencing in human cancer. *Brief. Funct. Genomics* (2017). doi:10.1093/bfgp/elx036

476 18. Yuan, Y., Lee, H. T., Hu, H., Scheben, A. & Edwards, D. Single-cell genomic
477 analysis in plants. *Genes* (2018). doi:10.3390/genes9010050
478 19. Srivastava, M. *et al.* The *Amphimedon queenslandica* genome and the
479 evolution of animal complexity. *Nature* (2010). doi:10.1038/nature09201
480 20. Tuovinen, V. *et al.* Two Basidiomycete fungi in the cortex of wolf lichens.
481 *Curr. Biol.* (2019). doi:10.1016/j.cub.2018.12.022
482 21. Cuomo, C. A. *et al.* Microsporidian genome analysis reveals evolutionary
483 strategies for obligate intracellular growth. *Genome Res.* (2012).
484 doi:10.1101/gr.142802.112
485 22. Gardner, M. J. *et al.* Genome sequence of the human malaria parasite
486 *Plasmodium falciparum*. *Nature* (2002). doi:10.1038/nature01097
487 23. Tan, X. *et al.* Diversity and bioactive potential of culturable fungal endophytes
488 of *Dysosma versipellis*; A rare medicinal plant endemic to China. *Sci. Rep.*
489 (2018). doi:10.1038/s41598-018-24313-2
490 24. Kaul, S., Sharma, T. & K. Dhar, M. “Omics” tools for better understanding the
491 plant–endophyte interactions. *Front. Plant Sci.* (2016).
492 doi:10.3389/fpls.2016.00955
493 25. Parniske, M. Arbuscular mycorrhiza: The mother of plant root endosymbioses.
494 *Nature Reviews Microbiology* (2008). doi:10.1038/nrmicro1987
495 26. Humphreys, C. P. *et al.* Mutualistic mycorrhiza-like symbiosis in the most
496 ancient group of land plants. *Nat. Commun.* (2010). doi:10.1038/ncomms1105
497 27. Bonfante, P. & Genre, A. Mechanisms underlying beneficial plant-fungus
498 interactions in mycorrhizal symbiosis. *Nat. Commun.* 1, 48 (2010).
499 doi:10.1038/ncomms1046
500 28. Tisserant, E. *et al.* Genome of an arbuscular mycorrhizal fungus provides

insight into the oldest plant symbiosis. *Proc. Natl. Acad. Sci.* 110, 20117–20122 (2013). doi:10.1073/pnas.1313452110

29. Lin, K. *et al.* Single Nucleus genome sequencing reveals high similarity among nuclei of an endomycorrhizal fungus. *PLoS Genet.* 10, (2014). doi:10.1371/journal.pgen.1004078

30. Chen, E. C. H. *et al.* High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*. *New Phytol.* (2018). doi:10.1111/nph.14989

31. Kobayashi, Y. *et al.* The genome of *Rhizophagus clarus* HR1 reveals a common genetic basis for auxotrophy among arbuscular mycorrhizal fungi. *BMC Genomics* (2018). doi:10.1186/s12864-018-4853-0

32. Sun, X. *et al.* Genome and evolution of the arbuscular mycorrhizal fungus *Diversispora epigaea* (formerly *Glomus versiforme*) and its bacterial endosymbionts. *New Phytol.* (2018). doi:10.1111/nph.15472

33. Morin, E. *et al.* Comparative genomics of *Rhizophagus irregularis*, *R. cerebriforme*, *R. diaphanus* and *Gigaspora rosea* highlights specific genetic features in Glomeromycotina. *New Phytol.* (2019). doi:10.1111/nph.15687

34. Zimin, A. V. *et al.* The MaSuRCA genome assembler. *Bioinformatics* (2013). doi:10.1093/bioinformatics/btt476

35. Bankevich, A. *et al.* SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* (2012). doi:10.1089/cmb.2012.0021

36. Grabherr, M. G. Lingon: A d-mer based genome assembly pipeline. (2018).

37. Koren, S. *et al.* Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* (2017).

526 doi:10.1101/gr.215087.116

527 38. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial
528 variant detection and genome assembly improvement. *PLoS One* (2014).
529 doi:10.1371/journal.pone.0112963

530 39. Grabherr, M. G. *et al.* Genome-wide synteny through highly sensitive sequence
531 alignment: Satsuma. *Bioinformatics* (2010). doi:10.1093/bioinformatics/btq102

532 40. Spits, C. *et al.* Whole-genome multiple displacement amplification from single
533 cells. *Nat. Protoc.* (2006). doi:10.1038/nprot.2006.326

534 41. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple
535 displacement amplification. *Proc. Natl. Acad. Sci.* (2002).
536 doi:10.1073/pnas.082089499

537 42. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov,
538 E. M. BUSCO: Assessing genome assembly and annotation completeness with
539 single-copy orthologs. *Bioinformatics* (2015).
540 doi:10.1093/bioinformatics/btv351

541 43. Simpson, J. T. Exploring genome characteristics and sequence quality without
542 a reference. *Bioinformatics* (2014). doi:10.1093/bioinformatics/btu023

543 44. Montoliu-Nerin, M. OSF Repository - From Single Nuclei To Whole Genome
544 Assemblies. (2018). Available at: https://osf.io/yvwur/.

545 45. Rinke, C. *et al.* Obtaining genomes from uncultivated environmental
546 microorganisms using FACS-based single-cell genomics. *Nat. Protoc.* (2014).
547 doi:10.1038/nprot.2014.067

548 46. Ihrmark, K. *et al.* New primers to amplify the fungal ITS2 region - evaluation
549 by 454-sequencing of artificial and natural communities. *FEMS Microbiol.*
550 *Ecol.* (2012). doi:10.1111/j.1574-6941.2012.01437.x

551 47. Herlemann, D. P. R. *et al.* Transitions in bacterial communities along the 2000
552 km salinity gradient of the Baltic Sea. *ISME J.* (2011).
553 doi:10.1038/ismej.2011.41
554 48. Andrews, S. FastQC: A quality control tool for high throughput sequence data.
555 *Http://Www.Bioinformatics.Babraham.Ac.Uk/Projects/Fastqc/* (2010).
556 doi:citeulike-article-id:11583827
557 49. Simpson, J. SGA-PreQC. (2013). Available at:
558 https://github.com/jts/sga/wiki/preqc. (Accessed: 26th November 2017)
559 50. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence
560 classification using exact alignments. *Genome Biol.* (2014). doi:10.1186/gb-
561 2014-15-3-r46
562 51. Grabherr, M. G. NBIS-UtilityCode. Available at:
563 https://github.com/NBISweden/NBIS-UtilityCode. (Accessed: 16th September
564 2018)
565 52. Bushnell, B. BBMap short read aligner. *Joint Genome Institute, department of*
566 *energy* (2014). doi:10.1016/j.avsg.2010.03.022
567 53. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2.
568 *Nat. Methods* (2012). doi:10.1038/nmeth.1923
569 54. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality
570 assessment tool for genome assemblies. *Bioinformatics* (2013).
571 doi:10.1093/bioinformatics/btt086
572 55. Camacho, C. *et al.* BLAST+: Architecture and applications. *BMC*
573 *Bioinformatics* (2009). doi:10.1186/1471-2105-10-421
574 56. Smit, A. & Hubley, R. RepeatModeler Open-1.0.
575 57. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0.

576 58. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M.
577 Gene prediction in novel fungal genomes using an ab initio algorithm with
578 unsupervised training. *Genome Res.* (2008). doi:10.1101/gr.081612.108
579 59. Bateman, A. *et al.* UniProt: The universal protein knowledgebase. *Nucleic*
580 *Acids Res.* (2017). doi:10.1093/nar/gkw1099
581 60. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for
582 emerging model organism genomes. *Genome Res.* (2008).
583 doi:10.1101/gr.6743907
584 61. Beaudet, D. *et al.* Ultra-low input transcriptomics reveal the spore functional
585 content and phylogenetic affiliations of poorly studied arbuscular mycorrhizal
586 fungi. *DNA Res.* (2018). doi:10.1093/dnares/dsx051
587 62. Tate, R., Hall, B., DeRego, T. & Geib, S. Annie: the ANNotation Information
588 Extractor (Version 1.0). (2014). Available at:
589 http://genomeannotation.github.io/annie. (Accessed: 16th September 2018)
590 63. Gardes, M. & Bruns, T. D. ITS primers with enhanced specificity for
591 basidiomycetes - application to the identification of mycorrhizae and rusts.
592 *Mol. Ecol.* (1993). doi:10.1111/j.1365-294X.1993.tb00005.x
593 64. Bowers, R. M. *et al.* Corrigendum: Minimum information about a single
594 amplified genome (MISAG) and a metagenome-assembled genome (MIMAG)
595 of bacteria and archaea (Nature Biotechnology (2017) 35 (725-731) DOI:
596 10.1038/nbt.3893). *Nature Biotechnology* (2018). doi:10.1038/nbt0718-660a
597

598 **Acknowledgments**

599 We thank, J. Bever and S. Bertilsson for scientific discussions, Y. Strid and M.
600 Zakieh for assistance in the lab, J. Morton and W. Wheeler at INVAM culture

collection, and funding from ERC (678792). Sequencing was performed by the SNP&SEQ Technology Platform at NGI Sweden and SciLife Laboratory, Uppsala, supported by the VR and the KAW. Computations were performed on resources provided by SNIC through UPPMAX.

**Author contributions**

AR initiated the project and developed the method together with MMN and HJ. MSG helped develop the bioinformatic analysis. CB was in charge of the single nuclei facility and did the MDA, BE was in charge of the Nanopore sequencing. MG designed Lingon and helped with analysis together with MK. VK was in charge of the annotation. MMN was responsible for the project and wrote the manuscript with AR and HJ, with with input from all the authors.

625

626

627

628

629