SCIENTIFIC REPORTS
nature research

OPEN

Building *de novo* reference genome assemblies of complex eukaryotic microorganisms from single nuclei

Merce Montoliu-Nerin[1], Marisol Sánchez-García[1], Claudia Bergin[2], Manfred Grabherr[3], Barbara Ellis[1], Verena Esther Kutschera[4], Marcin Kierczak[3], Hanna Johannesson[5] & Anna Rosling[1*]

The advent of novel sequencing techniques has unraveled a tremendous diversity on Earth. Genomic data allow us to understand ecology and function of organisms that we would not otherwise know existed. However, major methodological challenges remain, in particular for multicellular organisms with large genomes. Arbuscular mycorrhizal (AM) fungi are important plant symbionts with cryptic and complex multicellular life cycles, thus representing a suitable model system for method development. Here, we report a novel method for large scale, unbiased nuclear sorting, sequencing, and *de novo* assembling of AM fungal genomes. After comparative analyses of three assembly workflows we discuss how sequence data from single nuclei can best be used for different downstream analyses such as phylogenomics and comparative genomics of single nuclei. Based on analysis of completeness, we conclude that comprehensive *de novo* genome assemblies can be produced from six to seven nuclei. The method is highly applicable for a broad range of taxa, and will greatly improve our ability to study multicellular eukaryotes with complex life cycles.

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], and 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 analyses 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.

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

[1]Department of Ecology and Genetics, Evolutionary Biology, Uppsala University, Uppsala, Sweden. [2]Department of Cell and Molecular Biology, Uppsala University, and Microbial Single Cell Genomics Facility, Science for Life Laboratory, Uppsala, Sweden. [3]Department of Cell and Molecular Biology, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Uppsala University, Uppsala, Sweden. [4]Department of Biochemistry and Biophysics, National Bioinformatics Infrastructure Sweden, Science for Life Laboratory, Stockholm University, Solna, Sweden. [5]Department of Organismal Biology, Systematics biology, Uppsala University, Uppsala, Sweden. *email: Anna.Rosling@ebc.uu.se

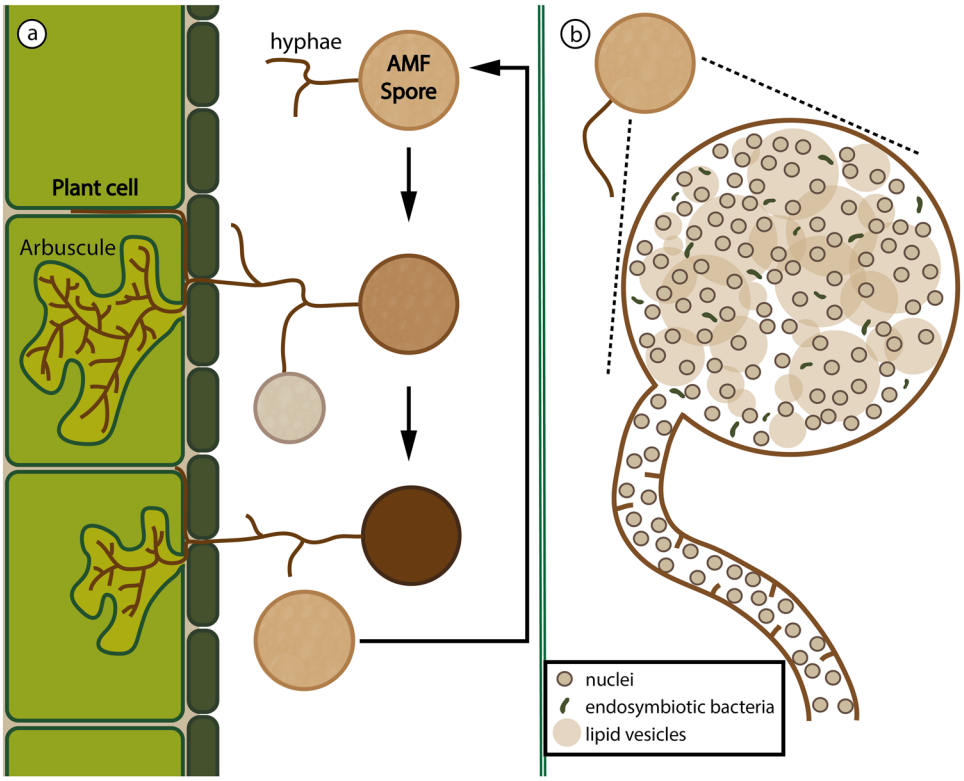

Figure 1. (**a**) Schematic representation of the life-cycle in AM fungi. A spore detects a plant root in the vicinity and grows hyphae towards it. The hyphae penetrate the plant cell wall and form the characteristically branching haustoria with the shape of arbuscules. The arbuscules are used to exchange 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.

with multinuclear asexual spores[28,29] (Fig. 1). Genomic research on AM fungi has been hampered by technical challenges involving isolation and culturing. Accordingly, reference nuclear genomes of only few species have been published[30–35], representing taxa that can be grown in axenic culture, *i.e., Rhizophagus irregularis*, *R. clarus*, *R. diaphanus*, *R. cerebriforme*, *Gigaspora rosea*, and *Diversispora epigaea*.

Methodological overview. A method was developed in which genomic fungal DNA can be obtained, free of plant and prokaryotic DNA, directly from individual nuclei of multinucleate spores. In brief, spores from a trap culture fungal strain of *Claroideoglomus claroideum/C. luteum* (SA101) were obtained from the INVAM pot culture collection. After visually confirming that nuclear size was appropriate for the method (Fig. S1), an initial trial to sort AM nuclei was carried out using pools of spores in order to assess optimal settings. Spores were cleaned, crushed vigorously, and stained with a DNA stain, before being analyzed by fluorescence-activated cell sorting (FACS), by recording level of fluorescence as a measure of DNA content and light scattering as proxy for size and particle granularity (Fig. 2a–h). A distinct cloud of particles was observed above the background in the scatter plot (Fig. 2h, inside the blue box), which by PCR verification with fungal and bacterial specific primers was confirmed to consist of biological structures containing mostly fungal DNA (Figs. S2–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 (Fig. S4, Table S2) and whole genome amplified (WGA) using multiple displacement amplification (MDA; Fig. 2I,j). The amplified DNA was screened for pure fungal origin by parallel amplification of rDNA barcode regions for both fungi and bacteria (Figs. 2k, S5). Twenty-four amplified nuclei samples confirmed to contain only fungi (Fig. S4, Table S3, S4), were sequenced with Illumina HiSeq X (Fig. 2l). Further, the MinION Nanopore-based sequencing device (Oxford Nanopore Technologies, ONT, UK) was used to obtain long read sequences from amplified DNA from multiple (5–100) nuclei separated from a pool of 30 spores of the same strain (Fig. 2i–k, m).

Three customized assembly workflows were developed 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 random coverage bias[36,37].

For the first two assembly workflows, individual nuclei assemblies were generated and subsequently combined to generate a consensus assembly using the workflow manager Lingon[38] (Fig. 2p), which consists of a

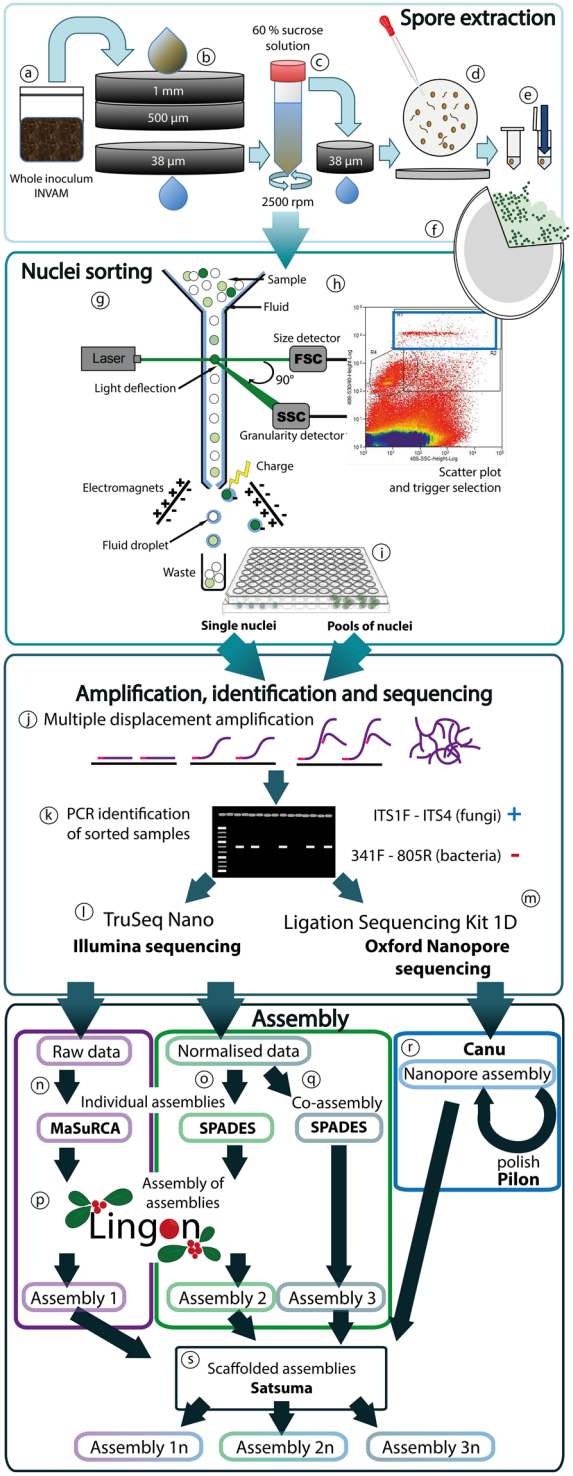

Figure 2. From a soil sample to AM fungal genome assemblies. (**a**) Whole inoculum from the culture collection INVAM is blended with water and (**b**) poured into a set of sieves; the material stuck in the 38 µm sieve is placed into a (**c**) tube that contains a solution of 60% sucrose, then centrifuged for 1 min. The supernatant is 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 ddH_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 a lighter signal for the background, organelles and microbes. (**g**) The stained spore content is loaded on the FACS, where 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 scatterplot, 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. 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. **(n)** In workflow 1, Illumina reads are assembled separately for individual nuclei using MaSuRCA[39]. **(o)** In workflow 2, reads from individual nuclei are normalized and assembled with SPADES[40]. **(q)** In workflow 3 reads from all nuclei are combined, then normalized and finally assembled with SPADES[40]. **(p)** Lingon[38] is used to produce a consensus assembly from individual nuclei assemblies in both workflows 1 and 2. **(r)** Nanopore data is assembled with Canu[41], polished with Pilon[53] using the Illumina raw-reads and used to **(s)** scaffold the three assemblies generated with workflows 1, 2 and 3 using Chromosemble, of Satsuma[55].

motif-distance based long sequence overlap finder that merges sequences based on mutual maximal overlaps. In the first assembly workflow raw Illumina reads were assembled using MaSuRCA[39] (Fig. 2n) resulting in 24 assemblies, ranging in size from 14 to 69 Mb (Tables S5). To overcome MDA-generated differences in coverage across the genome, the second workflow normalized raw reads to average 100X before assembling using SPADES[40] (Fig. 2o), generating 24 assemblies ranging in size from 11 to 50 Mb (Table S5). A third assembly was created using SPADES[40] after combining raw reads from 24 nuclei followed by normalization to 100X (Fig. 2q). One assembly with 24 nuclei was generated from each workflow and subsequently scaffolded with a Nanopore assembly built with Canu[41] (Fig. 2r,s). To evaluate the number of nuclei needed for a complete assembly, results from BUSCO[42] analyses, assembly size, and N50 were plotted across assemblies resulting from an increasing number of assembled nuclei. Data from different nuclei were merged in random combinations of two to twelve nuclei and one random combination for 13–23 nuclei. The analysis was performed separately for the three workflows and the results were compared with the single- and 24-nuclei assemblies.

Results

The different assembly workflows resulted in assemblies that vary in size, fragmentation and completeness (Table 1). Based on BUSCO analyses, workflow 3 generated the most complete assembly, with 89% for assembly 3n, compared to 2n at 80%, and 1n at 78% (Table 1). Of the core single copy genes identified by BUSCO, few were fragmented or duplicated in assembly 3n indicating that the set of 14,600 predicted genes is likely to be complete and a close representation of the genetic content in this strain (Table 1). This number is lower than the number of genes found in other sequenced AM fungi such as *R. irregularis*[30] and *R. clarus*[33], and also lower than those predicted in assemblies 1n and 2n (Table 1). Interestingly, assembly 3n is considerably smaller (70.8 Mb) than the other assemblies (92.4 Mb and 130.4 Mb for assembly 1n and 2n, respectively) and markedly smaller than the average estimated genome size of 119 Mb based on SGA-PreQC[43]. The smaller assembly size of 3n can be attributed to repeat sequences (20.6 Mb) that are captured to a lesser extent, compared to the other assembly workflows (41.3–58.6 Mb). Specifically, normalization is expected to disproportionally reduce high coverage genomic sequences such as repeat elements and collapse those regions when assembling. Note that this effect of normalization is eluded in assembly workflow 2, in which nuclei are normalized and assembled individually; repetitive regions will collapse but in different parts of the genome. Thus repeats end up being represented in the final assembly when single nuclei assemblies are 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). However, workflow 1 generates 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). Scaffolding with nanopore improves contiguity of all three assemblies by reducing the number of contigs and thus increasing N50. Furthermore, it decreases the number of genes, but does not affect BUSCO results or inferred repeat content in a major way (Table 1). Hence, in this study, nanopore data is not essential to produce biologically informative assemblies. The assembly from nanopore data alone gave a similar number of predicted genes compared to assembly 3, but captured more repeats (47.3 Mb). BUSCO results suggest a completeness of 77%, which is comparable to assemblies 1 and 2 (Table 1). It is important to notice that this nanopore assembly was polished with Illumina reads and that the completeness based on BUSCO results increased from 17% before polishing[44] to 77% after three rounds of polishing.

Combinations of increasing number (1–24) of randomly selected 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 Fig. 3, single nuclei assemblies are most complete when using normalized reads in workflow 2, with an average of 40% BUSCO estimated completeness compared to 25% in workflow 1. Interestingly, there is an increasing number of duplicated genes among the complete genes as more single nuclei assemblies are combined for method 2 compared to method 1 (Fig. 3a,b). Higher amount of duplicated genes 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. Assemblies of increasing number of nuclei result in increasing assembly size, N50, and BUSCO estimated completeness (Fig. 3). In both workflow 1 and 3, BUSCO results reach maximum performance when assembling random combinations of six - seven nuclei (Fig. 3a,c). The same pattern is observed for assembly size and N50 (Fig. 3d). In workflow 2, on the other hand, assembly size continuously increases with increasing number of combined nuclei assemblies (Fig. 3c). This pattern is reflected by an increasing number of duplicated genes in the BUSCO results (Fig. 3b).

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 F: 10	18068 (49.42)	40.39
1n	+Nanopore	92.38	3899	37258	176.652	27.91	C: 78 F: 9	16680 (69.54)	41.32
2	Normalized to 100×	124.96	21934	16055	155.09	28.07	C: 79 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 100×	68.31	11246	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
Nanopore	polished with Pilon	96.03	6409	20944	151.76	28.15	C: 77 F: 6	15858 (57.47)	47.31

Table 1. Comparative assessment of the 3 assembly workflows. [a]Completeness estimated in % of 290 single copy genes in fungi, scored as complete (C) or fragmented (F).

Discussion

Methodological challenges in assembling genomes from amplified single nuclei or cells can be tackled by careful analysis of generated assemblies[9,16,23]. In this study, it is suggested that different assembly strategies can be useful for different downstream analyses. A genome assembly with a high coverage and a high-quality dataset of single copy genes can already be generated from only six individually sequenced nuclei when reads are combined and normalized, as done in workflow 3 (Fig. 3). As demonstrated by Ahrendt *et al.*[16], such an assembly generates high coverage genome data and is ideally suited for phylogenomics studies. When using non-normalized data, as in assembly workflow 1, repeat elements are better represented and hence, this assembly is likely better suited for identification and classification of repeats, which are known to represent a large proportion of AM fungal genomes[34]. Comparative genetic analyses between single nuclei are best done using assemblies from workflow 2, where single nuclei assemblies are generated from normalized reads. Estimated completeness of these assemblies is comparable to results from single cell sequencing of fungi with smaller genomes[16]. However, single nuclei assemblies based on normalized reads should not be assembled into consensus assemblies since variable quality of contigs make them prone to duplication.

To conclude, sequence data from single cell sequencing presents itself as challenging, but as shown here, with the right combination of methods adapted to the data, *de novo* reference genomes can be generated, opening the door for an expansion in genomic and phylogenomic research in organisms like AM fungi, that have, for too long, evaded large scale genome sequencing efforts due too methodological limitations stemming from their complicated biology. With organism-specific modifications to the initial nuclei extraction step, the complete workflow can be adapted to investigate nuclei or other intraorganismal units, such as endosymbiotic bacteria or mitochondria, from taxonomically diverse groups of non-model organisms. Useful genomic information can be generated from a handful of single nuclei greatly improving our ability to study multicellular eukaryotes with complex life stages. The assembly method of choice will ultimately depend on the research questions asked and the kind of data needed or available.

Methods

Fungal strain and spore extraction. *C. claroideum/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. Due to the unclear taxonomic status of the strain we have decided to adhere to the current INVAM name throughout the text. 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 × 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 off with water. The contents were poured into 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 ddH_2O 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 ddH_2O 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 loading the sample on the FACS. The nuclei were sorted on 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 trigger channel was set to the forward scatter (FSC) at a threshold of 0.03% and sort regions were defined on SYBR Green I fluorescence (488–530/40) over side scatter (SSC). The samples were sorted in single cell 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 (Fig. 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).

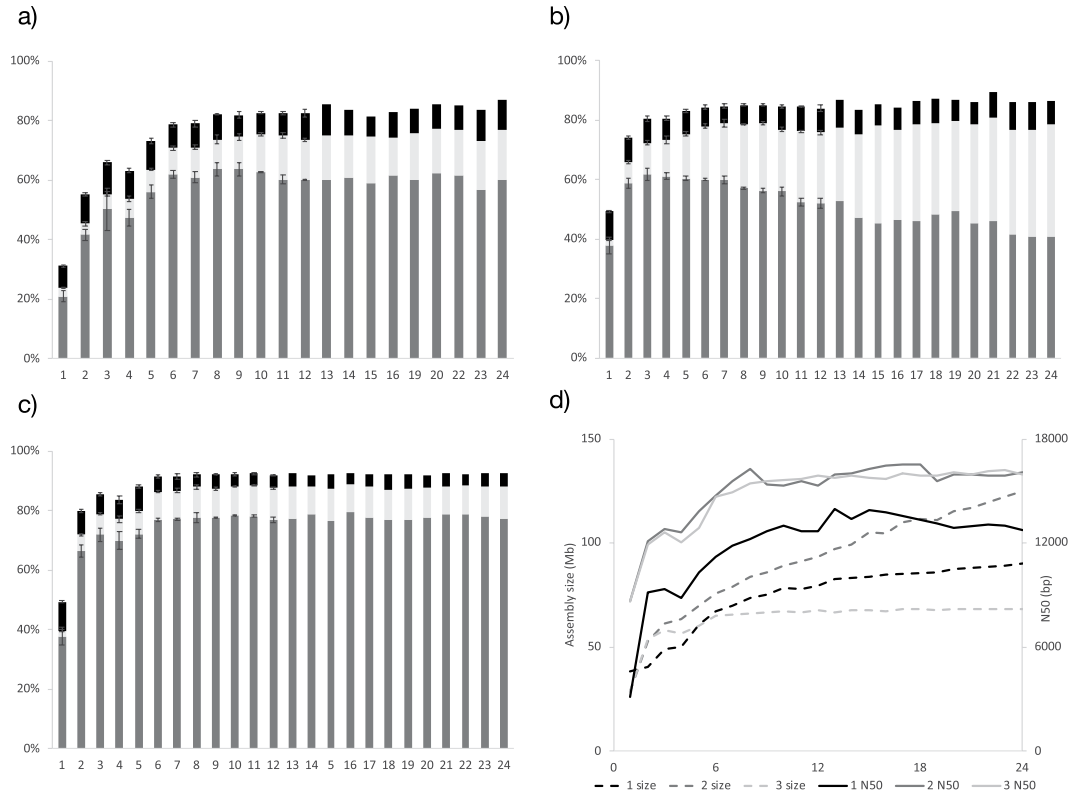

Figure 3. Summary statistics for different number of assembled nuclei (1–24) using three different assembly workflows. BUSCO estimates of completeness for (**a**) workflow 1: raw reads of individual nuclei assembled using MaSuRCa, consensus assembly using Lingon (**b**) workflow 2: normalized reads of individual nuclei assembled using SPADES, consensus assembly using Lingon and (**c**) workflow 3: reads from individual nuclei are pooled and normalized 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 three methods 1 (black), 2 (grey) and 3 (light grey).

Whole genome amplification. Sorted nuclei were lysed and neutralized followed by whole genome amplification using Phi29 and MDA as described by Rinke *et al.*[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.*[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 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 (40 ×) and screened for the presence of fungal and bacterial ribosomal genes using PCR. PCR reaction mixtures contained 10x Standard Taq Reaction buffer (Qiagen), 2 mM MgCl2, 0.2 mM deoxynucleoside triphosphates (dNTPs), 0.2 µM of each primer, and 1 U Taq DNA polymerase (Qiagen). The fungal-specific primers ITS9[46] and ITS4 were used. 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 341 F/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), 0.2 µM concentration of each primer and 1 U Taq DNA polymerase (Qiagen). DNA extracted from commercially available *Agaricus bisporus* provided by Dr. Ylva Strid (Uppsala University, Sweden), was included as a positive control, and ddH2O as negative control. The bacterial PCR protocol consisted of 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 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, Spain), and ddH_2O was used as negative control. The reaction was performed with a 2720 Thermocycler (Applied Biosystems, USA). The presence of amplification products was verified by gel electrophoresis by separation on a 2% agarose gel run for 35 min at 110 V (fungi) and 70 V (bacteria) including a Thermo Scientific GeneRuler 100 bp DNA Ladder (Fig. S5). The samples were identified as fungi positive, bacteria positive, fungi + bacteria positive or failed/empty (Table S3). From the samples that scored positive for presence of fungi, 24 undiluted samples were selected for sequencing and the DNA amount was measured using Qubit (Invitrogen, Austria) after addition of 30 µl ddH_2O (Table S4).

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

Spore sorting for Nanopore sequencing. Spores were picked in groups of 30 with the help of a P10 and P100 pipette, then washed five times in nuclease-free water and transferred to Eppendorf tubes in 30 uL nuclease-free water. For the FACS sorting spores were crushed, then 30 µl 1x PBS was added to the tube along with 1 µl of 200x SYBR Green for staining the DNA (20–50 mins). Sample volume was increased to 200 µl with 1x PBS before loading on the FACS. Pools of 5 and 100 nuclei were sorted into either individual 1.5 ml Eppendorf tubes or into multi-well plates. The above-described WGA protocol was run, and the 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 were made as described above. The fungal-specific ITS1F/ITS4 and bacteria-specific 341 F/805 R 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 using Qubit (Invitrogen, Austria). 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 analyses, 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 by MaSuRCA[39] using default options. The resulting assemblies were iteratively merged using Lingon[38], 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 100×. Normalized data were assembled individually into 24 assemblies using SPADES[40], and a consensus assembly was generated with Lingon[38], 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[40].

Nanopore assembly: Nanopore reads were assembled using Canu[41] 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 = 117 m -Nanopore-raw). The resulting assembly was polished with three rounds of Pilon[53] v.1.22 using the raw Illumina reads from the 24 nuclei mapped with Bowtie2[54]. The contigs of the final assemblies from single nuclei were scaffolded with the Nanopore assembly using Chromosemble from the Satsuma package[55].

Comparative assembly analyses. A quantitative assessment of the assemblies was done with Quast[56] v.4.5.4 and contamination was checked 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. (Figs. 3, S6)

Raw-reads were mapped to the individual assemblies of method 1 and 2 (Table S5) with Bowtie2[54] 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[57] (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[58] v1.0.8. The repeat library from RepeatModeler was used to mask the genome assembly using RepeatMasker[59] 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[60] 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[61] protein sequences (downloaded 8 May 2018) to the repeat-masked genome assembly using MAKER[62] v3.01.1-beta and provided the genomic locations of the protein alignments to GeneMark-ES. The previously published transcriptomic data from *C. claroideum*[63] was not used to due to the low mapping success of the reads to the assembly (25%), which could be related to the low BUSCO statistics shown in the study[63], and that could have negatively affected the annotation quality.

Protein and gene names were assigned to the gene predictions using a BLASTx[57] v2.6.0 search of predicted mRNAs against the UniProt/Swiss-Prot[61] database with default e-value parameters ($1 \times 10–5$). The ANNotation Information Extractor, Annie[64], was used to extract BLAST matches and to reconcile them with the gene predictions.

Sequences, assemblies and, annotations can be found in the BioProject: PRJNA528883.

Received: 2 August 2019; Accepted: 16 December 2019;
Published online: 28 January 2020

References

1. James, T. Y. & Berbee, M. L. No jacket required - new fungal lineage defies dress code: Recently described zoosporic fungi lack a cell wall during trophic phase. *BioEssays.* **34**, 94–102, https://doi.org/10.1002/bies.201100110 (2012).
2. Rosling, A. *et al.* Archaeorhizomycetes: Unearthing an ancient class of ubiquitous soil fungi. *Science.* **333**, 876–879, https://doi.org/10.1126/science.1206958 (2011).
3. Spribille, T. *et al.* Basidiomycete yeasts in the cortex of ascomycete macrolichens. *Science.* **353**, 488–492, https://doi.org/10.1126/science.aaf8287 (2016).
4. Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc. Natl. Acad. Sci.* **113**, 5970–5975, https://doi.org/10.1073/pnas.1521291113 (2016).
5. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048, https://doi.org/10.1038/nmicrobiol.2016.48 (2016).
6. Tyson, G. W. *et al.* Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43, https://doi.org/10.1038/nature02340 (2004).
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.* **370**, https://doi.org/10.1098/rstb.2014.0328 (2015).
8. Raghunathan, A. *et al.* Genomic DNA amplification from a single bacterium. *Appl. Environ. Microbiol.* **71**, 3342–3347, https://doi.org/10.1128/AEM.71.6.3342-3347.2005 (2005).
9. Yoon, H. S. *et al.* Single-cell genomics reveals organismal interactions in uncultivated marine protists. *Science* **332**, 714–717, https://doi.org/10.1126/science.1203163 (2011).
10. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: current state of the science. *Nat. Rev. Genet.* **17**, 175–188, https://doi.org/10.1038/nrg.2015.16 (2016).
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.* **18**, 159–173, https://doi.org/10.1111/1462-2920.12930 (2016).
12. Zaremba-Niedzwiedzka, K. *et al.* Asgard archaea illuminate the origin of eukaryotic cellular complexity. *Nature* **541**, 353–358, https://doi.org/10.1038/nature21031 (2017).
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.* **28**, 569–580, https://doi.org/10.1101/gr.228429.117 (2018).
14. Rinke, C. *et al.* Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**, 431–437, https://doi.org/10.1038/nature12352 (2013).
15. Woyke, T., Doud, D. F. R. & Schulz, F. The trajectory of microbial single-cell sequencing. *Nature Methods* **14**, 1045–1054, https://doi.org/10.1038/nmeth.4469 (2017).
16. Ahrendt, S. R. *et al.* Leveraging single-cell genomics to expand the fungal tree of life. *Nat. Microbiol.* **3**, 1417–1428, https://doi.org/10.1038/s41564-018-0261-0 (2018).

17. Rantalainen, M. Application of single-cell sequencing in human cancer. *Brief. Funct. Genomics* **17**, 273–282, https://doi.org/10.1093/bfgp/elx036 (2017).
18. Yuan, Y., Lee, H. T., Hu, H., Scheben, A. & Edwards, D. Single-cell genomic analysis in plants. *Genes* **9**, 50, https://doi.org/10.3390/genes9010050 (2018).
19. Srivastava, M. *et al*. The *Amphimedon queenslandica* genome and the evolution of animal complexity. *Nature* **466**, 720–726, https://doi.org/10.1038/nature09201 (2010).
20. Tuovinen, V. *et al*. Two basidiomycete fungi in the cortex of wolf lichens. *Curr. Biol.* **29**, 476–483, https://doi.org/10.1016/j.cub.2018.12.022 (2019).
21. Cuomo, C. A. *et al*. Microsporidian genome analysis reveals evolutionary strategies for obligate intracellular growth. *Genome Res.* **22**, 2478–2488, https://doi.org/10.1101/gr.142802.112 (2012).
22. Gardner, M. J. *et al*. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* **419**, 498–511, https://doi.org/10.1038/nature01097 (2002).
23. Tan, X. *et al*. Diversity and bioactive potential of culturable fungal endophytes of *Dysosma versipellis*; A rare medicinal plant endemic to China. *Sci. Rep.* **8**, 5929, https://doi.org/10.1038/s41598-018-24313-2 (2018).
24. Kaul, S., Sharma, T. K. & Dhar, M. "Omics" tools for better understanding the plant–endophyte interactions. *Front. Plant Sci.* **7**, 955, https://doi.org/10.3389/fpls.2016.00955 (2016).
25. Parniske, M. Arbuscular mycorrhiza: The mother of plant root endosymbioses. *Nat. Rev. Microbiol.* **6**, 763–775, https://doi.org/10.1038/nrmicro1987 (2008).
26. Humphreys, C. P. *et al*. Mutualistic mycorrhiza-like symbiosis in the most ancient group of land plants. *Nat. Commun.* **1**, 103, https://doi.org/10.1038/ncomms1105 (2010).
27. Bonfante, P. & Genre, A. Mechanisms underlying beneficial plant-fungus interactions in mycorrhizal symbiosis. *Nat. Commun.* **1**, 48 (2010).
28. Jany, J. L. & Pawlowska, T. E. Multinucleate spores contribute to evolutionary longevity of asexual glomeromycota. *Am. Nat.* **175**, 424–435, https://doi.org/10.1086/650725 (2010).
29. Marleau, J., Dalpé, Y., St-Arnaud, M. & Hijri, M. Spore development and nuclear inheritance in arbuscular mycorrhizal fungi. *BMC Evol. Biol.* **11**, 51, https://doi.org/10.1186/1471-2148-11-51 (2011).
30. Tisserant, E. *et al*. Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. *Proc. Natl. Acad. Sci.* **110**, 20117–20122, https://doi.org/10.1073/pnas.1313452110 (2013).
31. Lin, K. *et al*. Single Nucleus Genome Sequencing Reveals High Similarity among Nuclei of an Endomycorrhizal Fungus. *PLoS Genet.* **10**, e1004078 (2014).
32. Chen, E. C. H. *et al*. High intraspecific genome diversity in the model arbuscular mycorrhizal symbiont *Rhizophagus irregularis*. *New Phytol.* **220**, 1161–1171, https://doi.org/10.1111/nph.14989 (2018).
33. Kobayashi, Y. *et al*. The genome of *Rhizophagus clarus* HR1 reveals a common genetic basis for auxotrophy among arbuscular mycorrhizal fungi. *BMC Genomics* **19**, 465, https://doi.org/10.1186/s12864-018-4853-0 (2018).
34. Sun, X. *et al*. Genome and evolution of the arbuscular mycorrhizal fungus *Diversispora epigaea* (formerly *Glomus versiforme*) and its bacterial endosymbionts. *New Phytol.* **221**, 1556–1573, https://doi.org/10.1111/nph.15472 (2018).
35. Morin, E. *et al*. Comparative genomics of *Rhizophagus irregularis*, *R. cerebriforme*, *R. diaphanus* and *Gigaspora rosea* highlights specific genetic features in Glomeromycotina. *New Phytol.* **222**, 1584–1598, https://doi.org/10.1111/nph.15687 (2019).
36. Spits, C. *et al*. Whole-genome multiple displacement amplification from single cells. *Nat. Protoc.* **1**, 1965–1970, https://doi.org/10.1038/nprot.2006.326 (2006).
37. Dean, F. B. *et al*. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci.* **99**, 5261–5266, https://doi.org/10.1073/pnas.082089499 (2002).
38. Grabherr, M. G. Lingon: A d-mer based genome assembly pipeline, https://github.com/NBISweden/lingon (2018).
39. Zimin, A. V. *et al*. The MaSuRCA genome assembler. *Bioinformatics* **29**, 2669–2677, https://doi.org/10.1093/bioinformatics/btt476 (2013).
40. Bankevich, A. *et al*. SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* **19**, 455–477, https://doi.org/10.1089/cmb.2012.0021 (2012).
41. Koren, S. *et al*. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **27**, 722–736, https://doi.org/10.1101/gr.215087.116 (2017).
42. 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* **31**, 3210–3212, https://doi.org/10.1093/bioinformatics/btv351 (2015).
43. Simpson, J. T. Exploring genome characteristics and sequence quality without a reference. *Bioinformatics* **30**, 1228–1235, https://doi.org/10.1093/bioinformatics/btu023 (2014).
44. Montoliu-Nerin, M. OSF Repository - From single nuclei to whole genome assemblies of arbuscular mycorrhizal fungi, https://osf.io/yvwur/ (2018).
45. Rinke, C. *et al*. Obtaining genomes from uncultivated environmental microorganisms using FACS-based single-cell genomics. *Nat. Protoc.* **9**, 1038–1048, https://doi.org/10.1038/nprot.2014.067 (2014).
46. Ihrmark, K. *et al*. New primers to amplify the fungal ITS2 region - evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol. Ecol.* **82**, 666–677, https://doi.org/10.1111/j.1574-6941.2012.01437.x (2012).
47. Herlemann, D. P. R. *et al*. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J.* **5**, 1571–1579, https://doi.org/10.1038/ismej.2011.41 (2011).
48. Andrews, S. FastQC: A quality control tool for high throughput sequence data. (2010), https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (Accessed 21st October 2019).
49. Simpson, J. SGA-PreQC. (2013), https://github.com/jts/sga/wiki/preqc. (Accessed: 26th November 2017).
50. Wood, D. E. & Salzberg, S. L. Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biol.* **15**, R46, https://doi.org/10.1186/gb-2014-15-3-r46 (2014).
51. Grabherr, M. G. NBIS-UtilityCode, https://github.com/NBISweden/NBIS-UtilityCode. (Accessed: 16th September 2018).
52. Bushnell, BBMap: A Fast, Accurate, Splice-Aware Aligner. Joint Genome Institute, department of energy (2014). https://sourceforge.net/projects/bbmap/ (Accessed 21st October 2019).
53. Walker, B. J. *et al*. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963, https://doi.org/10.1371/journal.pone.0112963 (2014).
54. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359, https://doi.org/10.1038/nmeth.1923 (2012).
55. Grabherr, M. G. *et al*. Genome-wide synteny through highly sensitive sequence alignment: Satsuma. *Bioinformatics* **26**, 1145–1151, https://doi.org/10.1093/bioinformatics/btq102 (2010).
56. Gurevich, A., Saveliev, V., Vyahhi, N. & Tesler, G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* **29**, 1072–1075, https://doi.org/10.1093/bioinformatics/btt086 (2013).
57. Camacho, C. *et al*. BLAST+: Architecture and applications. *BMC Bioinformatics* **10**, 421, https://doi.org/10.1186/1471-2105-10-421 (2009).
58. Smit, A. & Hubley, R. RepeatModeler Open-1.0, http://www.repeatmasker.org/RepeatModeler/ (Accessed 21st October 2019).

59. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-4.0, http://www.repeatmasker.org/RMDownload.html (Accessed 21st October 2019).
60. Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O. & Borodovsky, M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* **18**, 1979–1990, https://doi.org/10.1101/gr.081612.108 (2008).
61. The UniProt Consortium. UniProt: The universal protein knowledgebase. *Nucleic Acids Res.* **45**, D158–D169, https://doi.org/10.1093/nar/gkw1099 (2017).
62. Cantarel, B. L. *et al.* MAKER: An easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Res.* **18**, 188–196, https://doi.org/10.1101/gr.6743907 (2008).
63. Beaudet, D. *et al.* Ultra-low input transcriptomics reveal the spore functional content and phylogenetic affiliations of poorly studied arbuscular mycorrhizal fungi. *DNA Res.* **25**, 217–227, https://doi.org/10.1093/dnares/dsx051 (2018).
64. Tate, R., Hall, B., DeRego, T. & Geib, S. Annie: the ANNotation Information Extractor (Version 1.0). (2014), http://genomeannotation.github.io/annie. (Accessed: 16th September 2018).

Acknowledgements

We thank J. Bever and S. Bertilsson for scientific discussions, Y. Strid and M. Zakieh for assistance in the lab, J. Morton and W. Wheeler at INVAM culture collection, and funding from ERC (678792). Nuclei sorting and whole genome amplification was done at the SciLifeLab Microbial Single Cell Genomics Facility at Uppsala University. Sequencing was performed by the SNP&SEQ Technology Platform at NGI Sweden and SciLife Laboratory, Uppsala, supported by the VR and the KAW. Computational analyses were performed on resources provided by SNIC through UPPMAX. MG, MK and VK were financially supported by the Knut and Alice Wallenberg Foundation as part of the National Bioinformatics Infrastructure Sweden at SciLifeLab. Open access funding provided by Uppsala University.

Author contributions

A.R. initiated the project and developed the method together with M.M.N. and H.J. C.B. did the nuclei sorting and whole genome amplification. B.E. was in charge of the Nanopore sequencing. M.M.N. performed the bioinformatic analyses together with M.S.G., M.G. and M.K. M.G. designed Lingon and V.K. was in charge of the annotation. M.M.N. wrote the manuscript with A.R. and H.J., with input from all the authors.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-58025-3.

Correspondence and requests for materials should be addressed to A.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.