Supplemental Material S3: Bioinformatic analyses

1. SNP calling

1.1 Genomic DNA preparation and sequencing

For each homokaryotic isolate, genomic DNA was extracted from a mycelium culture in liquid Hagem medium (20 °C for 10 days) and a genomic library was prepared with an insert length of 400 bp as previously described (Lind et al. 2005; Dalman et al. 2013). Briefly, DNA was extracted using Qiagen Genomic tips columns 100 according to the manufacturers' protocol. High molecular weight DNA are fished out after addition of isopropanol and transferred to an Eppendorf tube containing 70% ethanol. DNA was then transferred to a new tube and let to air dry. Each library was sequenced from both ends with a HiSeq 1500 apparatus (Illumina, San Diego, CA) in order to generate paired-end reads of 150 bp. In addition, the genome of the reference isolate Sä_159-5 was sequenced using the PacBio RS II system (Pacific Biosciences, Menlo Park, CA). Sequence reads were deposited to the European Nucleotide Archive (ENA) Sequence Read Archive (SRA) under the project accession PRJEB27090.

1.2 Filtering of sequencing pair-reads and removal of adaptor sequences

Removal of Illumina adaptor sequences and filtering of low quality reads was carried out using the <clip> tool of software Nesoni v0.97 (https://github.com/Victorian-Bioinformatics-Consortium/nesoni). Only reads larger than 75 bp were kept (--length 75) and output paired-end reads were stored in separate left/right files (--out-separate yes).

1.3 *De novo* assembly of the genome of the reference isolate Sä_159-5

The PacBio sequence reads of the genome of isolate Sä_159-5 were assembled *de novo* at the Uppsala Genome Center (National Genomics Infrastructure in Uppsala, SciLifeLab, Sweden) using Hierarchical Genome Assembly Process version 3 (HGAP3, Chin *et al.* 2013). This *de novo* assembly was corrected with the Illumina reads clipped and filtered with Nesoni coming from the same isolate. Reads were aligned to the HGAP3 assemby using Bowtie2 v2.2.4 (Langmead and Salzberg, 2012), and differences between the Illumina data and the HGAP3 assembly were identified and collected in a VCF file using Freebayes v1.0.0-19-gefg685d (Garrison and Marth, 2012). Any differences found where the read depth was 50x or above, 80% paired-end reads and 90% of the reads supporting the difference were incorporated into the reference assembly. This was done using an in-house python script (https://github.com/mikdur/assembly_corrector). *H. parviporum* genome size was estimated at 34.4 Mb from this *de novo* assembly of the genome sequence of isolate Sä_159-5, also available under project accession number PRJEB27090.

1.3 Alignment of short read sequences

Average sequencing depth for all other isolates was estimated at 123x and ranged from 79x to 239x (Table S1). Average coverage of the reference genome was 94% and ranged from 91.74 to 96.04% (Table S1). Reads of all isolates were mapped onto the corrected assembly of the genome of the reference isolate using Bowtie2 v2.2.4 (Langmead and Salzberg, 2012). An index was built from the consensus fasta file of the *de novo* genome assembly of the reference isolate (-build index). The program was run with an option favouring sensitive and accurate results (--very-sensitive), unpaired reads that had passed the Nesoni quality filter were kept for reference assembly (-U), and SAM files were created for the outputs (-S). The <view> tool of Samtools v1.2 (http://www.htslib.org/doc/samtools.html) was used to convert these SAM files into BAM files, and to mark headers and read groups (-bhSr). Alignments of each

BAM file were sorted using the <sort> tool of Samtools. Duplicate reads were marked with the <MarkDuplicates> tool of Picard-tools v1.140 (http://picard.sourceforge.net) and read groups created with the <AddOrReplaceReadGroups> tool of the same program.

1.4 Parallel SNP calling

SNPs were called in parallel from the genome assemblies of all isolates using Freebayes v1.0.0-19-gefg685d (Garrison and Marth, 2012): the Fasta file corresponding to the Illumina-corrected *de novo* assembly of Sä_159-5 was used as reference (-f), ploidy was set to 1 in order to reflect the haploid nature of homokaryotic isolates (-p 1), a list of file names corresponding to all BAM files with their headers, read groups and marked duplicates was provided (-L), and the program was allowed to proceed by windows of 100 kb, the coordinates of which were stored in a separate file (-r), created by partitioning every scaffold of Sä_159-5 reference genome assembly at that pace. The sizes of all scaffolds were measured with an in-house Perl script and stored in an intermediate file, which was used to create a list of windows of 100 kb by the <makewindows> tool (options -g and -w) of Bedtools v2.16.2 (Quinlan and Hall 2010) with the following command:

```
makewindows -g /scaffolds_sizes_file_name.txt -w 100000 | awk '{printf("%s:%s\n", $1, $2, $3)}'
>/windows_file_name.txt
```

The template of the Freebayes command used is:

```
freebayes -f /reference_genome_file_name.fas -p 1 -L /list_of _genomes_files_names.txt -r $(head -
$SGE_TASK_ID /windows_file_name.txt | tail -1) > ${SGE_TASK_ID}_VCF_files_name.vcf
```

1.5 Molecular control of homokaryosis

Major allele frequencies of the biallelic SNP called in parallel from the assembled genome sequences of the 30 isolates were used as molecular control of the homokaryotic phase. SNPs for which the major allele frequency is under a specific value (between 0.5 and 1) in a specific isolate were retrieved with an in-house Perl script. In the VCF format, for each SNP and for each isolate, RO is the number of reads bearing the reference allele, AO the number of reads bearing the alternative allele, and DP the read depth (DP = AO + RO). Major allele frequencies were calculated by dividing RO by DP if RO > AO, or by dividing AO by DP if AO > RO. For each isolate, less than 0.3% of the SNPs had a major allele frequency below 0.7, confirming that all isolates are homokaryons.

2. SNP filtering and determination of genetic distances between homokaryotic isolates

SNPs were filtered using a successive set of in-house Perl scripts: 1) An in-house Bash script was used to numerically sort out the VCF files resulting from parallel SNP calling, and a Perl script to sequentially extract from each of them biallelic SNPs with a QUAL phred-scale quality score above 10,000 only, and for which sequence reads were found in every isolate. All SNPs extracted from each numerically ordered VCF files were concatenated in a single file and the VCF file heading added to it. SNPs were subsequently selected only if: 2) the number of reads corresponding to the reference allele and to the alternative allele were both different from 0 (an infrequent technical failure due to Freebayes); 3) the genotype of each SNP was supported by more than 90% of the reads for each isolate; 4) the minor allele was found in at least two isolates among the 30 of the collection; 5) they were not found in two scaffolds belonging to the mitochondrial genome. These two scaffolds bearing 145 mitochondrial SNPs were identified in two steps: First as bearing only SNPs supported by a very high read depth in every isolate, then by carrying out nucleotide BLAST searches against the annotated genome sequence of *H. irregulare* (Olson et al, 2012) with DNA fragments of

100 bp bearing SNPs randomly chosen in these two scaffolds and retrieved from the reference genome. 6) SNPs were finally filtered in order to ensure homogeneity of the number of sequence reads supporting them. A stringent selection procedure was designed and repeated for each isolate in order to avoid those showing extreme values of read depth compared to average. Only those fulfilling the following conditions for every isolate were kept: Mean (μ) and standard deviation (σ) of the read depth of all SNPs were first calculated for each isolate. SNPs were subsequently selected only if, for each isolate, their read depth was higher than 20 or μ - 2σ, and lower than μ + 2σ. 7) The genetic distance between two isolates was then determined as the pairwise sequence divergence (distance in the sense of Hamming) over the entire collection of filtered SNPs. 8) For the analysis of population structure, the filtered SNP collection was additionally filtered to remove fully linked SNPs: only the first SNP from each stretch of contiguous SNPs having identical genotypes was kept.

3. Analysis of population structure

For population structure analyses, three isolates (RB48_B2, FSE_7, Br518_c2) were excluded from the list because they are closely related to three other isolates sampled in the same locations (RB48_9, FSE_3, Br244_4 respectively, Figure S1). Structure within the sampled population of homokaryotic isolates was investigated using Structure v2.3.4 (Pritchard *et al.* 2000) and unlinked SNPs extracted from the filtered SNP collection. SNPs were analysed with presumed population subgroups (K) ranking from 1 to 4, first without user pre-definition, in an admixture model assuming that the organism is haploid. SNPs were subsequently analysed using sampling locations as prior information to assist the detection of population structure (model LOCPRIOR with no admixture), by pre-defining four population subgroups corresponding to four large geographic areas (see Table S1). For each run, the initial burn-in period was set to 10,000 and 20,000 replicates were carried out. All values of K were tested independently three times for both models. Pairwise F_{ST} values between each of the four pre-defined population subgroups were computed with a random sample of 30,000 unlinked SNPs using the Gene Flow and Genetic Differentiation tool of DnaSP v5 (Librado and Rozas 2009).

Supplementary literature cited:

Chin C.-S., Alexander D. H., Marks P., Klammer A. A., Drake J., *et al.*, 2013 Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. Nat. Methods 10: 563.

Librado P., Rozas J., 2009 DnaSP v5: a software for comprehensive analysis of DNA polymorphism data. Bioinformatics 25: 1451–1452.

Pritchard J. K., Stephens M., Donnelly P., 2000 Inference of population structure using multilocus genotype data. Genetics 155: 945–959.

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