Future Science Ltd
BioTechniques
Volume 72, Issue 2, February 2022, Pages 39-50
https://doi.org/10.2144/btn-2021-0097

Reports

A high-quality fungal genome assembly resolved from a sample accidentally contaminated by multiple taxa

Janneke Aylward [1,2,*], Michael J Wingfield [1], Francois Roets [2] & Brenda D Wingfield [1]

[1]Department of Biochemistry, Genetics & Microbiology, Forestry & Agricultural Biotechnology Institute (FABI), University of Pretoria, Private Bag X20, Hatfield, Gauteng, 0028, South Africa
[2]Department of Conservation Ecology & Entomology, Stellenbosch University, Private Bag X1, Matieland, Western Cape, 7602, South Africa
[*]Author for correspondence: janneke.aylward@fabi.up.ac.za

ABSTRACT

Contamination in sequenced genomes is a relatively common problem and several methods to remove non-target sequences have been devised. Typically, the target and contaminating organisms reside in different kingdoms, simplifying their separation. The authors present the case of a genome for the ascomycete fungus *Teratosphaeria eucalypti*, contaminated by another ascomycete fungus and a bacterium. Approaching the problem as a low-complexity metagenomics project, the authors used two available software programs, BlobToolKit and anvi'o, to filter the contaminated genome. Both the *de novo* and reference-assisted approaches yielded a high-quality draft genome assembly for the target fungus. Incorporating reference sequences increased assembly completeness and visualization elucidated previously unknown genome features. The authors suggest that visualization should be routine in any sequencing project, regardless of suspected contamination.

METHOD SUMMARY

Complementary use of the BlobToolKit and anvi'o programs made it possible to resolve DNA sequences originating from closely related organisms. The authors applied *de novo* and reference-assisted filtering of contaminated raw genomic reads and visualized the filtering process to distinguish between the genomic sequences of two ascomycetous fungi and a bacterium.

DATA DEPOSITION

The genome of *Teratosphaeria eucalypti* (isolate CMW54005) has been deposited in the National Center for Biotechnology Information (NCBI) genome repository under the accession number JAIZZA000000000. The reference-filtered assembly of CMW55930 has been submitted as a metagenome-assembled genome under the accession number JAJADS000000000.

First draft submitted: 4 October 2021; **Accepted for publication**: 11 November 2021; **Published online**: 30 November 2021.

KEYWORDS
ascomycete • contamination • eukaryotic • filtering • fungal • genome • metagenomic

Reports of non-target taxa in genome assembly databases are increasingly common. In some cases, human DNA [1,2] or non-sterile sampling reagents [3] result in contaminant sequences. Microbial sequences have been identified in human DNA extracts [4] and are present even within the human reference genome [5]. In other cases, an inherent inability to separate an obligate associate from its host [6,7] or the need to use an entire organism for DNA extraction, including internal symbionts [8], necessitates sequencing a mixed sample.

As the speed of sequencing and the abundance of genomes increases, scrutinizing quality becomes increasingly important. Basic assembly statistics such as size, contig number and N50 [9] can reveal, though not diagnose, obvious contamination issues. The straightforward process of assessing genome completeness by considering the presence and copy number of lineage-specific housekeeping genes [10,11] has become common practice. In conjunction with the N50 score, an assembly with low completeness or high completeness coupled with unexpected duplications should raise concern [11,12]. Unfortunately, checking contamination is more complicated than assessing completeness, most likely the reason it has not yet become standard practice for most single-organism genome projects.

Several software tools have been developed to visualize genome assemblies and detect potential contaminant sequences [13–16]. These programs typically follow principles applied in shotgun metagenomic studies, where the inherent aim is to assemble a mixed sample and identify different compartments [17,18]. Because different taxa typically have different sequence compositions [16], predictors such as GC% and tetra-nucleotide frequency (or other *k*-mer lengths) can be applied to identify taxonomic groups within a dataset [19]. If the amount of starting genetic material differs among taxa, sequence coverage also becomes an invaluable separator [19,20]. While these methods are powerful, detection and identification of contaminants can be further supplemented with taxonomic annotations [16].

Previous studies that have resolved mixed assemblies have focused primarily on separating taxa from different kingdoms, approaching the process as a 'low-complexity metagenomics project' [8,20]. The majority of these studies have targeted Animalia [6,8,13] or Fungi [21] contaminated with Bacteria. Though less abundant, examples of archaeal and viral contamination in eukaryotic genomes are also available [7,22]. At these high-level taxonomic differences, sequence composition statistics vary considerably [23] and Basic Local Alignment Search Tool (BLAST) annotations easily differentiate taxa [19]. Where the taxonomic distances between target and contaminant organisms are smaller, annotations become an increasingly useful supplement to the sequence statistics [16], although they are only as accurate as the information available in databases. This presents a particular challenge for *de novo* sequencing projects considering underrepresented taxa.

In this study, the authors targeted the plant pathogenic Dothideomycete fungus *Teratosphaeria eucalypti* [24,25] for *de novo* genome sequencing. Due to its slow growth on agar, they experimented with liquid culture. Before sequencing, the sample harvested from this culture was, however, stored at 4°C instead of -20°C, resulting in undetected growth of contaminants, including another ascomycete fungus. Consequently, the authors evaluated

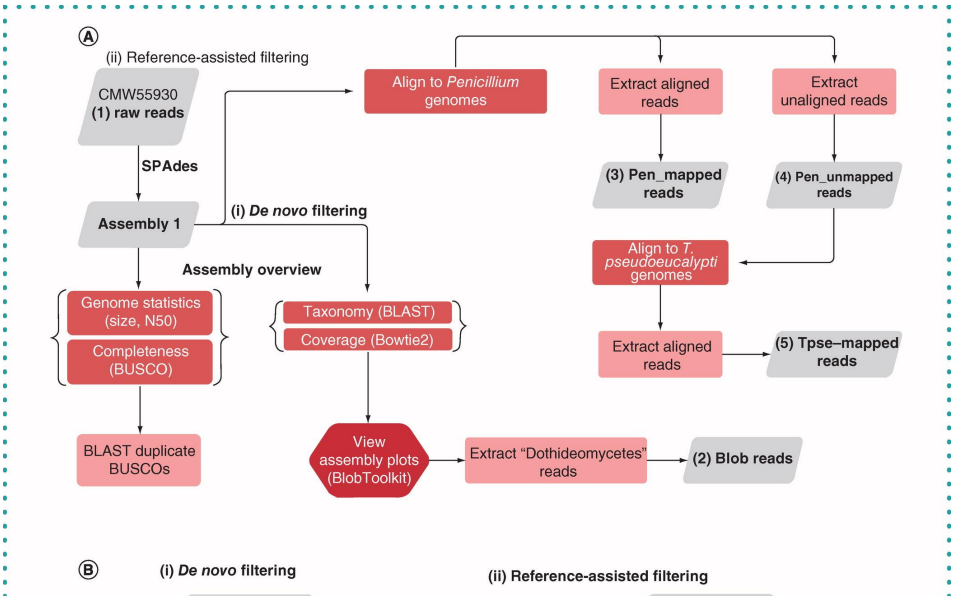
two software tools, BlobToolKit [16] and anvi'o [17], in an attempt to filter the contaminants within the genome. They illustrate a *de novo* as well as a reference-assisted filtering method and, using a second *T. eucalypti* genome as control, show that these methods can yield a high-quality draft genome assembly, even in the absence of reference sequences.

Materials & methods

Whole genome sequencing of *T. eucalypti* isolates CMW54005 and CMW55930 (BioProject PRJNA759074) was performed with the Ion Torrent™ GeneStudio™ S5 Prime System (Thermo Fisher Scientific, MA, USA) at the Central Analytical Facilities (CAF), Stellenbosch University, South Africa. Isolate CMW55930 was grown in nitrogen-deficient liquid media [26,27] for 4 weeks at room temperature. Harvested mycelium was washed in sterile water and kept at 4°C for approximately 2 weeks. Isolate CMW54005 was grown on Malt Extract Agar (Merck, Wadeville, South Africa) for 8 weeks at room temperature and harvested mycelium was freeze-dried immediately. DNA was extracted from both samples with the NucleoSpin Plant II kit (Machery-Nagel, Düren, Germany). Genome assemblies were computed with SPAdes 3.15.2 [28], applying the built-in error correction.

Assemblies were visualized as taxon-annotated Guanine-Cytosine (GC)-coverage plots [13,18] using BlobToolKit 1.2 [16] and as dendograms in anvi'o v7 [17]. Assembly coverage was determined by mapping reads with Bowtie 2.4.4 [29]. The putative taxonomic identity of contigs was determined with searches against the National Center for Biotechnology Information (NCBI) non-redundant nucleotide database using BLASTn 2.10.0 [30] and the Uniprot database with DIAMOND 2.0.9 [31], following the BlobToolKit instructions. Completeness was estimated with BUSCO 5.1.2 [10], using the Bacteria, Eukaryota and Fungi *OrthoDB* 10.1 datasets.

After detecting significant contamination of the initial assembly of isolate CMW55930 (Assembly 1), both *de novo* and reference-assisted filtering methods were applied (Figure 1A). For the *de novo* approach, putative eukaryotic and prokaryotic contigs were distinguished with EukRep 0.6.7 [23]. Contigs classified in BlobToolkit as Class 'Dothideomycetes' ('bestum' taxrule [16]) were extracted as a metagenome-assembled genome (MAG) and raw reads mapping to this MAG were obtained using BlobToolkit utilities.



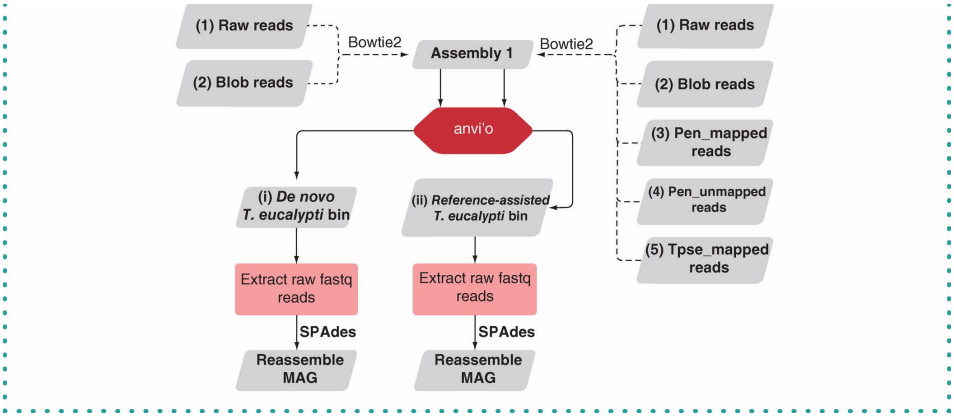


Figure 1. Workflow used to obtain filtered read sets.
(A) The initial raw reads of CMW55930 were filtered with a (i) *de novo* approach, using BlobToolKit taxonomy, and a (ii) reference-assisted approach that entailed alignment to the *Penicillium* and *Teratosphaeria pseudoeucalypti* genomes. (B) Read sets obtained from the two approaches were subsequently used to predict the boundaries of the *T. eucalypti* bin in anvi'o and extract the associated raw reads.

Reference-assisted filtering (Figure 1A) entailed using available genomes from species other than *T. eucalypti*. The raw reads were mapped to the putative fungal contaminant with Bowtie, using the *Penicillium camemberti* (GenBank assembly GCA_000513335.1), *P. crustosum* (GCA_014621375.1) and *P. solitum* (GCF_002072235.1) genomes. This was based on BLASTp hits of a subset of duplicated BUSCOs (Supplementary Table 1) and identification of the *Penicillium BenA* gene (Supplementary Table 2). Subsequently, reads that did not map to the contaminant were mapped to *T. pseudoeucalypti* (GCA_013403725.1 and GCA_013403725.1), the sister species of *T. eucalypti* [25].

All filtered CMW55930 read sets (Figure 1B) were mapped to Assembly 1. Using the coverage data of the *de novo* and *de novo* combined with the reference-filtered read sets, automatic binning of Assembly 1 was performed in CONCOCT 1.1.0 [32] and MetaBAT 2.15 [33]. Contig clustering, based on differential sequence coverage and tetra-nucleotide frequency, was viewed in anvi'o. With the guidance of the automated binning, the *T. eucalypti* and contaminant genome bins were selected manually. The default anvi'o Hidden Markov Models (HMMs) were used to identify single-copy genes predicted with Prodigal [34]. For both the *de novo* and reference-assisted approaches, the putative *T. eucalypti* MAG was extracted from Assembly 1 and the associated raw fastq reads were recovered with SAMtools 1.12 [35] and reassembled.

The efficacy of the filtering and re-assembly process was evaluated by comparing the filtered genome with the second genome sequence of *T. eucalypti* CMW54005. The raw reads for isolate CMW55930 were re-filtered by aligning them to the *T. eucalypti* reference and all filtered assemblies and MAGs were compared with this reference-based CMW55930 assembly and the CMW54005 reference assembly. Repeat content was estimated using custom repeat libraries constructed for each assembly with RepeatScout 1.0.6 and RepeatMasker 4.1.2 [36] and comparative assembly statistics were calculated with QUAST 5.0.2 [37].

Results & discussion

Mixed & reference genome assemblies for *T. eucalypti*

For isolate CMW55930, 13.62 million reads of 25–942 bp were obtained. The initial 76.5 Mb

assembly (Assembly 1; Table 1) was more than double the expected size, based on the 28–37 Mb assemblies of other *Teratosphaeria* species [38–42]. In contrast, the 12.09 million reads of isolate CMW54005 assembled a genome of 30.1 Mb, consistent with that expected for a *Teratosphaeria* species (Table 1).

Table 1. Genome assembly statistics† of the mixed and reference genome samples of *Teratosphaeria eucalypti*. (Table view)

	Assembly 1	Reference
Strain	CMW55930 (mixed sample)	CMW54005
Raw sequence reads	13618098	12092134
Assembly size (Mb)	76.45	30.06
Contigs ≥1000 bp	6184	2359
L50	168	141
N50 (kb)	113.91	63.97
Guanine-Cytosine (GC) content	49.37%	51.70%
Repeat content	8.24%	16.86%
BUSCO sequences identified		
Bacteria: single-copy	112 (90.32%)	27 (21.77%)
Bacteria: duplicate	4 (3.23%)	0
Eukaryota: single-copy	40 (15.69%)	248 (97.25%)
Eukaryota: duplicate	214 (83.92%)	0
Fungi: single-copy	152 (20.05%)	741 (97.76%)
Fungi: duplicate	606 (79.95%)	0

† All statistics are based on contigs ≥1000 bp.

BUSCO analysis of Assembly 1 detected high duplication levels (>79%) in the Eukaryota and Fungal datasets (Table 1). Additionally, 93.5% of Bacteria *odb10* BUSCOs were identified. A BLASTp of 16 randomly chosen duplicate Eukaryota BUSCOs revealed that half were consistent with a *T. eucalypti* origin, being most similar to Mycosphaerellales sequences, predominantly those from the Teratosphaeriaceae (Supplementary Table 1). The other half consistently hit to *Penicillium* proteins, with *P. camemberti* and *P. solitum* appearing as hits for 9/16 sequences. A BLASTn search for the *Penicillium* beta-tubulin (*BenA*) gene yielded a 729 bp high-scoring hit, with 100% similarity to *BenA* of *P. crustosum* (MN969379).

The BlobPlot (Figure 2) confirmed the presence of three main taxa in Assembly 1. Taxonomy filtering was applied at the highest rank separating the target and contaminant organisms and the majority of contigs were annotated as Eurotiomycetes (37.5 Mb; 49.1%), Dothideomycetes (27.9 Mb; 36.53%) and Bacilli (7.3 Mb; 9.6%). Consistent with a contaminated *T. eucalypti* sample, the Dothideomycete sequences had the highest mean coverage (*ca.* 130X), whereas the coverages of the two primary contaminants were approximately 15X (Eurotiomycetes) and 80X (Bacilli). The GC contents of the three taxa were distinct, although overlapping.

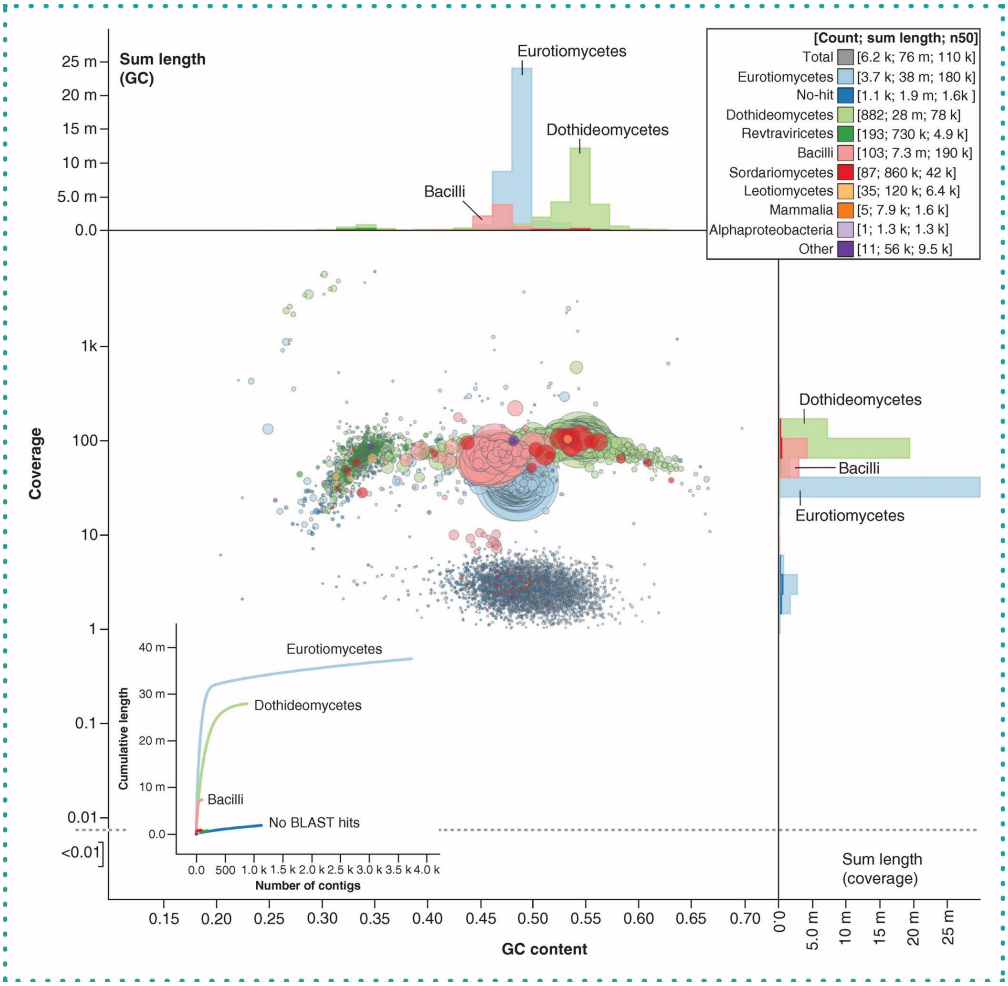


Figure 2. Taxon-annotated GC-coverage plot (BlobPlot) of CMW55930 Assembly 1.
Each circle represents a contig sequence, plotted relative to its base coverage and GC proportion. Circle diameter is proportional the size of the contig it represents. Circles are colored according to their assigned taxon at the Class level (see legend), using the 'bestsum' taxrule. Histograms show the distribution of the total assembly length along each axis. The inserted cumulative plot indicates the total length and number of contigs contributed by each taxon.

Based on HMM hits, anvi'o also estimated one bacterial and two eukaryotic genomes in Assembly 1. The genome sample was, therefore, contaminated by at least one *Penicillium* species, most likely *P. crustosum*, and a bacterium. All ribosomal sequences identified in anvi'o indicated the bacterium to be a *Paenibacillus* species. *P. crustosum* is a common food contaminant that occurs naturally in subglacial ice [43] and *Paenibacillus* bacteria have been isolated from laboratory surfaces [44]. Their growth in a poorly stored laboratory sample is, therefore, consistent with unintentional environmental introductions.

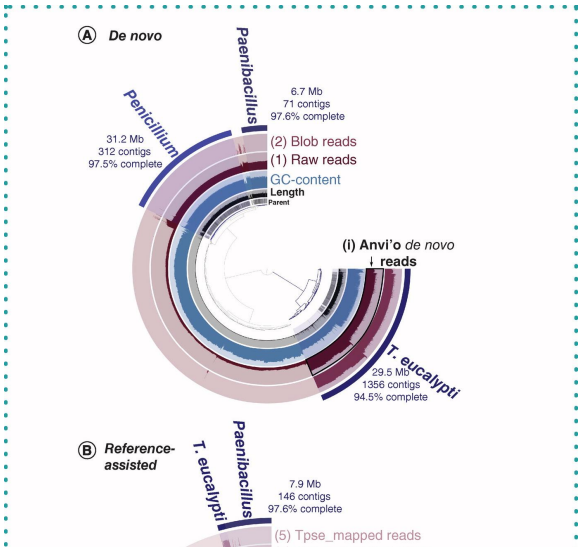
Read sets used to identify genome bins

The various read filtering methods suggested that approximately 60% of the raw reads originated from *T. eucalypti* (Table 2). The most stringent filtering method (alignment of already-filtered reads to *T. pseudoeucalypti*) retained 54.2% of the raw reads, whereas BlobToolKit-based filtering retained 60.8%. The *de novo* and reference-assisted approaches retained a similar number of raw reads (60–62%). Between the two approaches, 50 contigs of Assembly 1 were assigned differently, with an additional 32 contigs included in the reference-assisted approach.

Table 2. The number of raw and filtered reads in each read set used in this study. (Table view)

Read set	Number	Total (%)	Description
Isolate CMW55930			
(1) raw reads	13618098	100.00	Reads from the contaminated (mixed) genome sample
De novo filtering			
(2) blob reads	8273103	60.75	BlobToolKit taxonomy-filtered reads
Reference-assisted filtering			
(3) Pen_mapped reads	3045924	22.37	Reads that aligned to *Penicillium*
(4) Pen_unmapped reads	10572174	77.63	Reads that did not align to *Penicillium*
(5) Tpse_mapped reads	7375248	54.16	Reads from set 4 that aligned to *Teratosphaeria pseudoeucalypti*
Extracted from anvi'o			
(i) anvi'o *de novo*	8246161	60.55	Reads that mapped to the *T. eucalypti* bin
(ii) anvi'o reference-assisted	8576930	62.98	Reads that mapped to the T. eucalypti bin
Reads aligned to CMW54005	8618060	63.28	Reads that mapped to the uncontaminated *T. eucalypti* genome
Isolate CMW54005			
Raw reads	12092134	100.00	Reads from the uncontaminated genome sample

In the anvi'o dendograms of Assembly 1 (Figure 3), putative *T. eucalypti*, *Paenibacillus* and *Penicillium* bins were identified based on the tetra-nucleotide sequence composition of contigs and the coverage of each read set. The *T. eucalypti* bin comprised 38.6% (29.5 Mb) of Assembly 1 in the *de novo* approach (i) and 40.4% (30.9 Mb) in the reference-assisted approach (ii). *Penicillium* and *Paenibacillus* comprised 40–41% (30.9–31.2 Mb) and 9–10% (6.7–7.9 Mb), respectively. This was congruent with the 7.3 Mb Bacilli contigs and less than the 37.5 Mb Eurotiomycetes contigs identified by BlobToolkit. The remainder of the assembly consisted of the low coverage, short contigs also represented by the noisy data observed in Figure 2.



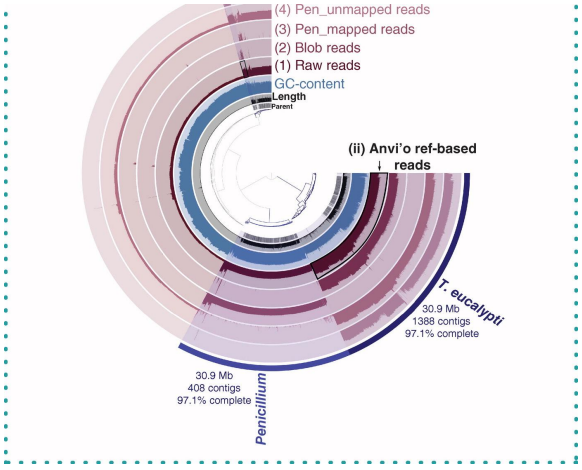


Figure 3. Anvi'o plots depicting the effect of the filtering process applied in the (A) *de novo* and (B) reference-assisted approaches.
In both plots, Assembly 1 is visualized as a circular dendogram clustered by tetra-nucleotide frequency and differential sequence coverage. Contigs >20,000 bp have been split as indicated by the innermost Parent layer. Subsequent layers depict the length, GC% and read coverage of these splits as histograms. **(A)** The *Teratosphaeria eucalypti* bin is identified by high read coverage in the BlobToolKit-filtered reads (set 2). The *Penicillium* and *Paenibacillus* (bacterial) bins were identified by high sequence coverage, different sequence compositions and the occurrence of bacterial single-copy genes. **(B)** The *T. eucalypti* bin has a high read coverage in all read sets apart from number 3. The raw read set (1) and read set 3 identify the *Penicillium* bin, whereas the *Paenibacillus* bin was identified by non-fungal reads present in set 4.

Teratosphaeria eucalypti metagenome-assembled genomes

De novo taxonomy-based filtering is conservative

The *de novo* approach (i) produced four iterations of a clean *T. eucalypti* CMW55930 assembly (Table 3 & Supplementary Figure 1). These consisted of two 'metagenome-assembled genomes' (MAG1-2) and their re-assemblies. A further three reference-assisted (ii) assemblies (MAG3, MAG3r and Tpse_filtered) were produced using reference genomes during the filtering stages. The *de novo* MAGs had similar BUSCO completeness (93–95%), slightly lower than that of the reference-assisted assemblies (>97%). For the fungal dataset, the completeness of the reference-assisted assemblies (>98%) was even higher than that of the CMW54005 reference-based assembly (97.0%).

Table 3. Filtered genome assemblies of *Teratosphaeria eucalypti* CMW55930. (Table view)

Assembly name	Metagenome-assembled genomes (MAGs)			Re-assembled MAGs			Assembled fro reads	
	MAG 1	MAG 2	MAG 3	MAG 1r	MAG 2r	MAG 3r	Tpse_filtered	R
Origin	BlobToolKit	anvi'o (*de novo*)	anvi'o (reference-based)	(2) blob reads	(i) anvi'o *de novo*	(ii) anvi'o reference-assisted	(5) Tpse_mapped reads	R a C
Raw reads used (%)	60.75	60.55	62.98	60.75	60.55	62.98	54.16	6
Size (Mb)	27.93	29.48	30.87	28.82	29.12	30.58	26.22	3
Contigs ≥1000 bp	882	1356	1388	1131	982	1016	2375	1
L50	103	113	120	107	104	107	276	1
N50 (kb)	78.19	76.91	76.81	77.32	84.84	81.71	25.82	6
Guanine-Cytosine (GC) content (%)	52.21	51.28	51.35	51.64	51.45	51.50	53.84	5

Assembly name	Metagenome-assembled genomes (MAGs)			Re-assembled MAGs			Assembled fro reads	
	MAG 1	MAG 2	MAG 3	MAG 1r	MAG 2r	MAG 3r	Tpse_filtered	R
Repeat content (%)	14.50	18.69	18.32	17.22	17.85	17.59	8.29	1
Comparison to reference								
– NG50 (kb)	74.58	74.58	77.37	74.77	83.18	84.44	21.30	6
– LG50	117	115	115	115	108	104	358	1
– Genome fraction (%)	88.13	91.46	95.56	89.85	90.90	95.10	84.32	9
– Mismatches/100 kbp (n)	191.35	235.85	233.14	237.22	232	231.29	161.73	2
– Total aligned length (Mb)	26.52	27.63	28.87	27.09	27.52	28.67	25.37	2
BUSCO results								
– Eukaryota: single copy	238 (93.33%)	238 (93.33%)	249 (97.65%)	238 (93.33%)	237 (92.94%)	249 (97.65%)	248 (97.25%)	2 (9
– Eukaryota: duplicate	0	0	0	0	0	0	0	0
– Fungi: single copy	716 (94.46%)	716 (94.46%)	748 (98.68%)	716 (94.46%)	715 (94.33%)	748 (98.68%)	749 (98.81%)	7 (9
– Fungi: duplicate	0	0	0	0	0	0	0	0

De novo taxonomy-based filtering of contigs in BlobToolKit produced the least complete *T. eucalypti* assemblies (MAG1 and 2). The authors applied the least conservative BlobToolKit 'taxrule' [16], yielding the most Dothideomycetes sequence classifications. The BlobPlot protocol was similarly reported to be more conservative than other approaches when identifying contigs of the nematode *Caenorhabditis remanei*, with a loss of some *C. remanei* sequence information [19].

A taxonomy-based approach relies on public databases and is likely to be less accurate as public sequence availability, contig length and taxonomic distance decrease [5,18]. For example, at taxonomic ranks below phylum, the taxonomies of the smaller Ascomycota contigs in Assembly 1 became increasingly divided so that less of the assembly was attributed to *T. eucalypti*. The loss of sequence information had a smaller effect on the *Penicillium* contigs, because large amounts of data are publicly available for this well-studied genus. Taxon-based filtering, however, proved immensely useful as a reference point for decontamination [18], despite limited *Teratosphaeria* sequence data.

The MAG2r, MAG3r and Tpse_filtered assemblies were re-assessed (Figure 4) and compared with the reference-based and reference assemblies. Approximately the same proportion of all assemblies (26–28 Mb) were classified as Dothideomycetes. The remaining *ca.* 5% 'contaminant' sequences identified by BLAST primarily originated from other Fungi, particularly Ascomycetes. While certain studies may need to consider whether sampling reagents have been contaminated [3], such contaminants would have insignificant coverage in single-genome projects. The consistent occurrence of apparently contaminant sequences in all *T. eucalypti* assemblies, and in other *Teratosphaeria* genomes

(Supplementary Figure 2), suggested that these were a result of incorrect taxonomic classification.

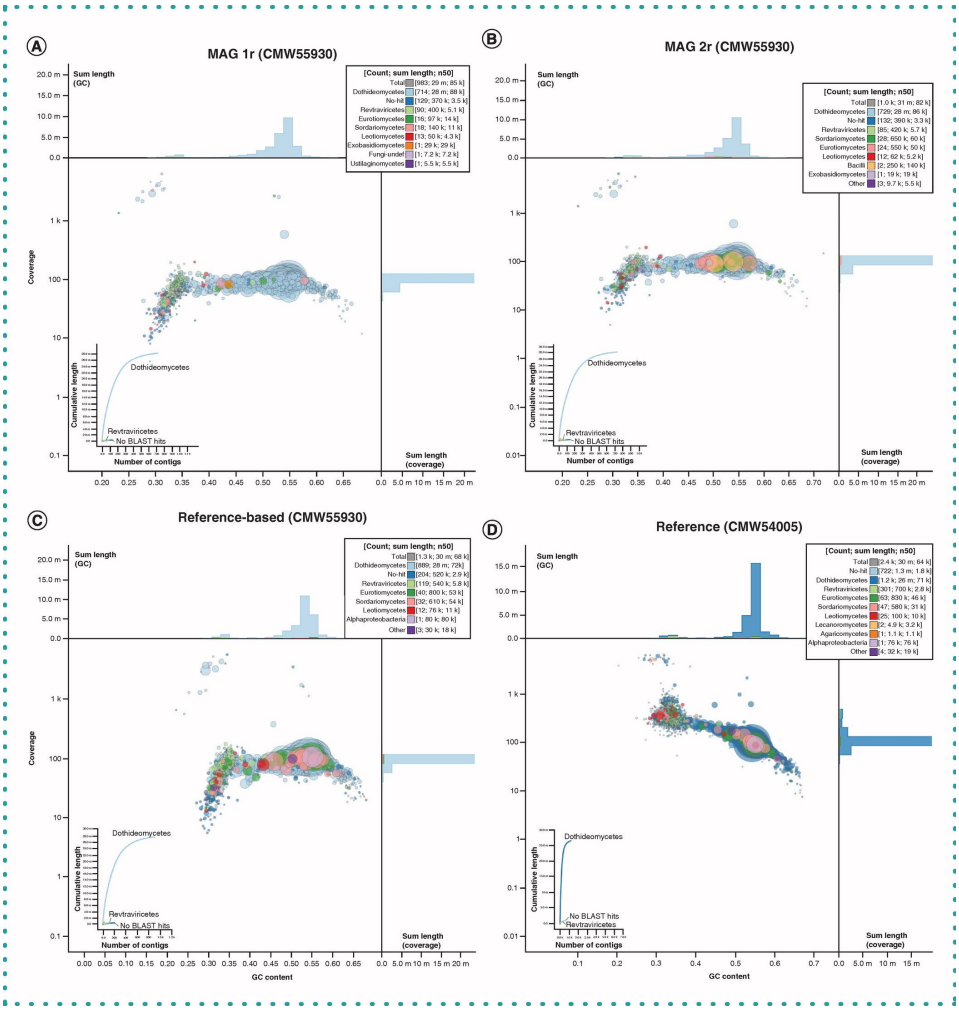


Figure 4. BlobPlots comparing the filtered and reference assemblies of *Teratosphaeria eucalypti*.
(A) CMW55930 MAG 2r represents the *de novo* re-assembly filtered without a reference genome. (B) MAG 3r was produced with reads filtered using the genomes of species closely related to the target and contaminant organisms, whereas (C) was filtered using the *T. eucalypti* CMW54005 reference genome (D).

In the reference genome, 8.0% (2.5 Mb) of sequences did not have a BLAST hit, which was more than double the amount of 'no-hits' in the filtered CMW55930 assemblies. Despite this, the repeat contents of the re-assembled MAGs, the reference-based assembly (Table 3) and the reference genome were similar (16–18%; Table 1). In contrast, the lower repeat content and high BUSCO completeness of the Tpse_filtered assembly indicated that its smaller genome size was due to a loss of repetitive sequences during the stringent read filtering process (Supplementary Figure 3).

A modified metagenomics approach

The reference-assisted anvi'o assemblies (MAG3 and MAG3r) appeared to be most similar to the reference *T. eucalypti* genome and MAG3r had the best genome statistics. Therefore, using the consensus of several different filtering methods to identify the different genome bins and extract target reads appeared to be most effective [20]. This mimicked the approach of a metagenomics study in which multiple samples are considered based on their sequence compositions and differential coverage [17]. The increased genome size, gene

complexity and taxonomic distribution of eukaryotic communities, however, have meant that few metagenomics studies consider this domain [11,23]. Subsequently, tools like anvi'o [17], that combine binning and taxonomic identification, have been developed for prokaryotes and lack automated functionality for eukaryotic projects.

The aims of metagenomic studies do not require MAG re-assembly, but the aim of this project was to produce a genome of the highest possible quality and completeness. Assembly algorithms might not be able to calculate the optimal single-organism assembly in a mixed sample and the initial mixed assembly may have contained chimeric contigs [13]. Re-assembly is thus an important step that improved the continuity, while slightly decreasing the repeat content, of the anvi'o MAGs. In contrast, although re-assembling the BlobToolKit MAG (MAG1) decreased continuity and increased repeat content, the re-assembled MAG more closely resembled the reference genome, further supporting the need for re-assembly.

Initially, the authors also distinguished putative eukaryotic and prokaryotic sequences, a technique that previously yielded high-quality eukaryotic genomes from metagenomic samples [23]. However, in the low-complexity mixed genome, this step overfiltered data that could be confidently distinguished based on other characteristics, such as taxonomy. Few of the *T. eucalypti*-assigned contigs were classified as prokaryotic and their removal resulted in poorer completeness estimates (Supplementary Table 3) and exclusion of sequences resembling *Teratosphaeria* mitochondrial genomes (Supplementary Table 4) [Kanzi, unpublished data]. The authors, therefore, conclude that EukRep [23] is best applied to complex samples, where MAG boundaries are less clearly defined.

The anvi'o plots of Assembly 1 (Figure 3) also provided insight into the structure of the *T. eucalypti* genome by indicating two main branches of tetra-nucleotide frequencies in the *T. eucalypti* MAG. Further investigation (Supplementary Figure 3) confirmed that these branches correspond to a division between highly repetitive and non-repetitive sequences, consistent with the bimodal GC% distribution observed for all *T. eucalypti* assemblies (Supplementary Figure 1A). In all investigated assemblies, a cluster of contigs with an unusually high coverage and low GC% was apparent (Supplementary Figure 3) and BLASTn confirmed that these represented mitochondrial sequences [Kanzi, unpublished data].

Conclusion

This study provided an overview of how BlobToolKit and anvi'o can be applied in complementary ways to decontaminate a mixed assembly, both in the presence and in the absence of closely related reference sequences. First, the statistics and completeness scores of the initial assembly must be scrutinized, while considering that BUSCO does not work equally well across the eukaryotic tree of life [11]. BUSCO is, however, well developed for fungal lineages and BLAST provides an easy and rapid method to identify the origin of fungal BUSCOs. A further 'first overview' is provided by BlobToolKit [16].

Once an overview of the initial contaminated assembly has been generated, the assembly and its reads should be filtered in various ways. BlobToolKit provides the first taxonomy-based screening step, without the need to identify available reference genomes. Applying only BlobToolKit as a filter already yields a high-quality genome sequence, though perhaps falsely excluding some coding sequences [19], leading to lower completeness. Further manual read filtering can be done by aligning the raw reads to genomes of the same or closely related species. In this study, we used reference genomes related to both

the contaminant and target species for filtering, and anvi'o provided excellent visualization of the impact of all filtering methods and, subsequently, of the boundaries of our target genome.

Based on the differential coverage in our filtered read sets, manual identification of MAGs in anvi'o was most effective. Automated binning programs, such as MetaBAT [33] and CONCOCT [32] applied here, could also be used and compared with the manual process, choosing the result that makes the most biological sense. Finally, the target MAG and its associated reads can be extracted, re-assembled and re-evaluated. Visualizing the cleaned genome in BlobToolKit and anvi'o has the added benefit of providing some insight into genome structure.

Future perspective

Sequencing pure DNA samples should be the goal of all single-organism genome sequencing projects. Obtaining such samples, however, is not always biologically possible or contaminants may be introduced unintentionally [3,8]. This study supplements a growing body of research [8,19,20] that illustrates why it is essential to assess the purity of single-organism genome assemblies. To the best of our knowledge, this is the first example of removing a fungal contaminant from a fungal assembly.

BlobToolKit plots sequence coverage and GC content, incorporating taxonomic annotation. We contend that such screening is essential for any genome project, regardless of whether contamination is suspected. We, consequently, recommend that all genome sequencing projects routinely create taxon-annotated GC-coverage plots. Similarly, visualization in anvi'o is immensely powerful, not only to detect different organisms but also to reveal different genome compartments within a single organism. Although the full functionality of this tool cannot yet be exploited for eukaryotic projects, anvi'o already benefits eukaryotes [20], as also illustrated in this study.

Executive summary

Background

- The authors illustrated the use of two available software programs, BlobToolKit and anvi'o, to resolve a fungal genome from a contaminated genome sample.

Experimental

- A *de novo* and a reference-assisted approach were implemented to identify different organisms within the mixed assembly.
- Both approaches followed that of a low-complexity metagenomics project and used various techniques to filter the raw reads and identify genome bins.
- Efficacy of the filtering process was evaluated through comparison with an uncontaminated genome sequence of the target organism.

Results & discussion

- BlobToolKit provides a taxonomy-based filtering method without the need to identify other available sequence data, but it tends to be conservative.
- The effect of various *de novo* and reference-assisted filtering techniques can be visualized in anvi'o, and a consensus of these results can be extracted to produce a filtered assembly.

- Using reference sequences for filtering is not essential but improves the completeness of the final filtered genome assembly.

Conclusion

- A high-quality draft fungal assembly can be obtained from an inadvertently mixed genome sample.
- Visualizing assemblies should be routine in all genome sequencing projects.

Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.future-science.com/doi/suppl/10.2144/btn-2021-0097

Author contributions

J Aylward conceptualized and designed the study, acquired and analyzed the data and wrote the manuscript. BD Wingfield and MJ Wingfield substantially contributed to the concept, data acquisition and analysis and revised the manuscript. F Roets substantially contributed to data acquisition and revised the manuscript. All authors have approved the final version of this manuscript and agree to the authorship statement in the included author disclosure form.

Acknowledgments

The authors thank A Kanzi for supplying unpublished mitochondrial sequences of *Teratosphaeria* species for validation of results.

Financial & competing interests disclosure

We thank members of the Tree Protection Co-operative Program (TPCP), the Department of Science and Technology (DST) – National Research Foundation (NRF) Centre of Excellence in Plant Health Biotechnology (CPHB) and the SARChI chair in Fungal Genomics for financial support. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

References

Papers of special note have been highlighted as: • of interest; •• of considerable interest

1. Breitwieser FP, Pertea M, Zimin AV, Salzberg SL. Human contamination in bacterial genomes has created thousands of spurious proteins. *Genome Res.* 29(6), 954–960 (2019).
2. Kryukov K, Imanishi T. Human contamination in public genome assemblies. *PLoS One* 11(9), e0162424 (2016).
3. Salter SJ, Cox MJ, Turek EM et al. Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* 12(1), 87 (2014).
4. Samson CA, Whitford W, Snell RG, Jacobsen JC, Lehnert K. Contaminating DNA in human saliva alters the detection of variants from whole genome sequencing. *Sci. Rep.* 10(1), 19255 (2020).
5. Steinegger M, Salzberg SL. Terminating contamination: large-scale search identifies more than 2,000,000 contaminated entries in GenBank. *Genome Biol.* 21(1), 115 (2020).

6. Reveillaud J, Bordenstein SR, Cruaud C et al. The *Wolbachia* mobilome in *Culex pipiens* includes a putative plasmid. *Nat. Commun.* 10(1), 1051 (2019).
7. Lu J, Salzberg SL. Removing contaminants from databases of draft genomes. *PLoS Comput. Biol.* 14(6), e1006277 (2018).
8. Koutsovoulos G, Kumar S, Laetsch DR et al. No evidence for extensive horizontal gene transfer in the genome of the tardigrade *Hypsibius dujardini*. *Proc. Natl Acad. Sci.* 113(18), 5053–5058 (2016).
• **Uses BlobToolKit to identify bacterial contamination of a tardigrade genome and disprove previous claims of horizontal gene transfer.**
9. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13(5), 329–342 (2012).
10. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31(19), 3210–3212 (2015).
11. Saary P, Mitchell AL, Finn RD. Estimating the quality of eukaryotic genomes recovered from metagenomic analysis with EukCC. *Genome Biol.* 21(1), 244 (2020).
• **Shows that BUSCO proteins are not necessarily an accurate measure of completeness in all eukaryotic lineages.**
12. Jauhal AA, Newcomb RD. Assessing genome assembly quality prior to downstream analysis: N50 versus BUSCO. *Mol. Ecol. Resour.* 21(5), 1416–1421 (2021).
13. Kumar S, Jones M, Koutsovoulos G, Clarke M, Blaxter M. Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots. *Front. Genet.* 4, 237 (2013).
14. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res.* 25(7), 1043–1055 (2015).
15. Mallet L, Bitard-Feildel T, Cerutti F, Chiapello H. PhylOligo: a package to identify contaminant or untargeted organism sequences in genome assemblies. *Bioinformatics* 33(20), 3283–3285 (2017).
16. Challis R, Richards E, Rajan J, Cochrane G, Blaxter M. BlobToolKit – interactive quality assessment of genome assemblies. *G3* 10(4), 1361–1374 (2020).
•• **Details the premise and implementation of the BlobToolKit pipeline.**
17. Eren AM, Esen ÖC, Quince C et al. Anvi'o: an advanced analysis and visualization platform for 'omics data'. *PeerJ* 3, e1319 (2015).
• **Details the premise and implementation of anvi'o.**
18. Laetsch D, Blaxter M. BlobTools: interrogation of genome assemblies. *F1000Research* 6, 1287 (2017).
19. Fierst JL, Murdock DA. Decontaminating eukaryotic genome assemblies with machine learning. *BMC Bioinform.* 18(1), 533 (2017).
• **Describes and applies different predictor variables to separate target and contaminant sequences in a *de novo* manner.**
20. Delmont TO, Eren AM. Identifying contamination with advanced visualization and analysis practices: metagenomic approaches for eukaryotic genome assemblies. *PeerJ* 4, e1839 (2016).
• **Uses anvi'o to identify bacterial contaminants in a tardigrade genome.**
21. Chiapello H, Mallet L, Guérin C et al. Deciphering genome content and evolutionary relationships of isolates from the fungus *Magnaporthe oryzae* attacking different host plants. *Genome Biol. Evol.* 7(10), 2896–2912 (2015).

22. Merchant S, Wood DE, Salzberg SL. Unexpected cross-species contamination in genome sequencing projects. *PeerJ* 2, e675 (2014).

23. West PT, Probst AJ, Grigoriev IV, Thomas BC, Banfield JF. Genome-reconstruction for eukaryotes from complex natural microbial communities. *Genome Res.* 28(4), 569–580 (2018).

24. Andjic V, Carnegie AJ, Pegg GS et al. 23 years of research on *Teratosphaeria* leaf blight of *Eucalyptus*. *For. Ecol. Manag.* 443, 19–27 (2019).

25. Andjic V, Pegg GS, Carnegie AJ, Callister A, Hardy GE, Burgess TI. *Teratosphaeria pseudoeucalypti*, new cryptic species responsible for leaf blight of *Eucalyptus* in subtropical and tropical Australia. *Plant Pathol.* 59(5), 900–912 (2010).

26. Havenga M, Wingfield BD, Wingfield MJ et al. Genetic recombination in *Teratosphaeria destructans* causing a new disease outbreak in Malaysia. *For. Pathol.* 51(3), e12683 (2021).

27. Havenga M, Wingfield BD, Wingfield MJ et al. Mating strategy and mating type distribution in six global populations of the *Eucalyptus* foliar pathogen *Teratosphaeria destructans*. *Fungal Genet. Biol.* 137, 103350 (2020).

28. Bankevich A, Nurk S, Antipov D et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J. Comput. Biol.* 19(5), 455–477 (2012).

29. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9(4), 357 (2012).

30. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. *J. Mol. Biol.* 215(3), 403–410 (1990).

31. Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat. Methods* 18(4), 366–368 (2021).

32. Alneberg J, Bjarnason BS, De Bruijn I et al. Binning metagenomic contigs by coverage and composition. *Nat. Methods* 11(11), 1144–1146 (2014).

33. Kang DD, Li F, Kirton E et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* 7, e7359 (2019).

34. Hyatt D, Chen G-L, Locascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform.* 11(1), 119 (2010).

35. Li H, Handsaker B, Wysoker A et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25(16), 2078–2079 (2009).

36. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0. http://www.repeatmasker.org

37. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. *Bioinformatics* 29(8), 1072–1075 (2013).

38. Wilken PM, Aylward J, Chand R et al. IMA Genome-F13: draft genome sequences of *Ambrosiella cleistominuta*, *Cercospora brassicicola*, *C. citrullina*, *Physcia stellaris*, and *Teratosphaeria pseudoeucalypti*. *IMA Fungus* 11(1), 1–17 (2020).

39. Wingfield BD, Fourie A, Simpson MC et al. IMA Genome-F 11 draft genome sequences of *Fusarium xylarioides*, *Teratosphaeria gauchensis* and *T. zuluensis* and genome annotation for *Ceratocystis fimbriata*. *IMA Fungus* 10, 13 (2019).

40. Wingfield BD, Liu M, Nguyen HD et al. Nine draft genome sequences of *Claviceps purpurea s. lat.*, including *C. arundinis*, *C. humidiphila*, and *C. cf. spartinae*, pseudomolecules for the pitch canker pathogen *Fusarium circinatum*, draft genome of *Davidsoniella eucalypti*, *Grosmannia galeiformis*, *Quambalaria eucalypti*, and *Teratosphaeria destructans*. *IMA Fungus* 9, 401 (2018).

41. Haridas S, Albert R, Binder M et al. 101 Dothideomycetes genomes: a test case for predicting lifestyles and emergence of pathogens. *Stud. Mycol.* 96, 141–153 (2020).

42. Duong TA, Aylward J, Ametrano CG et al. IMA Genome-F15: draft genome assembly of *Fusarium pilosicola*, *Meredithiella fracta, Niebla homalea, Pyrenophora teres* hybrid WAC10721, and *Teratosphaeria viscida*. *IMA Fungus* 12, 30 (2021).

43. Sonjak S, Frisvad JC, Gunde-Cimerman N. *Penicillium* mycobiota in Arctic subglacial ice. *Microb. Ecol.* 52(2), 207–216 (2006).

44. Sáez-Nieto JA, Medina-Pascual MJ, Carrasco G et al. *Paenibacillus* spp. isolated from human and environmental samples in Spain: detection of 11 new species. *New Microbes New Infect.* 19, 19–27 (2017).