

A practical approach to genome assembly and annotation of Basidiomycota using the example of *Armillaria*

Deborah L Narh Mensah^{*,1,2} , Brenda D Wingfield¹  & Martin PA Coetzee¹ 

¹Department of Biochemistry, Genetics & Microbiology, Forestry & Agricultural Biotechnology Institute (FABI), Faculty of Natural & Agricultural Sciences, University of Pretoria, Pretoria, Gauteng, South Africa; ²Council for Scientific and Industrial Research – Food Research Institute (CSIR-FRI), PO Box M20, Accra, Ghana; *Author for correspondence: deborah.narhmensah@fabi.up.ac.za

BioTechniques 75: 115–128 (September 2023) 10.2144/btn-2023-0023

First draft submitted: 17 March 2023; Accepted for publication: 17 August 2023; Published online: 8 September 2023

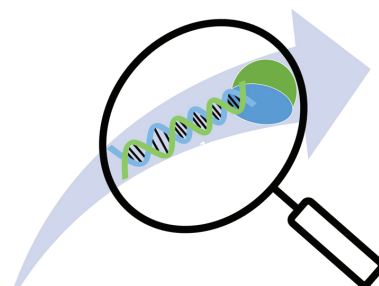
ABSTRACT

Technological advancements in genome sequencing, assembly and annotation platforms and algorithms that resulted in several genomic studies have created an opportunity to further our understanding of the biology of phytopathogens, including *Armillaria* species. Most *Armillaria* species are facultative necrotrophs that cause root- and stem-rot, usually on woody plants, significantly impacting agriculture and forestry worldwide. Genome sequencing, assembly and annotation in terms of samples used and methods applied in *Armillaria* genome projects are evaluated in this review. Infographic guidelines and a database of resources to facilitate future *Armillaria* genome projects were developed. Knowledge gained from genomic studies of *Armillaria* species is summarized and prospects for further research are provided. This guide can be applied to other diploid and dikaryotic fungal genomics.

TWEETABLE ABSTRACT

We reviewed methods used in *Armillaria* (Fungi) genome projects. Infographic guidelines and a resource database were developed to facilitate future genome projects. This guide can be applied to other diploid and dikaryotic genomics.

GRAPHICAL ABSTRACT



Steps, tools and guidelines for genome sequencing and assembly

Genome annotation tools and techniques

Database of useful resources

Knowledge gained from genomics and transcriptomics

Suggestions for future genomic and transcriptomic studies

Future perspectives for genome sequencing and annotation

KEYWORDS:

bioinformatics • DNA extraction • evolution • genomics • metabolomics • next-generation sequencing • phylogenomics • phytopathogens • proteomics • transcriptomics

Armillaria includes about 70 known species [1] that are distributed worldwide, as reviewed by Baumgartner *et al.* and Coetzee *et al.* [2,3]. Two species, *Armillaria ectypa* and *Armillaria tabescens*, that were previously placed in *Armillaria*, now reside in *Desarmillaria* based on absence of annulus at maturity and an evolutionary trajectory separated from the rest of the *Armillaria* species [4]. *Armillaria* spp. are ubiquitous, have a broad host range, parasitize both herbaceous and woody plants and survive under diverse environmental conditions, as reviewed by Heinzelmann *et al.* [5]. The ubiquity of *Armillaria* spp. has been attributed to their life cycles, lifestyles and other biotic factors. We refer the reader to previous reviews for further information regarding the ecology, life cycles, lifestyles, species boundaries and geographical distribution and other properties of *Armillaria* spp [2,3,5].

Armillaria spp. are of significant economic and agroecological importance due to various characteristics they exhibit. When they are parasitic, they cause massive devastation in horticulture and agriculture, as well as in natural and managed plantation forests [6–9]. However, *Armillaria* spp. enhance the forest ecosystem when they act as saprophytes [10]. The forest ecosystem is enriched when saprobic microorganisms including *Armillaria* spp. degrade plant biomass with their extensive enzyme systems, resulting in significant contribution in releasing sequestered organic carbon into the carbon cycle [11,12]. Moreover, like other saprobic Basidiomycota species, some species in the genus are also edible. Hence there is interest in domesticating species such as *Armillaria mellea* for culinary purposes, as reviewed by Ren *et al.* [13]. Additionally, the enzymes, secondary metabolites and other compounds produced by *Armillaria* spp. have the potential to be used in various industries such as the pharmaceutical industry. Therefore research on *Armillaria* spp. has focused on understanding the growth mechanisms in terms of *in situ* and *in vitro* mycelia, rhizomorph and fruit body growth, as well as pharmacological and other properties [13,14]. Other research on these fungi has focused on mechanisms which drive their pathogenicity and can be exploited and/or harnessed to halt their growth (in the case of pathogenicity of socioeconomically important plants) or enhance their production of enzymes and other compounds of interest for biotechnological applications [13,14]. Despite the research milestones made to better understand and limit or utilize these mushroom-forming fungi, there are still large knowledge gaps. Closing of these gaps is relevant as the current management practices have proven to be mostly inefficient in managing *Armillaria* root-rot disease [5,13].

Several factors facilitate survival and pathogenicity and/or virulence of *Armillaria* spp. To understand these mechanisms and to detect possible applications, genomics-based research has been conducted by various research groups on *Armillaria* spp. [11,15–22]. These studies have given some insights into various characteristics of *Armillaria* spp., mostly from samples collected from countries in the northern hemisphere. In the present review, the extent of genomic studies on *Armillaria* species is considered, with a focus on the diversity of species and strains for which genomes have been assembled and annotated. The steps, techniques and tools used for genome extraction, sequencing, assembly, evaluation and annotation are discussed. Infographic guidelines for future *Armillaria* genome projects with some relevant resources have also been provided. Knowledge gained from these genome projects is also highlighted. We further suggest potential areas for future *Armillaria* genome projects as well as potential comparative genomic and multi-omic studies. Although this review centers around genomic studies of *Armillaria* spp., the methods and information gathered from these studies can be extended to genome research on other fungi.

Status of genome sequencing & assembly of *Armillaria* spp.

There are presently (March 2023) 11 published whole genome sequences of *Armillaria* spp. in the National Center for Biotechnology Information's (NCBI) Genome [23] and US Department of Energy's Joint Genome Institute (JGI) fungal genomics resource MycoCosm databases (Table 1) [24]. Other genomes have been submitted to the JGI MycoCosm database but have not yet been published in journals and require permissions from the principal investigators for use. This review will therefore focus on the published genomes of *Armillaria* spp. only and the knowledge gained from their genomic studies. For concision, we refer the reader to links and references provided for all the tools discussed in this review in Supplementary Table 1.

Starting materials & DNA extraction methods used in *Armillaria* genome projects

The starting material can influence the quantity and quality of DNA extracted for downstream processes. The quality of DNA extracted, in terms of chemical purity, structural integrity and size of the DNA, influences the choice of sequencing technology that can be employed in a genome project. These parameters also affect the quality of obtained data for downstream processes, including genome assembly [27]. *Armillaria* genome projects have typically used fresh mycelia, mostly obtained as liquid cultures in different media, as the starting materials (Table 1) [16–22]. For the *Armillaria borealis* strain AB13-TR4-IP16 genome project, fresh mycelium isolated from under the bark of infested *Abies sibirica* stems 50 cm above the soil surface was used as starting material [21].

Relatively low-molecular-weight DNA will be sufficient for sequencing using second-generation sequencing (SGS) technologies (also known as short reads; e.g., Illumina, 454, SOLiD, Ion Torrent). High-molecular-weight DNA in sufficient quantity and quality is required for third-generation sequencing (TGS, also known as long reads; e.g., Pacific Biosciences [PacBio], Oxford Nanopore Technologies). Typically, the minimum amounts of DNA required for Illumina, PacBio, Oxford Nanopore and BioNano are approximately 3 ng, 20 µg, 1 µg

Table 1. Methods and tools used in sequencing and assembly of published genomes of *Armillaria* species to date (March 2023).

Species [strain or isolate]	Country of origin (lifestyle) [†]	DNA extraction methods/kits	Sequencing technologies (pre-assembly reads correction)	Genome assembly tools	Post-assembly processing	Genome completeness evaluation	NCBI accession number or genome link	Ref.
<i>Armillaria altimontana</i> (NABS X) [837–10]	USA (unknown)	MoBio (Qiagen) DNeasy PowerMax Soil Kit	PacBio SMRT cells (NA)	Pacific Biosciences SMRT Portal software	NA	QUAST, BUSCO 2.0b2 with Basidiomycota dataset	JAIWYR0000000000.1	[19]
<i>Armillaria borealis</i> [AB13-TR4-IP16]	Eastern Siberia (FN, weakly pathogenic, opportunistic)	Modified version of the hot-CTAB extraction at 65°C with chloroform washing	Illumina MiSeq (NA)	SPAdes v. 3.13.0 genome assembler	NA	BUSCO and RS BridgeMapper protocol	JAAGUC0000000000.1	[21]
<i>Armillaria cepistipes</i> [B5]	Italy (FN, weakly pathogenic)	PowerMax MOBIO DNA isolation kit	PacBio RS II platform; Illumina HiSeq 2000 (NA)	HGAP3 workflow of the SMRT Analysis suite v. 2.3	PBJelly2, FinisherSC, RS resequencing protocol and Pilon	CEGMA and BUSCO v. 3.0.1 with the lineage-specific profile library basidiomycota_odb9 and RS resequencing protocol	FTRY0000000000.1	[16]
<i>Armillaria fuscipes</i> [CMW 2740]	South Africa (FN, particularly pathogenic to exotic species)	DNeasy Plant Mini Kit (Qiagen)	Illumina HiSeq; PacBio sequencing with 20 SMRT cells (NA)	Celera Assembler pipeline, CLC Genomics Workbench v. 5.5.1 and MIRA	Velvet optimizer, Graph Accordance Assembly program and SSPACE	CEGMA and BUSCO using gene models from <i>Laccaria bicolor</i>	LWUHU0000000000.1	[18]
<i>Armillaria gallica</i> [012 m]	China (FN, weakly or secondarily pathogenic)	DNeasy plant kit (Qiagen)	PacBio Sequel; Illumina HiSeq (LoRDEC v.0.6, Trimmomatic v.0.36)	ALLPATHS-LG V52488 and MECAT2	Purge Haplotigs, FinisherSC v. 2.1 and NextPolish v. 1.0.21	BUSCO with the lineage-specific profile library basidiomycota_odb9	VTST0000000000.1	[20]
<i>Armillaria gallica</i> [Ar21-2]	USA (FN, weakly or secondarily pathogenic)	PowerMax MOBIO DNA isolation kit	Illumina MiSeq (NA)	HGAP3 workflow of the SMRT Analysis suite v. 2.3	PBJelly2, FinisherSC, RS Resequencing protocol and Pilon	CEGMA and BUSCO v. 3.0.1 with the lineage-specific profile library basidiomycota_odb9 and RS BridgeMapper protocol	NKEW0000000000.1	[16]
<i>Armillaria mellea</i> [DSM 3731]	France (FN, highly pathogenic)	ZR Fungal/Bacterial DNA Kit using a modified protocol	Illumina GAIi (NA)	Velvet v. 0.7.53	NA	NA	https://genome.jgi.doe.gov/portal/Armmel1.1/Armmel1_1.download.html	[17]
<i>Armillaria ostoyae</i> [C18/9]	Switzerland (pathogenic)	PowerMax MOBIO DNA isolation kit	PacBio RS II platform; Illumina HiSeq 2000 (NA)	HGAP3 workflow of the SMRT Analysis suite v. 2.3	PBJelly2, FinisherSC, RS Resequencing protocol and Pilon	CEGMA and BUSCO v. 3.0.1 with the lineage specific profile library basidiomycota_odb9 and RS BridgeMapper protocol	FUEG0000000000.1	[16]
<i>Armillaria solidipes</i> [C28-4]	USA (FN, highly pathogenic)	PowerMax MOBIO DNA isolation kit	Illumina HiSeq 2000 (NA)	HGAP3 workflow of the SMRT Analysis suite v. 2.3	PBJelly2, FinisherSC, RS Resequencing protocol and Pilon	CEGMA and BUSCO v. 3.0.1 with the lineage specific profile library basidiomycota_odb9 and RS BridgeMapper protocol	NKHM0000000000.1	[16]
<i>Armillaria solidipes</i> (form <i>A. ostoyae</i>) [ID001]	USA (FN, highly pathogenic)	MOBio DNeasy PowerMax Soil Kit	PacBio SMRT cells (NA)	Pacific Biosciences SMRT Portal software	NA	QUAST and BUSCO 2.0b2 with Basidiomycota dataset	JAIWYQ0000000000.1	[19]
<i>Armillaria</i> African Clade B sp. [CMW4456]	Zimbabwe (unknown)	Qiagen DNeasy Plant Pro Kit (50)	PacBio HiFi; Illumina HiSeq (Trimmomatic)	CLC Genomics Workbench v. 22.0.1	Pilon v. 1.23	QUAST v. 5.0.2 and BUSCO v. 5.3.2, using the agaricales_odb10 lineage dataset in the Galaxy platform	JANDKJ0000000000.1	[22]

[†] Lifestyle information for *A. ostoyae* was obtained from [25,26]. Lifestyle information for other species was summarized from [4].

CTAB: Cetyltrimethylammonium bromide; FN: Facultative necrotroph; NA: Not available; NCBI: National Center for Biotechnology Information.

and >200 ng, respectively [28]. High-molecular-weight DNA is typically obtained from fresh material [27]. Other considerations for DNA quality requirements for *de novo* sequencing have been discussed by Del Angel *et al.* [27].

DNA extraction for genome projects can be achieved using various protocols or DNA extraction kits. The cetyltrimethylammonium bromide method and modifications of it have been used to extract DNA from fungal and plant samples, which characteristically contain high concentrations of impurities (RNA, polysaccharides, proteoglycans, proteins, secondary metabolites, polyphenols and other contaminants) [29–33]. Akulova *et al.* used a modified cetyltrimethylammonium bromide method for extracting DNA from *A. borealis* strain AB13-TR4-IP16 [21]. Commercial DNA extraction kits have also been used in *Armillaria* genome projects, either according to the manufacturer's instructions or with some modifications [16–20,22]. These kits produce high-quality DNA extracts but may differ in their efficiency in terms of the quality and quantity of DNA extracted in relation to the specific organism [34,35]. Relatively high concentrations of some compounds considered to be impurities in DNA extracted for genome sequencing – including polysaccharides, proteins, secondary metabolites and polyphenols – have been reported in liquid cultures and fruiting bodies of *A. mellea* and *Armillaria gallica* grown under different conditions and extracted with different solvents [36–39]. For this reason, DNA extraction protocols or kits adopted for use in *Armillaria* genome projects should aim at reducing or eliminating these impurities to obtain high chemical purity, while maintaining the structural integrity and size of the DNA.

Choice of genome sequencing technologies

Armillaria genome projects have employed both SGS (i.e., Illumina HiSeq and Illumina MiSeq) and TGS (i.e., PacBio RS and PacBio SMRT) technologies (Table 1), similar to other fungal genome sequencing projects. These technologies have their respective strengths and limitations, as reviewed by several authors [27,40–42]. Although the SGS technologies are relatively cheaper, the generation of long contiguous DNA sequence reads during genome assembly is often impaired due to short read lengths, as well as the presence of repeat elements and polymorphisms [40,41]. Genome assemblies from reads obtained from SGS technologies also result in the loss of as much as 99.1% of validated duplicated sequences [38]. These challenges are largely circumvented by TGS, which provides long reads for better representation of the genome being assembled, although high rates (5–20%) of sequencing error of raw reads are obtained with this technology [40–42]. However, the technologies are being improved to reduce these error rates.

Five of the *Armillaria* genome projects utilized both SGS and TGS technologies for hybrid *de novo* genome assemblies (Table 1). These were *Armillaria cepistipes* strain B5, *Armillaria fuscipes* strain CMW 2740, *A. gallica* strain 012 m, *Armillaria ostoyae* strain C18/9 and *Armillaria* African Clade B sp. strain CMW4456. Hybrid approaches were followed to minimize the effects of individual use of SGS and TGS technologies on the quality and contiguity of the assembled genomes. Four of these genome projects used the Illumina reads for error correction or polishing after the PacBio read assembly [16,18,20], while the reverse approach was used for assembling the *A. fuscipes* strain CMW 2740 genome [22]. This hybrid genome assembly method has been used in other fungal genome projects such as the *Fusarium equiseti* genome project [43] and is the most frequently used combination of techniques in bacterial RefSeq genome projects [44]. Another combination of genome assembly techniques is the use of Illumina and MinION™ reads [44,45]. MinION is another Oxford Nanopore TGS technology [46]. This combination has not yet been used in *Armillaria* genome projects and might be useful for assembling good-quality genomes in future.

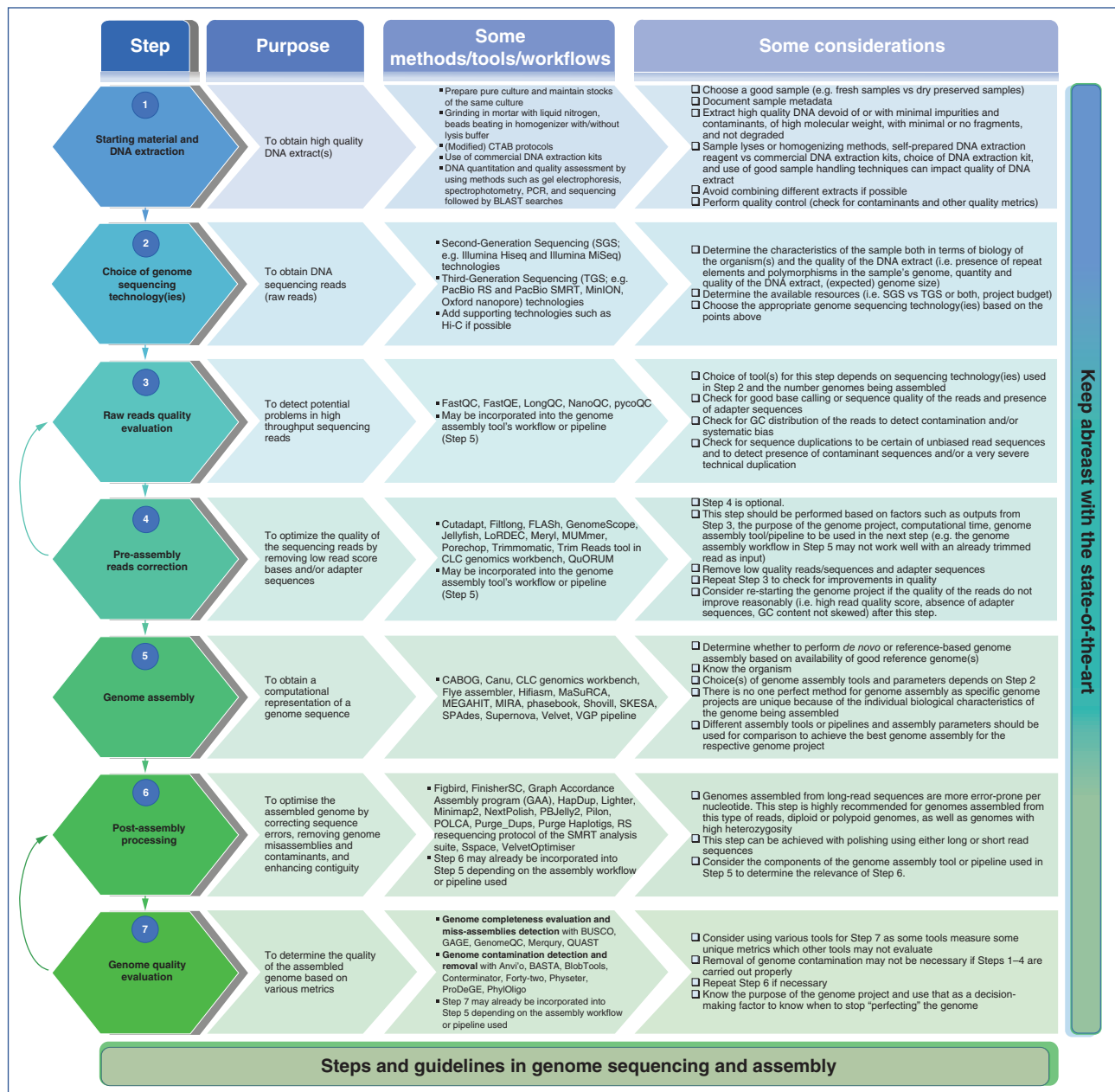
Raw reads quality control

Quality control (QC) pre-assembly is highly important as this can impact the quality of the final assembled genome. Methods or bioinformatic tools used for the QC steps in the published *Armillaria* genome projects were usually not reported. Nonetheless, various tools available for this purpose include FastQC, FastQE, Falco (FastQC Alternative Code), NanoPlot, PycoQC, LongQC and MultiQC (Figure 1 & Supplementary Table 1). These tools assess quality metrics including base or sequence quality of the reads, presence of adapter sequences and GC distribution. The choice of the tools used for raw-reads QC is usually informed by the sequencing technology used and the number of genomes being assembled. Additionally, raw-reads QC is sometimes integrated into genome assembly pipelines. As such, the decision of whether to perform a separate QC step should consider the genome assembly tool or pipeline to be used in the next steps.

Pre-assembly reads correction

Pre-assembly correction of raw reads is optional in genome projects because this step depends on the quality of the raw reads obtained, the purpose of the genome project and computational time, as well as the choice of genome assembly tool or workflow in the next step [27]. Some genome assembly tools or workflows incorporate reads correction; hence a separate reads correction is not needed in this case.

Tools such as LoRDEC and Trimmomatic have been used for reads correction in the *Armillaria* genome projects (Table 1). LoRDEC is mostly intended to correct PacBio reads, using Illumina reads as reference. Trimmomatic performs trimming of adapters and other Illumina-specific sequences, as well as quality and length trimming for Illumina paired-end and single-end reads. The order of the steps used during reads correction with Trimmomatic is important and depends on the quality of the raw sequence data. Other available tools for reads profiling and correction include the Trim Reads function in CLC Genomics Workbench (Qiagen, Aarhus, Denmark), Cutadapt, Filtlong and Porechop (Figure 1 & Supplementary Table 1). These tools also perform adapter trimming, quality trimming and length



Keep abreast with the state-of-the-art

Figure 1. Workflow, technological tools and decision-making considerations for genome sequencing and assembly.

trimming for reads obtained from SGS, TGS or both. The choice of which tool(s) to use depends on the sequencing technology used, the genome assembly pipeline and other considerations, some of which are outlined in Figure 1.

Genome assembly tools, pipelines or workflows

De novo genome assembly using various genome assembly tools has been used as opposed to reference-based assembly in *Armillaria* genome projects (Table 1). Genome assembly tools that were used include CLC Genomics Workbench v. 22.0.1, SPAdes, Velvet, MIRA, ALLPATHS-LG, MECAT2, PacBio SMRT Portal software and the now obsolete Celera Assembly Pipeline (Table 1). These tools have their respective strengths and limitations. For the specific functions and considerations for selecting and working with specific genome assembly tools, pipelines or workflows, see the links, references and workflow provided in Figure 1 & Supplementary Table 1.

De novo genome assembly is impacted by the computational methods used due to constant technological advancement and specific limitations of respective genome assemblers. The nature and quantity of sequence data, which rely on the features of the genome

being assembled, also impact *de novo* genome assembly [47,48]. Differences in various genome assemblers – in terms of memory consumption, total assembling time and accuracy based on genome quality matrices (contig number, N50 values and genome fractions) using prokaryotic and eukaryotic Illumina paired- and single-end reads – have been evaluated by Khan *et al.* [49]. The differences of the various assemblers and the unique biological characteristics of the genome being assembled require that correct assembler versions should be selected, and various assemblers and assembly parameters should be evaluated in genome projects. Outputs of the various assemblies should be compared in each genome project to determine the optimum assembly for the specific genome. Nevertheless, knowing when to stop while chasing perfection in a genome project is vital. It is essential to consider the complexity of the genome to be assembled (i.e., genome size, amount and distribution of repeat sequences, heterozygosity, ploidy and homogeneity of GC content), the research questions and available resources (e.g., availability of a reference genome, project budget, time and computational constraints) in the decision-making process [27,50].

Post-assembly processing

Post-assembly processing involves sequence error correction, polishing and haplotype phasing. Assembly errors are introduced especially when long reads are assembled. Hence post-assembly error correction and polishing is highly recommended to improve the quality of the overall local base accuracy and resolve some misassemblies of the assembled genome. This can be achieved by polishing the assembly with short or long reads [27]. Tools which have been used in *Armillaria* genome projects for this purpose include NextPolish, Pilon, FinisherSC, PBJelly2, RS Resequencing application (SMRT Analysis suite) and VelvetOptimiser (Table 1).

Haplotype phasing is the identification of alleles that are co-located on the same chromosome. Haplotype phasing and removal of the identified heterozygous overlaps in assembled genomes is relevant to diploid or polyploid genomes and in genomes which have regions of high heterozygosity [51]. This is because most algorithms in diploid-aware assemblers (e.g., Canu) will assemble these regions as separate contigs, rather than the expected single haplotype-fused contig, once a pair of allelic sequences exceeds a certain threshold of nucleotide diversity [52]. *Armillaria* spp. usually exist as prolonged diploid vegetative states [5]; hence most sequenced genomes are diploid and warrant removal of heterozygous overlaps in assembled *Armillaria* genomes. This can be achieved by utilizing standalone tools such as Purge Haplotigs, as demonstrated in the *A. gallica* strain 012 m genome project (Table 1) [20]. Purge Haplotigs automates the reassignment of allelic contigs and assists with manual curation of TGS-based genome assemblies specifically [52]. As outlined in Figure 1 & Supplementary Table 1, other tools are available for haplotype phasing and differ in how they function (see provided tool websites and/or references).

Genome quality evaluation

The quality of an assembled genome must be evaluated after genome assembly. This evaluation should be conducted in terms of genome completeness, the amount of misassemblies and the presence of contaminants. Metrics such as N50 sizes and L50 are indicative of the level of contiguity (high N50 and L50 values) of the genome assembly [50,53]. N50 sizes are the contig/scaffold size above which half the genome is represented, while L50 is the length in bases of the shortest contig for which 50% of the genome can be contained within contigs of that size or larger. These two metrics are computed based on the lengths of assembly scaffolds or contigs [50].

Various tools can be used for genome assembly evaluation. *Armillaria* genome projects have used tools such as BUSCO ('Benchmarking Universal Single-Copy Orthologs'), QUAST ('Quality ASsessment Tool'), RS BridgeMapper protocol (SMRT Analysis Suite) and the now defunct CEGMA ('Core Eukaryotic Genes Mapping Approach') to evaluate the completeness and quality of the respective genome assemblies (Table 1). BUSCO analyses of *Armillaria* genome projects have utilized different versions of BUSCO (5.3.2 and 3.0.1) and different lineage-specific datasets (agaricales_odb10 lineage and basidiomycota_odb9 datasets) (Table 1). The results generated by BUSCO depend on the lineage selected and version of database of the lineage used for evaluation of the sequenced organism, as well as the version of the tool itself. For example, the basidiomycota_odb9 dataset contained 1335 BUSCO groups, while the basidiomycota_odb10 dataset contains 1764 BUSCO groups. Moreover, the agaricales_odb10 and eukaryota_odb10 datasets contain 3870 and 255 BUSCO groups, respectively. BUSCO v. 5.4.4 is the current stable version [54]. Consequently, care should be taken when comparing genome completeness based on BUSCO analyses from different genome projects.

In addition to evaluating genome completeness and correctness, genomes should be evaluated for absence of contaminants (i.e., nontarget sequences). Genome contamination can occur and be detected at various stages of a genome project [55]. These contaminants can be detected and removed with various tools which function by using various parameters including detection of a highly skewed (i.e., extremely low or high) GC content; presence of an uncharacteristically high percentage of duplicated complete BUSCO genes; and results of BLAST searches of sections of the genome in the NCBI database [55]. BlobTools and Anvi'o have been used to detect and remove genome contamination from an assembled fungal genome [56]. Aside from BUSCO, none of the indicated tools were used in the *Armillaria* genome projects as genome contaminants may have been absent in the respective assemblies.

In summary, genome quality evaluation should be conducted in all genome projects. Genomes that seemingly have unsatisfactory completeness and correctness metrics can be reassembled with other tools or workflows with altered assembly parameters, and the best assembly in terms of these metrics selected as the final assembly. Minimal contamination can be resolved with tools such as the ones indicated in the review by Cornet and Baurain [55], some of which have been outlined in the present review (Figure 1 & Supple-

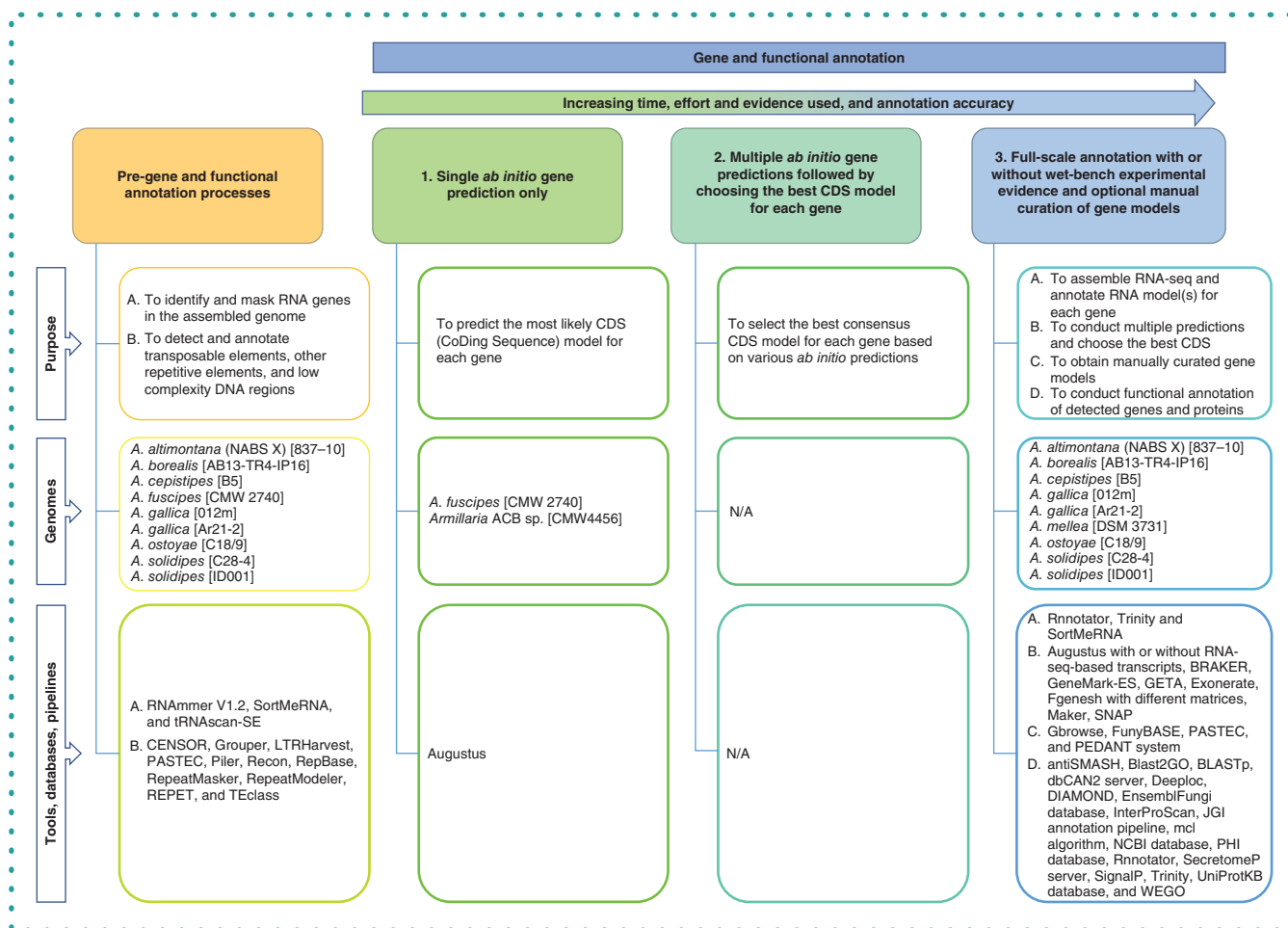


Figure 2. Steps and tools used in *Armillaria* genome annotation. Levels 1, 2 and 3 are based on the approaches to genome annotation relative to time, effort and evidence used and the resultant increase in annotation accuracy, as outlined by Yandell and Ence [53]. Level 2 can be performed with the *ab initio* prediction tools indicated in level 3 (point B under 'Tools, Databases, Pipelines'). ACB: African Clade B; NA: Not applicable.

mentary Table 1). The purpose for the genome project can be used to select the best assembly in genome assembly quality evaluation (e.g., selecting between the assembly with the highest contiguity or the assembly with the highest percentage of complete genes). This is important as assembly quality is not well correlated with assembly contiguity [50]. Additionally, it is apparent that genome quality evaluation technologies, in terms of both computation and datasets used, are being rapidly developed and/or enhanced. Hence research teams embarking on genome projects should keep abreast of the state-of-the-art tools for genome assembly and quality evaluation. Tools used for genome quality evaluation in *Armillaria* genome projects and other available tools compute different metrics which can be accessed via the information provided in Supplementary Table 1. In the case of very low genome completeness and correctness as well as high contamination, it would be more prudent to sequence and assemble a new genome.

Transposable elements & genome annotation

Genome annotation is required to attach biologically meaningful information to assembled genomes, which are otherwise of no significant value to biologists [27]. This process consists of annotation of groups of mobile genetic elements called transposable elements (TEs) as well as gene prediction through *ab initio* approaches with or without the addition of wet-bench experimental data [27]. Wet-bench experimental data for genome annotation include short-reads sequenced expressed sequence tags, RNA sequencing or even long-read sequenced transcripts [27]. Genome annotation also involves describing the structures and functions of the products of the predicted genes. The reader is referred to two book chapters by De Sá *et al.* [57] and Christoffels and Heusden [58] as well as the review articles by Yandell and Ence [53] and Del Angel *et al.* [27] for detailed steps used for genome annotation. As a summary, the methods and depth of *Armillaria* genome annotation have been outlined in Figure 2 and are described in this section.

Gene and functional annotation of assembled genomes is preceded by identification and masking of RNA genes, detection and annotation of TEs and other repetitive elements and detection of low-complexity DNA regions. TEs include long terminal repeats (LTRs),

and their annotation is considered a major task in any genome project [27,59]. RNA gene identification and masking, and detection and annotation of TEs, were carried out in all the *Armillaria* genome projects excluding the *Armillaria* African Clade B sp. (CMW4456) genome project using the tools indicated in Figure 2. Tools employed in *Armillaria* genome projects for identification and masking of RNA genes in the assembled genomes include RNAmmer v. 1.2, Rnnotator, SortMeRNA and tRNAscan-SE [16,19–21]. *De novo* detection and annotation of TEs, other repetitive elements and low-complexity DNA regions have been achieved with tools including BLASTER, Grouper, LTRHarvest, PASTEC, Piler, Recon, RepBase, RepeatMasker, RepeatModeler, REPET and TEclass [16,18–21]. Although not clearly indicated in the *Armillaria* African Clade B sp. (CMW4456) genome project, the AUGUSTUS annotation tool on the Galaxy platform has an option called Softmasking, which was used in the *Armillaria* African Clade B sp. (CMW4456) project. Repeated regions in the genome assembly are detected as lowercase letters when this option is enabled [60].

Gene and functional annotation of *Armillaria* genomes has either been restricted to single *ab initio* gene prediction only or has involved full-scale annotation with or without wet-bench experimental evidence and optional manual curation of gene models (Figure 2). For prediction and annotation of genes and proteins, tools including AUGUSTUS, Blast2GO, dbCAN2 server, Exonerate, Fgenesh, GETA, PHI base and SNAP have been employed [16–22]. Some of these tools have been utilized in pipelines such as BRAKER2 and Maker [16].

Functional annotation of some of the *Armillaria* genomes has incorporated the use of RNA-Seq data [16,20,21]. Annotation with wet-bench experimental data involves transcript assembly using RNA-Seq data. This has been achieved with Trinity and Rnnotator in the *Armillaria* genomes annotated by Sipos *et al.* [16] and with SortMeRNA v. 2.1 in the genome annotation reported by Akulova *et al.* [21]. The tool used for RNA-Seq data assembly was not specified in the genome project by Zhan *et al.* [20]. Following the use of multiple *ab initio* predictors, genome annotation in the genome projects reported by Sipos *et al.* [16] involved manual curation of the predicted gene models using tools or pipelines such as Gbrowse, FunyBASE, PASTEC and PEDANT systems. The authors used the JGI Annotation Pipeline for annotation of the *A. gallica* strain Ar21-2 and *Armillaria solidipes* strain C28-4 genomes.

Further genome annotation has been conducted in *Armillaria* genome projects to gain more insights about these fungi. For this purpose, tools such as BLASTp, GeneMark-ES and GeneMark-ET have been used for gene and protein ortholog finding [16,19,21]. Motifs and domains of proteins have been detected with two versions of InterProScan in the *Armillaria altimontana* (NABS X) strain 837–10, *A. gallica* strain 012 m and *A. solidipes* strain ID001 genome projects [19,20]. BLASTp, SecretomeP server and SignalP were used for putative gene function predictions based on databases and/or protein evidence [16,17,19,21]. Only the *A. altimontana* (NABS X) strain 837–10 and *A. solidipes* strain ID001 genome projects predicted subcellular localizations of the genes and annotated the secondary metabolite biosynthesis gene clusters [19]. These were achieved with Deeploc 1.0 and antiSMASH, respectively. Evaluation of the annotations was achieved using BUSCO in the genome projects reported by Sipos *et al.* [16] and Caballero *et al.* [19].

It is apparent that genome annotation in *Armillaria* genome projects has ranged from minimal annotation to in-depth annotation. The investments in terms of time, effort and evidence (with a consequence on the project budget) needed for genome annotation increase with increasing accuracy of the annotation, as explained by Yandell and Ence [53]. Therefore it is relevant to consider the project budget and other resources, including time and computational constraints, as well as the research questions, to determine whether to perform only *in silico* gene and functional annotation or to include wet-bench experimental evidence with extensive annotations. Further considerations for genome annotation have been discussed by Christoffels and van Heusden [58]. Annotations of all the *Armillaria* genomes have been used to gain some knowledge about the biology of these fungi. Some of these are discussed in the next section.

Knowledge gained from *Armillaria* genomics

Armillaria genome projects have provided draft and complete genome assemblies [15–22]. Standardized QUAST and BUSCO analyses of the presently published *Armillaria* genomes were conducted for accurate comparison in this review (Table 2). This comparison shows that the *Armillaria* genomes ranged between approximately 50.71 and 87.31 Mbp. These genome sizes are comparable among the *Armillaria* species and are within the reported genome sizes (~28–175 Mb) of other Agaricales genomes [61]. Even at the draft stage, *Armillaria* genomes are highly valuable resources with sufficiently high contiguity (N50 values >5000 bp) and completeness (>98% complete BUSCOs) and thus can be used for some comparative genomics and other studies.

Some of the *Armillaria* genomes have been extensively annotated. TE annotation of *Armillaria* genomes by Sipos *et al.* revealed considerable variation in TEs among the different genomes and suggested that unlike in other plant pathogens, TEs do not account for genome expansion in *Armillaria* [16]. Genome expansion of rust genomes, including that of *Austropuccinia psidii* which has an estimated haploid genome size of just over 1 Gb, have been attributed to TEs [63,64]. LTRs were the most highly represented TE family in the genome of *A. borealis* strain AB13-TR4-IP16, with Gypsy (32) being the most abundant LTR type followed by Copia (5) [21]. Other repetitive elements reported in this genome included long interspersed nuclear element retrotransposons and retrotransposons with a tyrosine recombinase. Similarly, the most abundant LTR type in the genome of *A. gallica* strain 012 m was Gypsy (2062) followed by Copia (321) [20]. This trend has been reported in other Basidiomycota genomes [65]. Although the functions of these TEs have not yet been identified in *Armillaria*, these mobile genetic elements have been speculated or shown to have various functions in the genomes of other organisms. These functions include acting as mutators, contributing to genome restructuring and participation in the evolutionary ‘arms race’ between hosts and pathogens [66]. Nevertheless, the contribution of mobile genetic elements to size variation in the mitochondrial genomes (mitogenomes) of *A. borealis*, *A. gallica*, *Armillaria sinapina* and *A. solidipes* was discovered through comparative mitogenomics [67]. It

Table 2. QUAST and BUSCO evaluation of published *Armillaria* genomes.

Assembled isolate	QUAST evaluation metrics [‡]						Complete BUSCOs [§] (%, n = 3870)
	Contigs (n)	Largest contig (bp)	Total length (bp)	GC (%)	N50 (bp)	Ns per 100 kbp (n)	
<i>Armillaria altimontana</i> (NABS X) 837–10	100	5,843,527	73,739,702	47.77	1,930,169	14	98.8 (98.2, 0.6)
<i>Armillaria borealis</i> AB13-TR4-IP16]	10,402	2,136,877	56,994,024	47.76	75,348	5501	97.9 (97.4, 0.5)
<i>Armillaria cepistipes</i> B5	182	6,523,365	75,519,607	47.59	2,288,895	0	98.9 (97.8, 1.1)
<i>Armillaria fuscipes</i> CMW 2740	17,117	157,180	50,706,099	47.68	6152	982	92.6 (88.9, 3.7)
<i>Armillaria gallica</i> 012 m	63	6,431,929	87,305,441	47.38	2,159,699	0	98.7 (95.9, 2.8)
<i>A. gallica</i> Ar21-2	319	4,779,317	85,336,812	47.35	1,035,263	8165	98.6 (97.9, 0.7)
<i>Armillaria mellea</i> DSM 3731 [†]	9,867	639,705	75,786,598	46.99	26,609	16,520	97.9 (97.5, 0.4)
<i>Armillaria ostoyae</i> C18/9	106	6,405,655	60,106,801	48.33	2,283,935	0	98.8 (98.2, 0.6)
<i>Armillaria solidipes</i> C28-4	229	3,399,694	58,009,494	48.35	715,667	3906	98.9 (98.3, 0.6)
<i>A. solidipes</i> (form <i>A. ostoyae</i>) ID001	72	4,463,803	55,735,298	48.26	2,424,439	5	98.7 (98.2, 0.5)
<i>Armillaria</i> African Clade B sp. CMW4456	840	1,463,441	54,954,005	46.53	128,967	0	98.2 (97.0, 1.2)

Unless otherwise stated, the table was created with analysis of genome statistics of Fasta sequences of published genomes at the National Center for Biotechnology Information Genome [62].

[†] Downloaded from JGI MycoCosm.

[‡] QUAST analyses conducted in Galaxy using the genome assembly mode, and 'Fungus use of GeneMark-ES for gene finding, Barrnap for ribosomal RNA genes prediction'.

[§] BUSCO analyses performed using BUSCO v.5.3.2 in genome mode using AUGUSTUS gene predictor with agaricales_odb10 lineage dataset. BUSCO data are presented as total complete BUSCOs (single-copy BUSCOs, duplicated BUSCOs). There were <5% fragmented and missing BUSCOs each in all the genomes.

BUSCO: Benchmarking universal single-copy ortholog; JGI: Joint Genome Institute.

is relevant to conduct further studies on the functions of TEs in the genomes of *Armillaria* spp. to gain a better understanding of the effects of TEs on the evolution, ecology and other characteristics of these fungi.

The number of protein-coding genes identified in all the genomes ranged from 13,600 in *Armillaria* African Clade B sp. (CMW4456) to 26,261 in *A. gallica* 012 m [15–22]. The predicted gene contents of the *Armillaria* genomes are within the range reported by Ruiz-Duenas *et al.* [61] for other Agaricales genomes (5000–29,000 genes). Although the effect of gene contents (in terms of number of genes) and genome size have not been evaluated in *Armillaria*, Ruiz-Duenas *et al.* reported that these properties do not significantly correlate to the lifestyles exhibited by the Agaricales [61].

Information gained from *Armillaria* genomics has contributed to the body of knowledge regarding their genome evolution and architecture as well as their biology. For instance, Sipos *et al.* reported that there is a significant genome expansion in *Armillaria* spp. compared with other related fungi, with a consequent effect on various properties, including gene family expansion of pathogenicity-related genes [16]. There are also records of many relatively small differences in gene contents of genomes of *Armillaria* spp. which potentially contribute to differences in their ecological lifestyles [11,15,16,19,68]. All these records point to genome plasticity of *Armillaria* spp. which can mediate adaptability to changing environments and resistance to, or evasion of, host defense mechanisms. These changes can occur through gene inactivation, altered gene sequence or structure and the birth of new genes, which often encode proteins involved in host–pathogen interactions, as reviewed by Raffaele and Kamoun [69].

The availability of genome sequences and consequent genome annotations has allowed for phylogenomic studies on *Armillaria* spp. In this way, the position of *Armillaria* in the Physalacriaceae and existence of *Guyanagaster* and *Cylindrobasidium* as the closest relatives were confirmed using 835 conserved single-copy genes (188,895 amino acid sites) [16]. Furthermore, the age of pathogenic *Armillaria* spp. was estimated at 21 million years and their divergence from *Guyanagaster* is reckoned to be 42 million years [16]. Phylogenomics, by virtue of using large numbers of orthologous genes, provides an advantage over traditional phylogenetic methods that employ few loci for obtaining well-supported and accurate species delimitation [70]. This is explained by the fact that individual gene genealogies may differ from each other and from the organismal phylogeny and therefore affect the phylogeny obtained when the analysis is based on few genes [70]. However, the use of more genes or loci is not always better in phylogenomics. This is because the possibility of obtaining highly supported but incorrect phylogenetic results, that are due to inconsistency resulting from systematic error from nonphylogenetic signals or model inadequacy, increases when large amounts of data are included in phylogenomic analyses [71]. To minimize this risk, it has been suggested that the number of genes or loci used in phylogenomic analyses should be proportional to the number of taxa being studied [71]. In the context of *Armillaria*, the need for more resources and techniques for species delimitation has been highlighted in various review articles, including those by Baumgartner *et al.* [2], Coetzee *et al.* [3] and Heinzemann *et al.* [5]. The availability of genome sequences and the possibility of sequencing more genomes offer more opportunities for bridging this gap. Several genomes of *Armillaria* spp. with diverse geographical distribution will be required for further phylogenomic and/or phylogeographic studies to fully delimit species and species complexes among *Armillaria*, detect trends in horizontal gene transfer, decipher other evolutionary history of these fungi, and use multispecies phylogenetic comparisons to infer putative functions for DNA or protein sequences.

Genomic studies have also resulted in deciphering some of the molecular basis of sexual reproduction in *Armillaria* spp. The mating (*MAT*) loci of *A. solidipes*, *A. ostoyae*, *A. cepistipes* and *A. gallica* have been reported [16]. Further insights were gained about the chromosomal locations of the *MAT-A* and *MAT-B* loci with the establishment of the only near-chromosome-scale genome of an *Armillaria* spp., namely *A. ostoyae*. This research led to the discovery of the location of the *MAT* loci on pseudochromosomes LG1 and LG6, respectively [15]. Fungal sexual reproduction influences genetic recombination, with a consequent effect on genetic diversity. These mechanisms can play a role in the evolutionary 'arms race' between fungal phytopathogens and their hosts, as well as adaptation to hostile environments [72,73]. Hence more information about the *MAT* loci is relevant for understanding some of the mechanisms underlying the broad host range, host–pathogen interactions and global distribution of *Armillaria* spp.

Discoveries made about some important characteristics of *Armillaria* spp. have benefited from multidisciplinary approaches which incorporated genomics with other experimental investigations. For instance, through multi-omics studies including transcriptomics and proteomics, Sipos *et al.* reported that *A. cepistipes*, *A. ostoyae* and *A. gallica* possess more pectinolytic gene families than lignolytic gene families compared with other white-rot fungi [16]. The authors suggest that these gene families may enable *Armillaria* spp. to avoid competition with other white-rot fungi, enhancing their ability to colonize and thrive on some hosts. Differential regulation of the genes encoding these biotic factors was also reported in various tissues (mycelia, rhizomorphs, fruiting bodies at different developmental stages and in different tissues of a mature fruiting body) of *A. ostoyae*, with a potential influence on multicellular development as well as cap and stipe differentiation [16].

Other important discoveries have been made about *Armillaria* spp. based on multi-omics and biological investigations. Combining genomics with *in vitro* analyses of recombination rate variation, Heinzelmann *et al.* discovered 1984 crossover events in the progeny of *A. ostoyae*, highly variable recombination rates along chromosomes, and regions close to the telomeres serving as gene-poor recombination hotspots [15]. Multiple glycol-degradative enzyme systems which can be useful for recalcitrant carbon compound degradation, and potent antifungal properties of *A. mellea*, were also discovered based on genomic and proteomic studies [17]. Moreover, comparative genomics coupled with *in vitro* targeted metabolomics studies resulted in records of conserved putative nonribosomal peptide synthetases-dependent siderophore synthetase gene clusters and *in vitro* species- and strain-independent biosynthesis of different types and quantities of siderophores by strains of various *Armillaria* species [74]. Hence, while genomic data can provide some answers, using genome data in combination with *in vitro*, *in situ* and *in plantae* studies, including other bioinformatics studies, can provide a much greater understanding of the species in this genus. This multidisciplinary approach to investigating fungal phytopathogens has also been recommended in the review by Aylward *et al.* [75].

Suggestions for future genomics & transcriptomics studies on the genus *Armillaria*

Armillaria genome projects have yielded 11 high-quality publicly available genomes at the contig or scaffold levels on the NCBI database. These genomes have been applied in further research to gain a better understanding of these fungi and of phytopathogens in general. Nonetheless, there is still more to be achieved in terms of genomics and multi-omics research on *Armillaria* spp. In this section, prospects in *Armillaria* genomics are suggested.

So far, the genome of *A. ostoyae* strain SBI C18/9 is the only *Armillaria* genome that has been assembled to pseudochromosome (near-chromosome) level. The pseudochromosome was constructed based on the order of scaffolds within genetic map linkage groups of haploid progenies of *A. ostoyae* strain C15 [15]. Efforts should be made toward obtaining more chromosome-scale *Armillaria* genomes in order to derive the benefits of such data, as has been achieved for vertebrate genomes using pipelines such as the Vertebrate Genomes Project assembly pipeline, which was designed for generating high-quality, near-error-free, gap-free, chromosome-level, haplotype-phased, annotated reference genome assemblies for every vertebrate species [76]. This pipeline is hosted on the Galaxy platform and uses PacBio or Nanopore reads and additional techniques such as Hi-C sequencing [60,76]. Hi-C sequencing is a supporting technology used to improve the contiguity of genome assemblies [27,77]. It is expected that the Vertebrate Genomes Project pipeline will be suitable for assembling diploid, highly heterozygous *Armillaria* genomes to chromosomal level and this should be explored in future. It is imperative that this pipeline be validated using raw long- and short-read data of *Armillaria* sequences as well as additional sequences obtained using other sequencing techniques. In the absence of Hi-C and other supporting data, the workflows in Figures 1 & 2 can be used as guidelines for embarking on more *Armillaria* genome projects to achieve optimum genome assemblies and annotation given available resources. Resources for accessing the tools indicated in Figures 1 & 2 as well as other resources have been provided in Supplementary Table 1.

Other discoveries can be achieved with more studies based on *Armillaria* genomes. These discoveries can be facilitated by combining genomics with other experimental studies. The benefits of availability of genomes of numerous strains of fungal phytopathogens coupled with other wet-bench experimental studies have been discussed [75,78]. In the context of *Armillaria*, embarking on assembling more high-quality genomes of several strains belonging to various species and with a worldwide geographical sampling promises to be highly resourceful, especially if a multidisciplinary approach is adopted. Such studies will help bridge the wide knowledge gap about the mechanisms used by these fungi in their evolutionary success. Additionally, the considerable variation in aggressiveness, pathogenicity and virulence both intra- and inter-species among pathogenic members of *Armillaria* is still not clearly understood [25,26]. The mechanisms underlying these characteristics can also be deciphered with the multidisciplinary approach.

Some other possible research areas that can be facilitated by the multidisciplinary approach include:

1. Putatively establishing genes encoding enzymes involved in secondary metabolite biosynthesis and deciphering the expression patterns of genes encoding secondary metabolites in vegetative mycelia, rhizomorphs and fruiting bodies at different developmental stages of different *Armillaria* spp. under different growth conditions. This will help further understand the role that different developmental stages and different tissues play in *Armillaria* pathogenicity.
2. Further research aimed at establishing and understanding the mechanisms of the antibacterial, antifungal and cytotoxic properties of secondary metabolites produced by *Armillaria* spp.
3. Further investigation of the molecular basis of the mating system of *Armillaria* spp. or strains; for instance, to unravel the mechanisms of diploidization of the haploid mycelia and the signals that trigger these mechanisms.
4. Developing and/or optimizing tools for functional validation of genes in *Armillaria* for the study of gene function and annotation in *Armillaria*.
5. Conducting more research on the secretion, presence, structure and activity of ligninolytic enzymes, other plant cell wall degrading enzymes (PCWDEs) and other carbohydrate-active enzymes of *Armillaria* spp. *in silico*, *in vitro* and *in situ*.
6. Comparing the quantity and diversity of pathogenicity-related gene products such as PCWDE in *Armillaria* spp. with various lifestyles *in silico*, *in vitro* and *in plantae* to establish the effect of enzyme machineries possessed by different *Armillaria* spp. on their pathogenicity and how they circumvent or outcompete other sympatrically existing *Armillaria* spp. as well as microbes normally used as biocontrol agents of *Armillaria* root-rot disease. This information can be used for developing enhanced biocontrol methods of *Armillaria* root-rot.

Conclusion

Armillaria genome projects have achieved several milestones toward understanding the biology and evolution of these fungi. Advantage should be taken of the fast pace of technological advancement in genome sequencing, assembly and annotation, and the reducing cost of these technologies, to increase genome projects in *Armillaria* and other Basidiomycota. These should be coupled with other wet-bench experimental studies in a timely manner. Results from these studies can afford the scientific, agroindustrial and pharmaceutical sectors the benefits of the knowledge and technologies that can be gained from these projects.

Future perspective

Given the rapid advances in techniques and technologies relating to genome sequencing, assembly and annotation, as well as the continually reducing costs, we expect an exponential increase in the number of genome sequences in the next year or two. The information provided in this review will help facilitate this. The processes described toward producing high-quality *Armillaria* genome sequences can be extrapolated to genomic studies of other diploid and dikaryotic fungi and other organisms. It is expected that the availability of high-quality genome sequences will enable a better understanding of the biology and biochemistry of organisms when multidisciplinary approaches are followed. This has the potential to result in the discovery of novel biotechnological products with applications in agriculture, bioremediation, forestry and medicine. Consequently, this would facilitate the achievement of several of the 17 Sustainable Development Goals to sustain our world [79].

Executive summary

Background

- The current status of genomic studies on *Armillaria* species, with a focus on the diversity of species and strains for which genomes have been assembled and annotated, is considered in this comprehensive overview of the research carried out on this fungal genus.

Status of genome sequencing, assembly & annotation in *Armillaria*

- Various steps, techniques and tools have been used for *Armillaria* genome extraction, sequencing, assembly, evaluation and annotation.
- Infographic guidelines for future *Armillaria* genome projects with some relevant resources are provided.
- The guidelines and resources provided can be applied to other fungal genera and beyond mycology.

Knowledge gained from *Armillaria* genomics

- *Armillaria* genome projects have provided draft and complete genome assemblies which have been used to gain further insights into the biology and chemistry of these fungi.
- The range of predicted gene contents of the *Armillaria* genomes are within the range of between 5,000 and 29,000 genes predicted for other Agaricales genomes.

Suggestions for future genomics studies on the genus *Armillaria*

- Potential areas for future *Armillaria* genome projects, as well as areas for potential comparative genomics and multi-omics studies, are outlined.

Conclusion

- Although this review centers around genomic studies of *Armillaria* spp., the methods, guidelines and information gathered from these studies provide valuable information that can be utilized for genomic and transcriptomic research on other fungi and beyond mycology.

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-2023-0023

Author contributions

MPA Coetzee and BD Wingfield supervised the project; DL Narh Mensah collected the information, designed the tables and figures and wrote the manuscript. All authors read and agreed to publish the final version of the manuscript.

Financial disclosure

Funding was received from the University of Pretoria, the South African Department of Science and Innovation (DSI)-National Research Foundation (NRF) Centre of Excellence in Plant Health Biotechnology (CPHB, previously the CTHB), the DSI–NRF South African Research Chairs Initiative (SARChI) in Fungal Genomics (grant no. 98353) and the Tree Protection Cooperative Program (TPCP). The grant holders acknowledge that opinions, findings and conclusions or recommendations expressed in this publication are those of the authors, and that the funding bodies accept no liability whatsoever in this regard. 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.

Competing interests disclosure

The authors have no competing interests or relevant affiliations with any organization or entity with the subject matter or materials discussed in the manuscript. This includes employment, consultancies, honoraria, stock ownership or options, expert testimony, grants or patents received or pending, or royalties

Writing disclosure

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

Open access

This work is licensed under the Attribution-NonCommercial-NoDerivatives 4.0 Unported License. To view a copy of this license, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>

References

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

- Sipos G, Anderson JB, Nagy LG. *Armillaria*. *Curr. Biol.* 28(7), R297–R298 (2018).
- Baumgartner K, Coetzee MPA, Hoffmeister D. Secrets of the subterranean pathosystem of *Armillaria*. *Mol. Plant Pathol.* 12(6), 515–534 (2011).
- Gives a good overview of the genus *Armillaria*.
- Coetzee MPA, Wingfield BD, Wingfield MJ. *Armillaria* root-rot pathogens: species boundaries and global distribution. *Pathogens* 7(4), 83 (2018).
- Provides further information on *Armillaria* species.
- Koch RA, Wilson AW, Séné O, Henkel TW, Aime MC. Resolved phylogeny and biogeography of the root pathogen *Armillaria* and its gasteroid relative, *Guyanagaster*. *BMC Evol. Biol.* 17(1), 33 (2017).
- Heinzelmann R, Dutech C, Tsykun T, Labbé F, Soularue J-P, Prospero S. Latest advances and future perspectives in *Armillaria* research. *Can. J. Plant Pathol.* 41(1), 1–23 (2019).
- General overview of the latest advances of research on *Armillaria* and insights for future research on the genus.
- Baumgartner K, Rizzo DM. Ecology of *Armillaria* spp. in mixed-hardwood forests of California. *Plant Dis.* 85(9), 947–951 (2001).
- Baumgartner K, Rizzo DM. Spread of *Armillaria* root disease in a California vineyard. *Am. J. Enol. Vitic.* 53(3), 197–203 (2002).
- Guillaumin J-J, Mohammed C, Anselmi N *et al.* Geographical distribution and ecology of the *Armillaria* species in Western Europe. *Eur. J. For. Pathol.* 23(6–7), 321–341 (1993).
- Labbé F, Marçais B, Dupouey J-L *et al.* Pre-existing forests as sources of pathogens? The emergence of *Armillaria ostoyae* in a recently planted pine forest. *For. Ecol. Manage.* 357, 248–258 (2015).
- Legrand P, Guillaumin J-J. *Armillaria* species in the forest ecosystems of the Auvergne (Central France). *Acta Oecol.* 14, 389–403 (1993).
- Sahu N, Merényi Z, Bálint B *et al.* Hallmarks of basidiomycete soft- and white-rot in wood-decay – omics data of two *Armillaria* species. *Microorganisms* 9(1), 149 (2021).
- Lundell TK, Mäkelä MR, De Vries RP, Hildén KS. Genomics, lifestyles and future prospects of wood-decay and litter-decomposing Basidiomycota. In: *Advances in Botanical Research*. Martin FM (Ed.). 329–370, Academic Press, Oxford, UK (2013).
- Ren S, Gao Y, Li H *et al.* Research status and application prospects of the medicinal mushroom *Armillaria mellea*. *Appl. Biochem. Biotechnol.* 195(5), 3491–3507 (2022).
- Kedves O, Shahab D, Champramary S *et al.* Epidemiology, biotic interactions and biological control of armillarioids in the northern hemisphere. *Pathogens* 10(1), 76 (2021).
- Heinzelmann R, Rigling D, Sipos G, Münsterkötter M, Croll D. Chromosomal assembly and analyses of genome-wide recombination rates in the forest pathogenic fungus *Armillaria ostoyae*. *Heredity* 124(6), 699–713 (2020).
- The first paper to report genome assembly up to pseudochromosome level for *Armillaria* spp.
- Sipos G, Prasanna AN, Walter MC *et al.* Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi *Armillaria*. *Nat. Ecol. Evol.* 1(12), 1931–1941 (2017).
- Collins C, Keane TM, Turner DJ, O’Keeffe G, Fitzpatrick DA, Doyle S. Genomic and proteomic dissection of the ubiquitous plant pathogen, *Armillaria mellea*: toward a new infection model system. *J. Proteome Res.* 12(6), 2552–2570 (2013).
- Wingfield BD, Ambler JM, Coetzee M *et al.* Draft genome sequences of *Armillaria fuscipes*, *Ceratocystopsis minuta*, *Ceratocystis adiposa*, *Endoconidiophora laricicola*, *E. polonica* and *Penicillium freii* DAOMC 242723. *IMA Fungus* 7(1), 217–227 (2016).
- Caballero JRI, Lalande BM, Hanna JW, Klopfenstein NB, Kim MS, Stewart JE. Genomic comparisons of two *Armillaria* species with different ecological behaviors and their associated soil microbial communities. *Microb. Ecol.* doi:10.1007/s00248-022-01989-8 (2022) (Epub ahead of print).
- Zhan M, Tian M, Wang W *et al.* Draft genomic sequence of *Armillaria gallica* 012 m: insights into its symbiotic relationship with *Gastrodia elata*. *Braz. J. Microbiol.* 51(4), 1539–1552 (2020).
- Akulova VS, Sharov VV, Aksyonova AI *et al.* De novo sequencing, assembly and functional annotation of *Armillaria borealis* genome. *BMC Genom.* 21(Suppl. 7), 534 (2020).

22. Wingfield BD, Berger DK, Coetzee MPA et al. IMA genome-F17: draft genome sequences of an *Armillaria* species from Zimbabwe, *Ceratocystis colombiana*, *Elsinoe necatrix*, *Rosellinia necatrix*, two genomes of *Sclerotinia minor*, short-read genome assemblies and annotations of four *Pyrenophora teres* isolates from barley grass, and a long-read genome assembly of *Cercospora zeina*. *IMA Fungus* 13(1), 19 (2022).
23. NIH. Genome. www.ncbi.nlm.nih.gov/data-hub/genome
24. JGI. MycoCosm: the functional genomics resource. www.jgi.doe.gov/fungi/Agaricales
25. Prospero S, Holdenrieder O, Rigling D. Comparison of the virulence of *Armillaria cepistipes* and *Armillaria ostoyae* on four Norway spruce provenances. *Forest Pathol.* 34(1), 1–14 (2004).
26. Labbé F, Lung-Escarmant B, Fievet V et al. Variation in traits associated with parasitism and saprotrophism in a fungal root-rot pathogen invading intensive pine plantations. *Fungal Ecol.* 26, 99–108 (2017).
27. Del Angel VD, Hjerde E, Sterck L et al. Ten steps to get started in genome assembly and annotation. *F1000Res.* 7(ELIXIR) 148 (2018).
28. Jung H, Winefield C, Bombarely A, Prentis P, Waterhouse P. Tools and strategies for long-read sequencing and *de novo* assembly of plant genomes. *Trends Plant Sci.* 24(8), 700–724 (2019).
29. Gawel N, Jarret R. A modified CTAB DNA extraction procedure for *Musa* and *Ipomoea*. *Plant Mol. Biol. Rep.* 9(3), 262–266 (1991).
30. Porebski S, Bailey LG, Baum BR. Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. *Plant Mol. Biol. Reporter* 15(1), 8–15 (1997).
31. Cubero OF, Crespo A, Fatehi J, Bridge PD. DNA extraction and PCR amplification method suitable for fresh, herbarium-stored, lichenized, and other fungi. *Plant Syst. Evol.* 216(3–4), 243–249 (1999).
32. Huang X, Duan N, Xu H, Xie TN, Xue YR, Liu CH. CTAB–PEG DNA extraction from fungi with high contents of polysaccharides. *Mol. Biol.* 52(4), 718–726 (2018).
33. Van Burik JA, Schreckhise RW, White TC, Bowden RA, Myerson D. Comparison of six extraction techniques for isolation of DNA from filamentous fungi. *Med. Mycol.* 36(5), 299–303 (1998).
34. Karstens L, Siddiqui NY, Zaza T, Barstad A, Amundsen CL, Sysøeva TA. Benchmarking DNA isolation kits used in analyses of the urinary microbiome. *Sci. Rep.* 11(1), 6186 (2021).
35. Whitehouse CA, Hottel HE. Comparison of five commercial DNA extraction kits for the recovery of *Francisella tularensis* DNA from spiked soil samples. *Mol. Cell Probes* 21(2), 92–96 (2007).
36. Lung M-Y, Chang Y-C. Antioxidant properties of the edible Basidiomycete *Armillaria mellea* in submerged cultures. *Int. J. Mol. Sci.* 12(10), 6367–6384 (2011).
37. Zavastin DE, Mircea C, Aprotosoia AC, Gherman S, Hancianu M, Miron A. *Armillaria mellea*: phenolic content, *in vitro* antioxidant and antihyperglycemic effects. *Rev. Med. Chir. Soc. Med. Nat. Iasi* 119(1), 273–280 (2015).
38. Engels B, Heinig U, Grothe T, Stadler M, Jennewein S. Cloning and characterization of an *Armillaria gallica* cDNA encoding protoilludene synthase, which catalyzes the first committed step in the synthesis of antimicrobial melleolides. *J. Biol. Chem.* 286(9), 6871–6878 (2011).
39. An S, Lu W, Zhang Y, Yuan Q, Wang D. Pharmacological basis for use of *Armillaria mellea* polysaccharides in Alzheimer's disease: antiapoptosis and antioxidation. *Oxid. Med. Cell. Longev.* 2017, 4184562 (2017).
40. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat. Methods* 8(1), 61–65 (2011).
41. Giani AM, Gallo GR, Gianfranceschi L, Formenti G. Long walk to genomics: history and current approaches to genome sequencing and assembly. *Comput. Struct. Biotechnol. J.* 18, 9–19 (2020).
42. Athanasopoulou K, Boti MA, Adamopoulos PG, Skourou PC, Scorilas A. Third-generation sequencing: the spearhead towards the radical transformation of modern genomics. *Life* 12(1), 30 (2022).
43. Li X, Xu S, Zhang J, Li M. Assembly and annotation of whole-genome sequence of *Fusarium equiseti*. *Genomics* 113(4), 2870–2876 (2021).
44. Segerman B. The most frequently used sequencing technologies and assembly methods in different time segments of the bacterial surveillance and RefSeq genome databases. *Front. Cell Infect. Microbiol.* 10, 527102 (2020).
45. Minei R, Hoshina R, Ogura A. *De novo* assembly of middle-sized genome using MinION and Illumina sequencers. *BMC Genom.* 19(1), 700 (2018).
46. Tyler AD, Mataseje L, Urfano CJ et al. Evaluation of Oxford Nanopore's MinION sequencing device for microbial whole genome sequencing applications. *Sci. Rep.* 8(1), 10931 (2018).
47. Gupta AK, Kumar M. Benchmarking and assessment of eight *de novo* genome assemblers on viral next-generation sequencing data, including the SARS-CoV-2. *Omics* 26(7), 372–381 (2022).
48. Forouzani E, Maleki MSM, Karkhane AA, Yakhchali B. Evaluation of nine popular *de novo* assemblers in microbial genome assembly. *J. Microbiol. Methods* 143, 32–37 (2017).
49. Khan AR, Pervaz MT, Babar ME, Naveed N, Shoaib M. A comprehensive study of *de novo* genome assemblers: current challenges and future prospective. *Evol. Bioinform. Online* 14, 1176934318758650 (2018).
50. Salzberg SL, Phillippy AM, Zimin A et al. GAGE: a critical evaluation of genome assemblies and assembly algorithms. *Genome Res.* 22(3), 557–567 (2012).
51. Zhang X, Wu R, Wang Y, Yu J, Tang H. Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.* 18, 66–72 (2020).
52. Roach MJ, Schmidt SA, Borneman AR. Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinf.* 19(1), 460 (2018).
53. Yandell M, Ence D. A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.* 13(5), 329–342 (2012).
54. University of Geneva Faculty of Medicine. BUSCO. <http://busco.ezlab.org/>
55. Cornet L, Baurain D. Contamination detection in genomic data: more is not enough. *Genome Biol.* 23(1), 60 (2022).
56. Aylward J, Wingfield MJ, Roets F, Wingfield BD. A high-quality fungal genome assembly resolved from a sample accidentally contaminated by multiple taxa. *BioTechniques* 72(2), 39–50 (2022).
57. De Sá PHCG, Guimarães LC, Das Graças DA et al. Next-generation sequencing and data analysis: strategies, tools, pipelines and protocols. In: *Omics Technologies and Bio-Engineering*. Barh D, Azevedo V (Eds). 191–207, Academic Press, Oxford, UK, (2018).
- This book chapter provides relevant information about sequencing platforms, structural and functional genomics, as well as metagenomics.
58. Christoffels A, Van Heusden P. Genome annotation: perspective from bacterial genomes. In: *Encyclopedia of Bioinformatics and Computational Biology*. Ranganathan S, Gribskov M, Nakai K, Schönbach C (Eds). 152–156, Academic Press, Oxford, UK, (2019).
59. Jung H, Ventura T, Chung JS et al. Twelve quick steps for genome assembly and annotation in the classroom. *PLoS Comput. Biol.* 16(11), e1008325 (2020).
60. The Galaxy Community. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2022 update. *Nucleic Acids Res.* doi:10.1093/nar/gkac247 (2022).
61. Ruiz-Duenas FJ, Barrasa JM, Sanchez-Garcia M et al. Genomic analysis enlightens Agaricales lifestyle evolution and increasing peroxidase diversity. *Mol. Biol. Evol.* 38(4), 1428–1446 (2021).
62. NIH. Genome. *Armillaria*. www.ncbi.nlm.nih.gov/data-hub/genome/?taxon=47424
63. Tobias PA, Schwessinger B, Deng CH et al. *Austropuccinia psidii*, causing myrtle rust, has a gigabase-sized genome shaped by transposable elements. *G3 (Bethesda)* 11(3), jkaa015 (2021).
64. Tan MK, Collins D, Chen Z, Englezou A, Wilkins MR. A brief overview of the size and composition of the myrtle rust genome and its taxonomic status. *Mycology* 5(2), 52–63 (2014).
65. Castanera R, Borgognone A, Pisabarro AG, Ramirez L. Biology, dynamics, and applications of transposable elements in basidiomycete fungi. *Appl. Microbiol. Biotechnol.* 101(4), 1337–1350 (2017).
66. Kidwell MG, Lisch DR. Perspective: transposable elements, parasitic DNA, and genome evolution. *Evolution* 55(1), 1–24 (2001).
67. Kolesnikova AI, Putintseva YA, Simonov EP et al. Mobile genetic elements explain size variation in the mitochondrial genomes of four closely-related *Armillaria* species. *BMC Genom.* 20(1), 351 (2019).
68. Heinzelmann R, Croll D, Zoller S et al. High-density genetic mapping identifies the genetic basis of a natural colony morphology mutant in the root rot pathogen *Armillaria ostoyae*. *Fungal Genet. Biol.* 108, 44–54 (2017).
69. Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10(6), 417–430 (2012).
70. Pamilo P, Nei M. Relationships between gene trees and species trees. *Mol. Biol. Evol.* 5(5), 568–583 (1988).
71. Young AD, Gillung JP. Phylogenomics – principles, opportunities and pitfalls of big-data phylogenetics. *Syst. Entomol.* 45(2), 225–247 (2020).
72. Grandaubert J, Dutheil JY, Stukenbrock EH. The genomic determinants of adaptive evolution in a fungal pathogen. *Evol. Lett.* 3(3), 299–312 (2019).

73. Zhan J, Mundt CC, McDonald BA. Sexual reproduction facilitates the adaptation of parasites to antagonistic host environments: evidence from empirical study in the wheat-*Mycosphaerella graminicola* system. *Int. J. Parasitol.* 37(8–9), 861–870 (2007).
74. Narh Mensah DL, Wingfield BD, Coetzee MPA. Nonribosomal peptide synthetase gene clusters and characteristics of predicted NRPS-dependent siderophore synthetases in *Armillaria* and other species in the Physalacriaceae. *Curr. Genet.* 69(1), 7–24 (2023).
75. Aylward J, Steenkamp ET, Dreyer LL, Roets F, Wingfield BD, Wingfield MJ. A plant pathology perspective of fungal genome sequencing. *IMA Fungus* 8(1), 1–15 (2017).
 - Provides some valuable information regarding genome sequencing projects of plant pathogenic fungi in the Ascomycota.
76. Rhie A, McCarthy SA, Fedrigo O et al. Towards complete and error-free genome assemblies of all vertebrate species. *Nature* 592(7856), 737–746 (2021).
77. Lieberman-Aiden E, Van Berkum NL, Williams L et al. Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science* 326(5950), 289–293 (2009).
78. Coetzee MPA, Santana QC, Steenkamp ET, Wingfield BD, Wingfield MJ. Fungal genomes enhance our understanding of the pathogens affecting trees cultivated in southern hemisphere plantations. *South. For.* 82(3), 215–232 (2020).
 - Provides some high-level information regarding genome sequencing projects of tree pathogenic fungi.
79. Department of Economic and Social Affairs. The 17 Goals. <https://sdgs.un.org/goals>