scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Fully resolved assembly of *Fusarium proliferatum* DSM106835 genome

Gouthaman P. Purayil[1], Amal Y. Almarzooqi[1], Khaled A. El-Tarabily[1] ✉, Frank M. You[2] ✉ & Synan F. AbuQamar[1] ✉

In the United Arab Emirates, sudden decline syndrome (SDS) is a destructive disease of date palm caused by the soil-borne fungal pathogen *Fusarium proliferatum* (*Fp*) DSM106835. Here, a high-resolution genome assembly of *Fp* DSM106835 was generated using PacBio HiFi sequencing with Omni-C data to provide a high-quality chromatin-organised reference genome with 418 scaffolds, totalling 58,468,907 bp in length and an N50 value of 4,383,091 bp from which 15,580 genes and 16,321 transcripts were predicted. The assembly achieved a complete BUSCO score of 99.2% for 758 orthologous genes. Compared to seven other *Fp* strains, *Fp* DSM106835 exhibited the highest continuity with a cumulative size of 44.26 Mbp for the first ten scaffolds/contigs, surpassing the assemblies of all examined *Fp* strains. Our findings of the high-quality genome of *Fp* DSM106835 provide an important resource to investigate its genetics, biology and evolutionary history. This study also contributes to fulfill the gaps in fungal knowledge, particularly the genes/metabolites associated with pathogenicity during the plant-pathogen interaction responsible for SDS.

Background & Summary

Date palm (*Phoenix dactylifera*) is considered as one of the most economically important fruit crops grown in arid lands of the Arabian Peninsula, the Middle East and North Africa. This evergreen tree is well-adapted to harsh desert conditions of long hot summers, little rainfall and low humidity. The United Arab Emirates (UAE) has the largest number of date palms in the world, and is considered among the top global exporters of dates[1]. On the other hand, date palm orchards in the UAE have recently been suffering from serious diseases caused by fungal pathogens[2,3], including sudden death syndrome (SDS; also known as date palm wilt disease)[4].

Although researchers have reported several *Fusarium* species that are associated with disease symptoms of SDS worldwide[3,5–7], *Fusarium oxysporum* f.sp. *cumini* (*Foc*) DSM106834, *F. proliferatum* (*Fp*) DSM106835 and *F. solani* (*Fs*) DSM106836 are the causal agents of SDS on date palm in the UAE[4]. In North Africa, Bayoud is the most destructive fungal disease of date palm that is linked with *F. oxysporum* f.sp. *albedinis* (*Foa*)[8,9]. *Fs* was, however, found associated with declined date palm trees in Pakistan[10]. In the UAE, *Fp* was identified the main *Fusarium* spp. causing SDS in Saudi Arabia, Iraq, Jordan and Tunisia[11–14].

The soil-borne filamentous fungus *Fp* is a plant pathogen that belongs to the family Nectiraceae from the division Ascomycota. *Fp* is part of the *F. fujikuroi* species complex (FFSC) that is composed of around 60 different phylogenetic species with phytopathological and clinical relevance[15,16]. As other *Fusarium* spp., *Fp* has the ability to produce the mycotoxin, fumonisin[17,18]. Fumonisins are carcinogenic, estrogenic and immune suppressive in mammals and may cause birth defects of the brain and spinal cord[18,19]. Other mycotoxins, such as beauvericin, enniatins and moniliformin, can also be produced by *Fp* and act as virulence factors and specific effectors to elicit resistance to SDS in date palm[11,13,14].

Although SDS has been reported to negatively affect date palm plantations in the UAE and elsewhere, the genetic information of the causal agent is still meager. Therefore, we developed a whole genome sequencing of *Fp* DSM106835 using PacBio® to provide high throughput sequencing with highly accurate long HiFi reads. Here, we presented a highly contiguous and complete *de novo* genome assembly for *Fp* DSM106835, the main causal agent of SDS on date palm in the UAE, using PacBio HiFi long-reads and Omni-C data. The final genome is about 58.5 Mbp across 418 scaffolds, with a scaffold N50 of 4.4 Mbp and a Benchmarking Universal Single-Copy Orthologs (BUSCO)[20] score of 99.2%. This genome adds a valuable resource for studying the evolutionarily

[1]Department of Biology, College of Science, United Arab Emirates University, Al Ain, 15551, United Arab Emirates. [2]Ottawa Research and Development Centre, Agriculture and Agri-Food Canada, 960 Carling Avenue, Ottawa, ON, K1A 0C6, Canada. ✉e-mail: ktarabily@uaeu.ac.ae; frank.you@agr.gc.ca; sabuqamar@uaeu.ac.ae

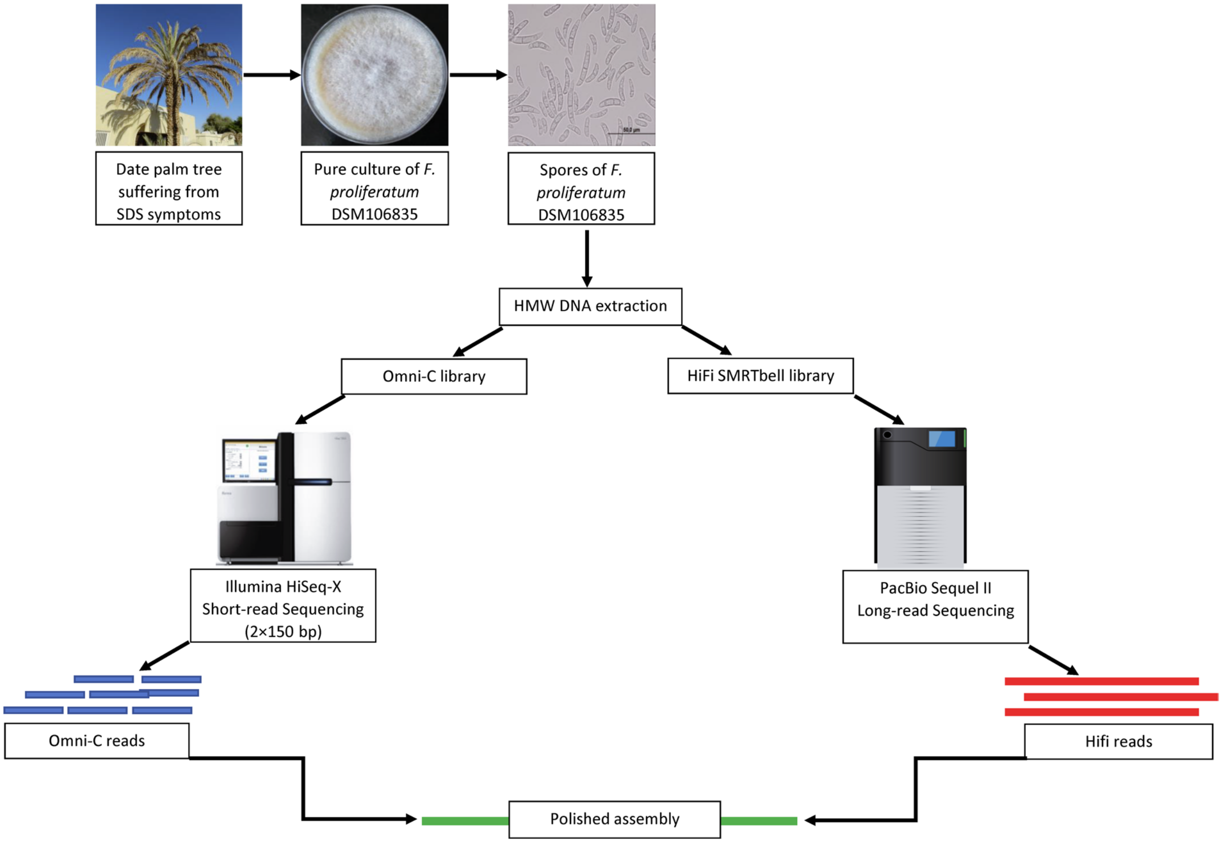
Date palm tree suffering from SDS symptoms
Pure culture of *F. proliferatum* DSM106835
Spores of *F. proliferatum* DSM106835
HMW DNA extraction
Omni-C library
HiFi SMRTbell library
Illumina HiSeq-X Short-read Sequencing (2×150 bp)
PacBio Sequel II Long-read Sequencing
Omni-C reads
Hifi reads
Polished assembly

Fig. 1 Flow diagram of the isolation, genome sequencing and assembly of *Fusarium proliferatum* DSM106835. Date palm trees showing symptoms of SDS were used to establish a pure culture of *F. proliferatum* DSM106835. Spores produced by the fungal pathogen were observed under light microscopy and further used for HMW DNA extraction. Omni-C and HiFi SMRbell libraries were prepared for Illumina HiSeq-X (short-read sequencing) and PacBio® Sequel II (long-read sequencing), respectively. HiFi and Omni-C reads were merged to develop a long-read-only assembly where all chromosomes were present as single contigs without the introduction of artificial gaps (Courtesy of Illumina, Inc., Pacific Biosciences of California, Inc.). SDS, sudden death syndrome; HMW, high molecular weight.

relationships and elucidating the molecular mechanisms for host specificity to further improve our understanding of *Fp* DSM106835-date palm interaction.

Methods

Growth and culture maintenance of *F. proliferatum* DSM106835. The pathogen, *Fp* DSM106835, was previously isolated from date palm trees showing SDS symptoms from Al Wagan area in Al Ain, Abu Dhabi, UAE, grown and maintained in potato dextrose agar plates (PDA; Lab M Limited, Lancashire, UK) supplemented with 25 mg/L penicillin-streptomycin (Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany) at 25°C[4]. Plates were subcultured every 14 days on PDA plates until pure *Fp* DSM106835 colonies were obtained. A flow scheme of the isolation and culturing of *Fp* DSM106835 can be found in Fig. 1.

DNA extraction and PacBio HiFi sequencing. High molecular weight (HMW) DNA was extracted by first scraping all visible fungal material from the Petri dish, which was then transferred to a 50-ml tube with 2-ml of autoclaved ddH_2O, flash frozen to create a pellet of ~500 mg, and ground to become powder. In the ground sample, 10 ml of cetyltrimethyl ammonium bromide (CTAB) and 100 µl of β-mercaptoethanol (BME) were added and incubated at 68°C for 15 minutes. After incubation, 10 µl of protease and 1 µl of RNase were added to the sample and incubated at 60°C for 30 minutes. Phenol/chloroform/isoamyl-alcohol was used to extract DNA from the cell lysate, which was then centrifuged into a pellet. The formed pellet was resuspended in 200 µl Tris-EDTA buffer (TE buffer). DNA samples were first sequenced using the PacBio Sequel II sequencer at Dovetail Genomics (Scotts Valley, California, USA). This sequencing step was carried out by preparing PacBio SMRTbell libraries (~20 kbp) using the SMRTbell Express Template Prep Kit 2.0 (PacBio, Menlo Park, CA), according to the manufacturer's protocol.

Omni-C sequencing. Omni-C sequencing is a chromatin conformation capture technology that allows the investigation of the genome's three-dimensional (3D) organisation. The Omni-C library was prepared using the

Reads	Total HiFi CCS reads	1,754,151
	Total HiFi CCS read size (bp)	26,392,037,220
	Average coverage (X)	503
	Mean read size (bp)	15,045
	Median read size (bp)	14,761
	N50 read size (bp)	15,154
	Max read size (bp)	49,366
	Total Omni-C reads	114,895,105
	Total Omni-C read size (bp)	17,234,265,750
Assembly	Genome size (bp)	58,468,907
	Estimated genome size by K-mer analysis (Mbp)	47.07
	No. of scaffolds	418
	Scaffold N50	4,383,091

Table 1. Information on the assembly of *Fusarium proliferatum* DSM106835.

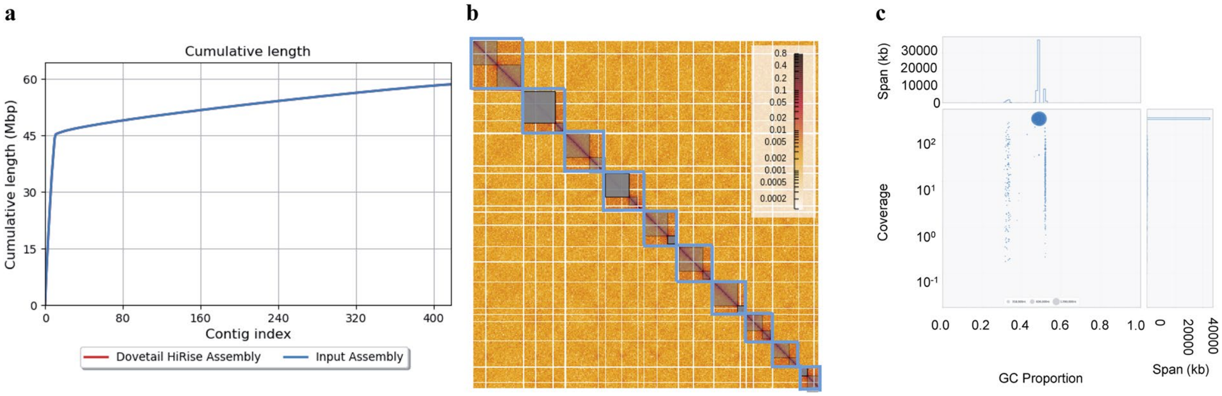

Fig. 2 Taxonomic partitioning, average read length of the raw data and Omni-C contact map of *Fusarium proliferatum* DSM106835. (**a**) The Cumulative length of scaffolds for the assembly; (**b**) Omni-C contact map showing the intensity of the physical interaction between genome regions; and (**c**) Taxonomic partitioning of *F. proliferatum* DSM106835 raw reads generated using blobtools2. In (**b**), the primary 10 chromosome-length scaffolds are highlighted in blue. In (**c**), blue represents Ascomycota while grey represents the reads with no-hits.

Dovetail® Omni-C® Kit according to the manufacturer's protocol. Briefly, the chromatin was fixed with disuccinimidyl glutarate (DSG) and formaldehyde in the nucleus. The crosslinked chromatin was *in situ* digested with DNaseI.

After digestion, chromatin fragments attached to Chromatin Capture Beads were released by lysing the cells with sodium dodecyl sulfate (SDS) buffer. The chromatin ends were repaired followed by ligation to a biotinylated bridge adapter. After proximity ligation, crosslinks were reversed and DNA was purified. The sequencing libraries using Illumina-compatible adaptors were generated. Biotin-containing fragments were isolated using streptavidin beads before PCR amplification. The library was sequenced on an Illumina HiSeq-X platform. A flow scheme of HMW DNA extraction, library preparations and genome assembly of *Fp* DSM106835 can be found in Fig. 1.

De novo genome assembly.

The genome assembly was carried out by first using 26.9 Gbp of PacBio Circular Consensus Sequencing (CCS) reads as an input to the hifiasm assembler[21] with default parameters to create the initial *de novo* assembly. Omni-C sequencing resulted in a paired-end set of raw reads, each 11,489,515 bp in length and GC content of 49% (Table 1). These reads, along with the *de novo* assembly, were used as input data for HiRise[22], a software pipeline explicitly designed for using proximity ligation data to scaffold genome assemblies (Fig. 2a). Dovetail Omni-C library sequences were aligned to the draft input assembly using BWA[23], and pairtools[24] was used to remove the PCR duplicates from the assembly; followed by SAMtools[25] to generate the final bam file. Quality control using the script get_qc.py part of the HiRise package found 88,132,543 (76.71%) of read pairs were mapped and 12,232,575 (10.65%) were unmapped. The HiRise pipeline was used to identify misassemblies, and to break and sort scaffolds (only those above the threshold) in accordance with the likelihood model used by HiRise. Omni-C contact maps were created from the output of HiRise using Juicer[26], and the contact map was configured to identify Topologically Associated Domains and A/B genome compartments. The configured contact map was visualised using Juicebox[27] (Fig. 2b). The final *de novo* assembly of 58,468,907 bp in length had an N50 value of 4,383,091. This assembly was used as a query to perform a BLASTN[28] search against the National Center for Biotechnology Information (NCBI) nucleotide database[29] as an input for blobtools2[30] to visualise the assembly and its taxonomic partitioning (Fig. 2c). The HiCanu[31] assembler was also used to assemble

Type of Element	Number of Element	Length (bp)	Percentage of Sequence (%)
Retroelements	248	482,156	0.82
LINEs	66	215,895	0.37
LTR elements	182	266,261	0.46
Gypsy/DIRS1	46	222,508	0.38
DNA transposons	339	391,106	0.67
Tc1-IS630-Pogo	269	161,039	0.28
Rolling-circles	3239	1,522,651	2.60
Unclassified	1610	1,560,876	2.67
Small RNA	2490	2,632,552	4.50
Simple repeats	6204	248,208	0.42
Low complexity	649	31,602	0.05
Total	15342	7,734,854	13.23

Table 2. Repeat sequence analysis of the genome of *Fusarium proliferatum* DSM106835.

the genome to compare and validate the hifiasm assembly. The completeness of the final assembly was assessed using BUSCO with fungi_odb10 lineage-specific profile[32].

Transposable element analysis, gene prediction and annotation. The assembly of *Fp* DSM106835 was subjected to transposable element (TE) analysis using a customised repeat annotation pipeline. This pipeline incorporated multiple *de novo* TE discovery tools, including RepeatModeler[33], HelitronScanner[34], MITE Tracker[35], SINEScan[36], and RepeatMasker. In brief, RepeatModeler integrates RECON[37], RepeatScout[38], and LTRHavest/LTRretriver[39]. These tools obtained a comprehensive representation of TEs, leading to a relatively complete TE library. Subsequently, RepeatMasker was employed with this library to identify genome-wide TEs and mask all the repeats and tandem sequences. The resulting masked genome sequences were then subjected to *de novo* gene prediction and annotation using BRAKER 2[40]. In the BRAKER 2 pipeline, Augustus[41] was trained with protein sequences of orthologous genes in fungi genomes to help in gene prediction. The genome was then subjected to functional annotation and Gene Ontology (GO) analysis using Blast2GO[42], and the prediction of secondary metabolites was performed using fungal-antiSMASH[43].

Assessment of completeness and continuity of the genome assembly. For assembly continuity comparison, the genome sequences of seven *Fp* strains with gene annotations, ET1 (FJOF00000000)[44], FFSC RH7 (JAJALB000000000)[45], Fp_A8 (MRDB00000000)[46], ITEM2341 (PKMI00000000)[47], MPVP328 (PKMJ00000000)[48], NRRL62905 (FCQG00000000)[49], and R16 (PKMG00000000)[50] were downloaded from the NCBI database. These strains were compared against *Fp* DSM106835 by comparing the sequence length of each assembly with the average scaffold length, and completeness analysis was performed by comparing the results of BUSCO analysis of each genome against fungi_odb10 lineage-specific profile.

Data Records

All sequence data, including raw HiFi long reads and Omni-C short reads, were deposited to the NCBI database under BioProject PRJEB64160, with accessions ERR11733479[51] and ERR11733478[52], respectively. The genome assembly is available through NCBI GenBank with the accession CAUHTQ000000000[53]. The genome annotation information was deposited in the Figshare database[54].

Technical Validation

Evaluating the quality of the genome assembly. The PacBio sequencing produced 1,754,151 raw HiFi long reads with an average read length of 15,045.5 bp, resulting in 26.4 Gbp, mostly falling between 5,000–25,000 bp in length and approximately 560x coverage (Supplementary Fig. S1). By utilising the hifiasm and HiRise software, the assembly of HiFi reads with Omni-C reads generated 418 scaffolds, amounting to 58.47 Mbp. The N50 value was 4.38 Mbp. The largest 11 scaffolds had a combined size of 45.18 Mbp, which accounted for 77.3% of the entire genome (Table 1). Similar results were obtained when the assembly of HiCanu was compared to that using hifiasm (Supplementary Fig. S2). The assembly achieved a completeness rate of 99.2% for the 758 orthologous genes in fungi_odb10 using BUSCO, similar to the genome assembly of *Fp* strain Fp_A8 (99.3%; Table 1).

Genome annotation. A total of 3.96 Mbp of transposable repeat sequences were detected in the genome of *Fp* DSM106835, including retroelements (0.48 Mbp), DNA transposons (0.39 Mbp), rolling-circle replicates (Helitrons; 1.52 Mbp), and some unclassified repeats (1.56 Mbp), collectively constituting 6.76% of the total genome (Table 2; Fig. 3). Notably, the genome of *Fp* DSM106835 also included long terminal repeat (LTR) retroelements that belong to Gypsy superfamily. Heitron rolling-circle elements and unclassified elements accounted for a significant part of repeat sequences. The gene prediction using BRAKER2[45] resulted in 15,580 putative genes, of which 267 were TE and 15,313 were non-TE genes. We also detected 16,321 transcripts, where the average gene length was about 1,580 bp. After performing functional annotation on the predicted sequences, GO terms distribution for cellular components, molecular function, and biological processes was identified (Fig. 4a) with the highest number of annotations belonging to GO levels 3–7. The evidence code distribution was calculated, and mostly they received a hit from Inferred from Electronic Annotation (IEA) and Inferred from Biological aspect

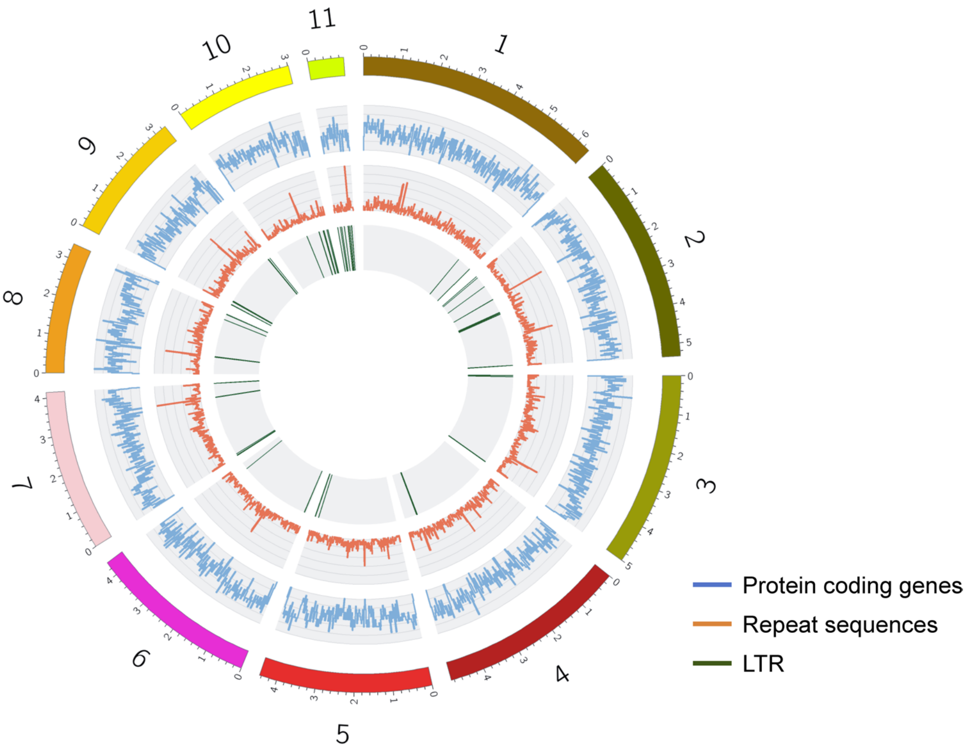

Fig. 3 Circos map of the 11 significant scaffolds for *Fusarium proliferatum* DSM106835. Outer track represents the ideogram of 11 scaffolds. The bin size of each track was 20 Kbp. LTR, long terminal repeats.

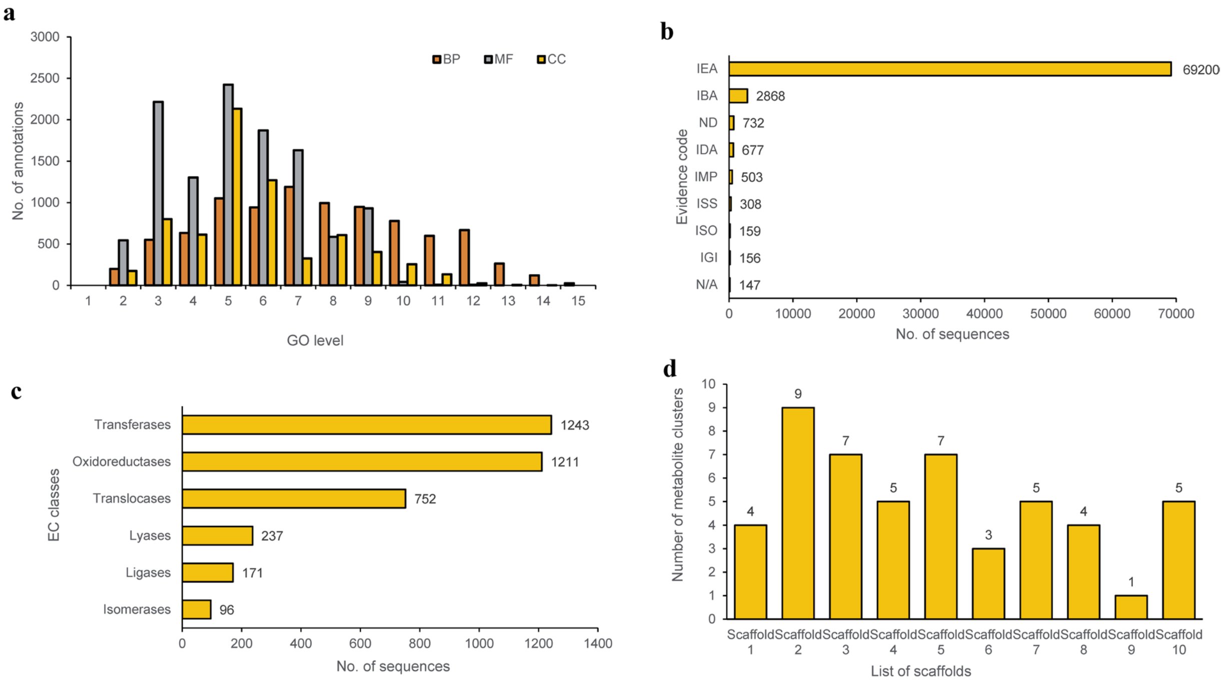

Fig. 4 Functional annotation and Gene Ontology distribution for *Fusarium proliferatum* DSM106835. (**a**) Distribution of GO generated from the genome of *F. proliferatum* DSM106835; and (**b**) evidence code distribution for the obtained sequences. (**c**) EC classification for sequences present in the assembly; and (**d**) the number of secondary metabolite biosynthesis gene clusters identified from the first 11 scaffolds of the genome of *F. proliferatum* DSM106835. In (**b**), the distribution of evidence code for functional terms was obtained during the mapping step. GO, Gene Ontology; BP, biological process; MF, molecular function; CC, cellular component; EC, enzyme code.

Region	Type	From	To	Cluster similarity	Similarity (%)
2.1	T1PKS	81,001	185,855	asperfuranone	18%
2.3	T1PKS,NRPS	275,348	382,994	gibepyrone-A	100%
2.7	NRPS,T1PKS	2,708,378	2,756,301	NG-391	83%
2.8	NRPS	4,987,936	5,037,349	beauvericin	20%
2.9	T1PKS	5,121,665	5,169,546	fumonisin	52%
3.6	NRPS,T1PKS	4,661,484	4,713,770	equisetin	54%
3.7	NRPS-like,T1PKS	4,897,251	4,974,345	fusaric acid	72%
4.3	terpene	3,396,575	3,426,909	squalestatin S1	40%
4.5	T1PKS	4,082,972	4,130,027	oxyjavanicin	100%
5.2	NRPS-like,NRPS	94,150	150,568	destruxin A	9%
5.3	terpene	205,874	230,579	gibberellin	100%
5.6	T1PKS	4,268,976	4,313,388	bikaverin	71%
5.7	NRPS-like	4,318,228	4,372,942	fusaridione A	12%
7.5	NRPS	4,056,001	4,098,456	acetylaranotin	40%
8.1	NRPS,T1PKS	620,775	672,633	ACT-Toxin II	100%
8.2	terpene	940,349	985,876	koraiol	100%
8.3	T1PKS	1,882,799	1,926,993	fujikurin A/B/C/D	100%
8.4	T1PKS	2,802,293	2,844,647	solanapyrone D	33%
10.1	T1PKS	396,309	444,038	neurosporin A	20%
10.4	terpene	2,119,715	2,146,446	α-acorenol	100%

Table 3. List of secondary metabolite biosynthetic gene clusters identified from the genome of *Fusarium proliferatum* DSM106835 using antiSMASH.

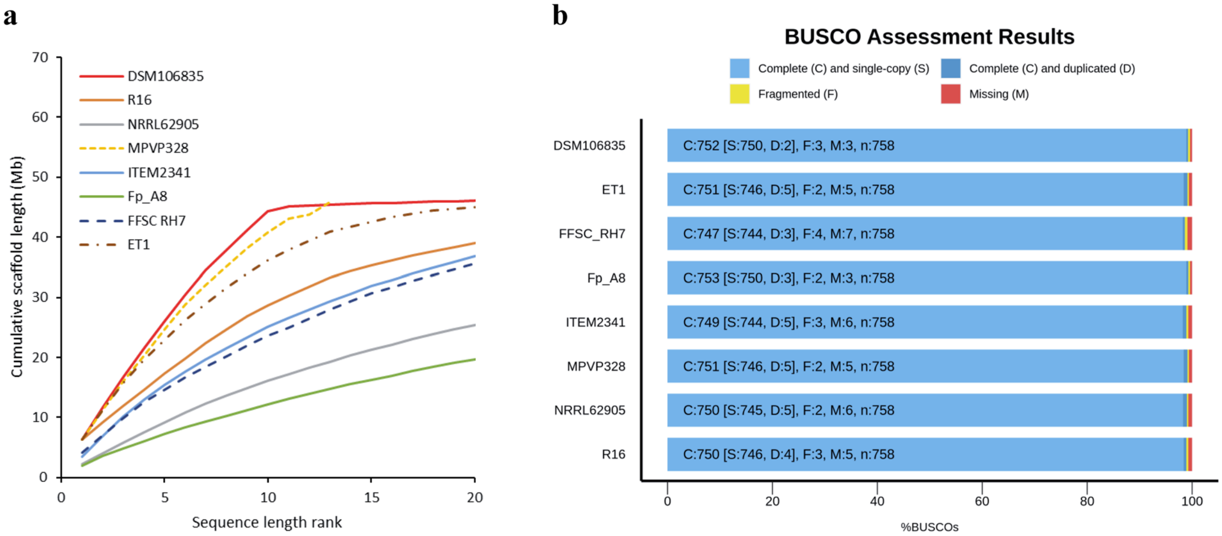

Fig. 5 Contiguity and completeness of the assembly of *Fusarium proliferatum* DSM106835. (**a**) Contiguity; and (**b**) completeness of the assembly of *F. proliferatum* DSM106835 compared with assemblies of seven *F. proliferatum* strains. In (**a**), only the first 20 longest scaffolds were presented.

of Ancestor (IBA) sections (Fig. 4b). Similarly, the enzyme code (EC) classification was carried out, from which most of the sequences were found to be either transferases or oxidoreductases (Fig. 4c).

The number of secondary metabolite biosynthesis gene clusters was also identified (Fig. 4d). In general, various gene clusters ranging from clinically relevant fumonisins, virulence-related ACT-Toxin II, and phytotoxic destruxin A were present in the genome. Gene clusters of secondary metabolites were found to belong to the biosynthesis of fusaric acid, oxyjavanicin, gibberellin, bikaverin, ACT-Toxin II, koraiol, Fujikurin A, α-acorenol, NG-391 and Gibepyrone A (Table 3).

Genome continuity and completeness analysis. The continuity analysis revealed that *Fp* DSM106835 exhibited the highest continuity among the seven *Fp* strains collected from NCBI. The cumulative size of the first 10 scaffolds/contigs was 44.26 Mbp, which surpassed the assemblies of all other *Fp* strains ranging from 12.19 Mbp in *Fp* Fp_A8) to 36.19 Mbp in *Fp* ET1 (Fig. 5a). The same genomes were compared for their completeness using BUSCO[19], and *Fp* DSM106835 achieved a completeness rate of 99.2% for the 758 orthologous genes in the Fungi_odb10 database, which is comparable to *Fp* Fp_A8 (99.3%; Fig. 5b).

Code availability

This work did not utilise a custom script. Data processing was carried out using the protocols and manuals of the relevant bioinformatics software.

Received: 25 July 2023; Accepted: 28 September 2023;
Published online: 16 October 2023

References

1. FAO. *World Food and Agriculture – Statistical Yearbook 2021*. https://doi.org/10.4060/cb4477en (FAO, 2021).
2. Saeed, E. E. *et al*. Chemical control of dieback disease on date palm caused by the fungal pathogen, *Thielaviopsis punctulata*, in United Arab Emirates. *Plant Dis.* **100**, 2370–2376 (2016).
3. Alhammadi, M. S., Al-Shariqi, R., Maharachchikumbura, S. & Al-Sadi, A. M. Molecular identification of fungal pathogens associated with date palm root diseases in the United Arab Emirates. *J. Plant Pathol.* **99**, 1–7 (2018).
4. Alwahshi, K. J. *et al*. Molecular identification and disease management of date palm sudden decline syndrome in the United Arab Emirates. *Int. J. Mol. Sci.* **20**, 923 (2019).
5. Armengol, J. *et al*. Identification, incidence and characterization of *Fusarium proliferatum* on ornamental palms in Spain. *Eur. J. Plant Pathol.* **112**, 123–131 (2005).
6. Mansoori, B. & Kord, H. Yellow death: A disease of date palm in Iran caused by *Fusarium solani*. *J. Phytopathol.* **154**, 125–127 (2006).
7. Al-Otibi, F., Al-Zahrani, R. M. & Marraiki, N. Biodegradation of selected hydrocarbons by *Fusarium* species isolated from contaminated soil samples in Riyadh, Saudi Arabia. *J. Fungi* **9**, 216 (2023).
8. Tantaoui, A., Ouinten, M., Geiger, J. P. & Fernandez, D. Characterization of a single clonal lineage of *Fusarium oxysporum* f.sp. *albedinis* causing Bayoud disease of date palm in Morocco. *Phytopathology* **86**, 787–792 (1996).
9. El Hassni, M. *et al*. Biological control of bayoud disease in date palm: selection of microorganisms inhibiting the causal agent and inducing defense reactions. *Environ. Exp. Bot.* **59**, 224–234 (2007).
10. Maitlo, W. A., Markhand, G. S., Abul-Soad, A. A., Lodhi, A. M. & Jatoi, M. A. Chemcial control of sudden decline disease of date palm (*Phoenix dactylifera* L.) in Sindh, Pakistan. *Pak. J. Bot.* **45**, 7–11 (2013).
11. Abdalla, M. Y., Al-Rokibah, A., Moretti, A. & Mulè, G. Pathogenicity of toxigenic *Fusarium proliferatum* from date palm in Saudi Arabia. *Plant Dis.* **84**, 321–324 (2000).
12. Hameed, M. A. Inflorescence rot disease of date palm caused by *Fusarium proliferatum* in Southern Iraq. *Afr. J. Biotechnol.* **11**, 8616–8621 (2012).
13. Alananbeh, K., Tahat, M. M. & Al-Taweel, H. First report of *Fusarium proliferatum* on date palm (*Phoenix dactylifera* L.) in Jordan. *Plant Dis.* https://doi.org/10.1094/PDIS-06-20-1219- (2021).
14. Rabaaoui, A. *et al*. Phylogeny and mycotoxin profile of pathogenic *Fusarium* species isolated from sudden decline syndrome and leaf wilt symptoms on date palms (*Phoenix dactylifera*) in Tunisia. *Toxins* **13**, 463 (2021).
15. Niehaus, E.-M. *et al*. Comparative "omics" of the *Fusarium fujikuroi* species complex highlights differences in genetic potential and metabolite synthesis. *Genome Biol. Evol.* **8**, 3574–3599 (2016).
16. Yilmaz, N. *et al*. Redefining species limits in the *Fusarium fujikuroi* species complex. *Persoonia - Mol. Phylogeny Evol. Fungi* **46**, 129–162 (2021).
17. Rheeder, J. P., Marasas, W. F. O. & Vismer, H. F. Production of fumonisin analogs by *Fusarium* species. *Appl. Environ. Microbiol.* **68**, 2101–2105 (2002).
18. Kamle, M. *et al*. Fumonisins: impact on agriculture, food, and human health and their management strategies. *Toxins* **11**, 328 (2019).
19. Chen, J. *et al*. Fumonisin B1: Mechanisms of toxicity and biological detoxification progress in animals. *Food Chem. Toxicol.* **149**, 111977 (2021).
20. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).
21. Cheng, H., Concepcion, G. T., Feng, X., Zhang, H. & Li, H. Haplotype-resolved *de novo* assembly using phased assembly graphs with hifiasm. *Nat. Methods* **18**, 170–175 (2021).
22. Putnam, N. H. *et al*. Chromosome-scale shotgun assembly using an *in vitro* method for long-range linkage. *Genome Res.* **26**, 342–350 (2016).
23. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
24. Open2C *et al*. Pairtools: from sequencing data to chromosome contacts. 2023.02.13.528389 Preprint at https://doi.org/10.1101/2023.02.13.528389 (2023).
25. Danecek, P. *et al*. Twelve years of SAMtools and BCFtools. *GigaScience* **10**, giab008 (2021).
26. Durand, N. C. *et al*. Juicer provides a one-click system for analyzing loop-resolution Hi-C experiments. *Cell Syst.* **3**, 95–98 (2016).
27. Durand, N. C. *et al*. Juicebox provides a visualization system for Hi-C contact maps with unlimited zoom. *Cell Syst.* **3**, 99–101 (2016).
28. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
29. Sayers, E. W. *et al*. Database resources of the national center for biotechnology information. *Nucleic Acids Res.* **50**, D20–D26 (2022).
30. Challis, R., Richards, E., Rajan, J., Cochrane, G. & Blaxter, M. BlobToolKit – Interactive quality assessment of genome assemblies. *G3-Genes Genom. Genet.* **10**, 1361–1374 (2020).
31. Nurk, S. *et al*. HiCanu: accurate assembly of segmental duplications, satellites, and allelic variants from high-fidelity long reads. *Genome Res.* **30**, 1291–1305 (2020).
32. Kriventseva, E. V. *et al*. OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res.* **47**, D807–D811 (2019).
33. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc. Natl. Acad. Sci.* **117**, 9451–9457 (2020).
34. Xiong, W., He, L., Lai, J., Dooner, H. K. & Du, C. HelitronScanner uncovers a large overlooked cache of Helitron transposons in many plant genomes. *Proc. Natl. Acad. Sci.* **111**, 10263–10268 (2014).
35. Crescente, J. M., Zavallo, D., Helguera, M. & Vanzetti, L. S. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. *BMC Bioinformatics* **19**, 348 (2018).
36. Mao, H. & Wang, H. SINE_scan: an efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. *Bioinformatics* **33**, 743–745 (2017).
37. Bao, Z. & Eddy, S. R. Automated *de novo* identification of repeat sequence families in sequenced genomes. *Genome Res.* **12**, 1269–1276 (2002).
38. Price, A. L., Jones, N. C. & Pevzner, P. A. *De novo* identification of repeat families in large genomes. *Bioinforma. Oxf. Engl.* **21**(Suppl 1), i351–358 (2005).

39. Ellinghaus, D., Kurtz, S. & Willhoeft, U. LTRharvest, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18 (2008).
40. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genomics Bioinforma.* **3**, lqaa108 (2021).
41. Stanke, M. *et al.* AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).
42. Götz, S. *et al.* High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Res.* **36**, 3420–3435 (2008).
43. Blin, K. *et al.* antiSMASH 6.0: improving cluster detection and comparison capabilities. *Nucleic Acids Res.* **49**, W29–W35 (2021).
44. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:FJOF00000000.1 (2016).
45. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:JAJALB000000000.1 (2022).
46. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:MRDB00000000.1 (2018).
47. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:PKMI00000000.1 (2018).
48. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:PKMJ00000000.1 (2021).
49. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:FCQG00000000.1 (2016).
50. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:PKMG00000000.1 (2021).
51. *NCBI Sequence Reads Archive*, https://identifiers.org/ncbi/insdc.sra:ERR11733479 (2023).
52. *NCBI Sequence Reads Archive*, https://identifiers.org/ncbi/insdc.sra:ERR11733478 (2023).
53. *NCBI GenBank*, https://identifiers.org/ncbi/nucleotide:CAUHTQ000000000 (2023).
54. Purayil, G. P., Almarzooqi, A. Y., El-Tarabily, K. A., You, F. M. & AbuQamar, S. F. Fully resolved assembly of *Fusarium proliferatum* DSM106835 genome., *figshare*, https://doi.org/10.6084/m9.figshare.23731314 (2023).

Acknowledgements

We would like to express our gratitude to Dr. C. Zheng for her contribution in preparing some of the figures. This work is supported by Abu Dhabi Education and Knowledge (Grant #: 21S105) to K. El-Tarabily, and Khalifa Center for Genetic Engineering and Biotechnology-UAEU (Grant #: 31R286) to S. AbuQamar.

Author contributions

G. Purayil: data curation, methodology, software, and writing – original draft; A. Almarzooqi: methodology, review, and editing; K. El-Tarabily: conceptualization, resources, and supervision; F. You: data curation, methodology, software, and writing – original draft; S. AbuQamar: conceptualization, data curation, writing – review, editing, and supervision.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41597-023-02610-4.

Correspondence and requests for materials should be addressed to K.A.E.-T., F.M.Y. or S.F.A.

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

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