Data in Brief 57 (2024) 110951

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Draft genome sequence data of *Fusarium verticillioides* strain REC01, a phytopathogen isolated from a Peruvian maize

Richard Estrada[a,*], Liliana Aragón[b], Wendy E. Pérez[c], Yolanda Romero[a], Gabriel Martínez[a], Karina Garcia[a], Juancarlos Cruz[c], Carlos I. Arbizu[d,*]

[a] *Dirección de Desarrollo Tecnológico Agrario, Instituto Nacional de Innovación Agraria (INIA), Av. La Molina 1981, Lima 15024, Peru*
[b] *Facultad de Agronomía, Universidad Nacional Agraria La Molina (UNALM), Av. La Molina s/n, Lima 15024, Peru*
[c] *Dirección de Supervisión y Monitoreo en las Estaciones Experimentales Agrarias, Instituto Nacional de Innovación Agraria (INIA), Av. La Molina 1981, Lima 15024, Peru*
[d] *Facultad de Ingeniería y Ciencias Agrarias, Universidad Nacional Toribio Rodríguez de Mendoza de Amazonas (UNTRM), Cl. Higos Urco 342, Amazonas 01001, Peru*

ARTICLE INFO

Article history:
Received 20 April 2024
Revised 11 September 2024
Accepted 12 September 2024
Available online 23 September 2024

Dataset link: WGS of Fusarium verticillioides isolated from peruvian root corn (Original data)

Keywords:
Genome sequencing
Vascular rot
SSR data mining

ABSTRACT

Fusarium verticillioides represents a major phytopathogenic threat to maize crops worldwide. In this study, we present genomic sequence data of a phytopathogen isolated from a maize stem that shows obvious signs of vascular rot. Using rigorous microbiological identification techniques, we correlated the disease symptoms observed in an affected maize region with the presence of the pathogen. Subsequently, the pathogen was cultured in a suitable fungal growth medium and extensive morphological characterization was performed. In addition, a pathogenicity test was carried out in a DCA model with three treatments and seven repetitions. De novo assembly from Illumina Novaseq 6000 sequencing yielded 456 contigs, which together constitute a 42.8 Mb genome assembly with a GC % content of 48.26. Subsequent

* Corresponding authors.
E-mail addresses: richard.estrada.bioinfo@gmail.com (R. Estrada), carlos.arbizu@untrm.edu.pe (C.I. Arbizu).
Social media: @arbizu_carlos (C.I. Arbizu)

https://doi.org/10.1016/j.dib.2024.110951
2352-3409/© 2024 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/)

comparative analyses were performed with other *Fusarium* genomes available in the NCBI database.

© 2024 The Author(s). Published by Elsevier Inc.
This is an open access article under the CC BY-NC license (http://creativecommons.org/licenses/by-nc/4.0/)

Specifications Table

Subject	Biology
Specific subject area	Genomics, Microbiology, Bioinformatics.
Data format	Raw sequence reads (fastq) Analyzed Filtered
Type of data	Genomic sequence, Table, Figure
Data collection	The strain was collected from a rotten stem of Peruvian maize and subsequently isolated on Potato Dextrose Agar (PDA). Then a microscopic inspection was carried out to observe morphological characteristics of the pathogen. The pathogenicity analysis was performed in a complete randomised design (CRD) model with three treatments; T1 (substrate without inoculum), T2 (1 $g.kg^{-1}$ of substrate) and T3 (5 $g.kg^{-1}$ of substrate) with 7 repetitions. Genomic DNA was extracted using the E.Z.N.A kit. Sequencing was performed using the Illumina Novaseq 6000 platform and *de novo* assembly MaSuRCA version 4.0.6. To evaluate the quality and integrity of the genome assembly, BUSCO version 5.4.2 was used. Additionally, identification of Simple Sequence Repeats (SSR) regions was performed using MISA software.
Data source location	Institution: EEA Vista Florida, Instituto Nacional de Innovación Agraria City/Town/Region: Chiclayo (6.72°S, 79.77°W) Country: Perú
Data accessibility	Repository name: NCBI (National Center for Biotechnology Information) Data identification number: SRR21935339 Direct URL to data: https://www.ncbi.nlm.nih.gov/sra/?term=SRR21935339 The Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accession JAWRVJ000000000.1

1. Value of the Data

- The genomic sequence data provide valuable information for conducting comparative genomic studies aimed at exploring the core genome of the *Fusarium* genus.
- The genomic sequence data allows a detailed examination of genes related to pathogenicity in the affected plant, paving the way for the development of genetic engineering strategies aimed at reducing the pathogen's virulence toward the host plant.
- The genome data is essential for advancing our understanding of the phylogeny of the *Fusarium genus*, contributing to the analysis of evolutionary relationships within this group of phytopathogenic fungi.

2. Background

The incidence of *Fusarium verticillioides* as a phytopathogen poses a substantial threat to the production and quality of maize, with significant economic repercussions [1]. The presence of mycotoxins in this crop adversely impacts human and animal health, leading to economic losses and affecting international trade relations [2]. F. verticillioides has multiple routes of infection, including systemic infection in seedlings, entry through the stigma, and infection of stems and ears due to mechanical damage [3].

In this study, *F. verticillioides* was isolated from maize samples (INIA 627 - Pátapo variety) collected in Lambayeque, Peru, which exhibited symptoms of pink vascular rot. The tissues were disinfected and cultured on PDA medium, where a change in mycelial morphology from white to violet was observed, along with the presence of macroconidia and microconidia. The phytopathogenic activity was confirmed by a reduction in seed germination and the number of seedlings as the inoculum concentration increased, demonstrating the pathogen's negative impact.

Sequencing of F. verticillioides was conducted using the Illumina Novaseq 6000 platform, with de novo assembly performed using MaSuRCA v. 4.0.6. The quality and integrity of the assembled genome were assessed using BUSCO v 5.4.2, and Simple Sequence Repeat (SSR) regions were identified with MISA software. This genomic analysis is crucial for characterising local strains and understanding their genetic variability, which can influence the pathogen's virulence and resistance. Such comparisons are essential for identifying regional differences and developing targeted management strategies to protect local agriculture.

3. Data Description

3.1. Sample collection

Maize samples of crop season 2022 were taken from a plot at Experimental Station Vista Florida in Lambayeque, Peru (6°43′33"S 79°46′44"W). Then, the samples were placed in paper envelopes and maintained at room temperature. Sample characterization and analysis was carried out within the next 10 days.

3.2. Microbiological identification

The pieces of plant tissues (stems and roots) exhibiting symptoms of vascular lesion were processed to detect infection by *Fusarium* sp. A characterization of affected tissues of Peruvian maize (INIA 627 - Pátapo variety) was conducted, involving segments of stems and roots displaying a pinkish vascular rot (Fig. 1A-B). The stem tissue samples were disinfected with 1.0 % sodium hypochlorite, washed with sterilised distilled water, and aseptically transferred to Petri dishes containing Potato Dextrose Agar (PDA) medium. The plates were incubated for 7 days at 28°C, then purified on new Petri dishes with PDA medium [4]. Initially, the strain exhibited white mycelial morphology (Fig. 1C) that transitioned to a violet hue (Fig. 1D) with age. Microscopic examination facilitated the observation of detailed morphological features, including macroconidia (Fig. 1E), microconidia (Fig. 1F) and conidiogenous cells (Fig. 1G).

3.3. Strain phytopathogenic assay

The phytopathogenic activity was verified through the germination of seeds inoculated with this strain. Table 1 shows the result of the germination percentage for each treatment 15 days

Table 1
Germination percentage of maize seeds in substrate inoculated with *F. verticillioides* isolated from maize, 15 days after sowing under semi-controlled mesh house conditions.

Treatment	Inoculum concentration	Phytopathogenic activity
T1	0g.Kg^{-1}	100 %
T2	1g.Kg^{-1}	57.10 %*
T3	5g.Kg^{-1}	33.30 %*

* significance $p < 0.05$

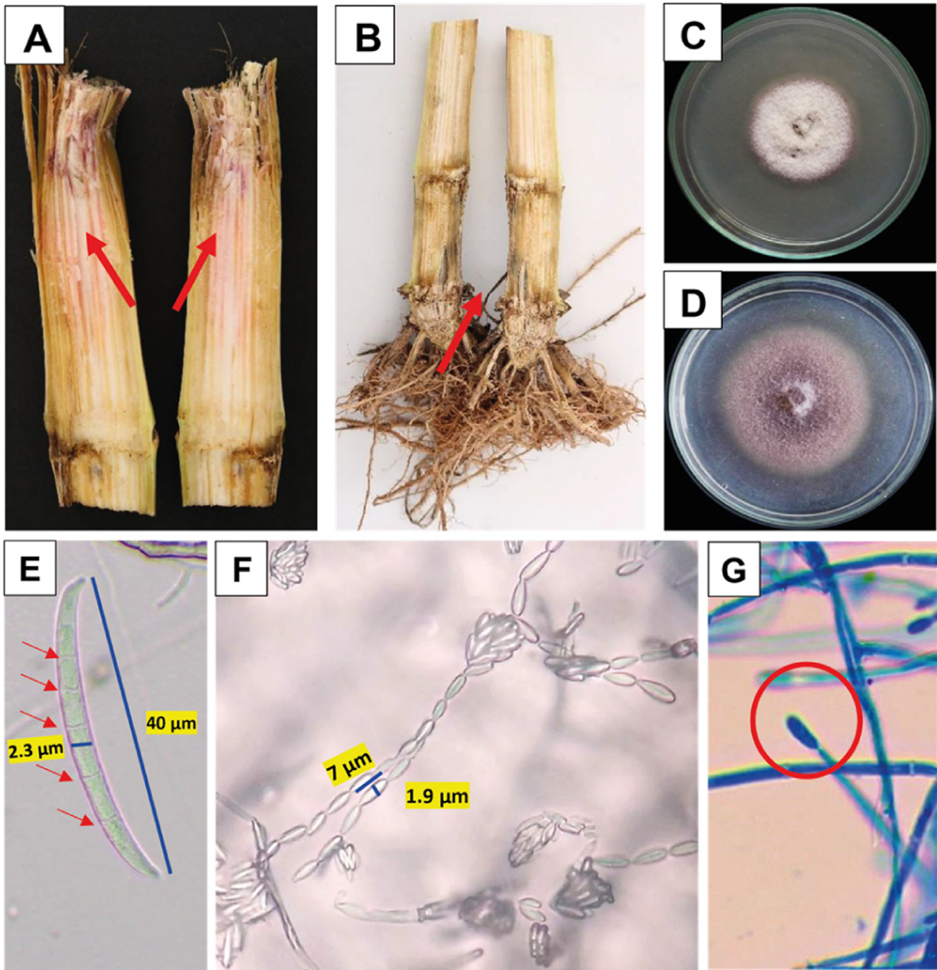

Fig. 1. A-B Maize stems and roots with vascular rot caused by *F. verticillioides*, the injured tissue was pink (red arrows). C. Colonies 6 days after inoculation on sterile petri dish. D. Colonies 10 days after inoculation. E. Macroconidia with 5 septa (red arrows). F. Microconidia in chains. *G. Conidiogenous* cell in monophyalid (red circle).

after sowing the maize seeds. Fig. 2 compares the reaction of the treatments during the experiment. Because it was necessary to define the behaviour of the pathogenicity with respect to the inoculum concentration, two doses were carried out. In both inoculation treatments, the germination reaction was reduced by the pathogen; and therefore, there was a lower number of maize seedlings, as the inoculum concentration increased.

3.4. Genomic survey

Low heterozygosity and repetition (0.7 %) were obtained and the estimated genome size (41.73 Mb) was close to the reported *F. verticillioides* genome references (7600: 41.9 Mb, HN2: 42.8 Mb, BRIP53590: 42.2 Mb, BRIP53263: 42.3 Mb, BRIP14953: 42.3 Mb, Fv10027_t1: 44.6 Mb, S1123A: 43.1 Mb and NRRL 20984: 41.9 Mb). (Fig. 3). Before quality control, a total of 23,102,402 reads were generated, which were reduced to 22,481,566 reads after trimming and filtering. The

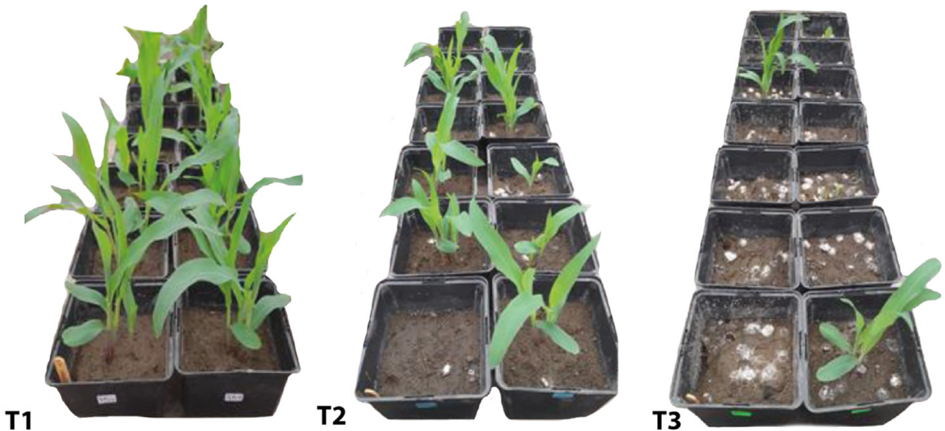

Fig. 2. Maize seedlings 15 days after germination in a substrate inoculated with *F. verticillioides* isolated from maize, under semi-controlled screen house conditions. T1 = 0 g.Kg^{-1}, T2 = 1 g.Kg^{-1} and T3 = 5 g.Kg^{-1}.

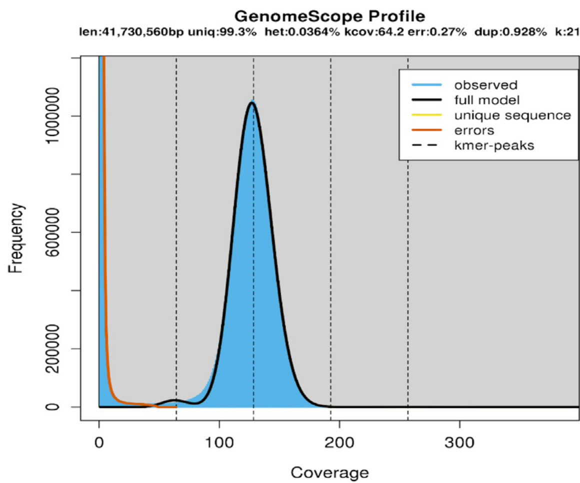

Fig. 3. Distribution of k-mers in the draft REC01 genome.

low heterozygosity and repetition rate suggest a relatively homogeneous genome with limited genetic variation. Furthermore, the genome size aligns closely with the reported sizes of other *F. verticillioides* genomes.

3.5. Assembly de novo*, reference-assisted scaffolding, and validation*

The scaffold assembled using MaSuRCA has a total length of 42.8 Mb, consisting of 42,796,516 contigs (≥1000 bp) with a GC content of 48.26 %. The longest contig was 1,769,758 bp (Table 2). Additionally, BUSCO was obtained: 4384 complete (S), 4377 simple copies, 7 complete and duplicates (D), 27 fragmented and 3 missing (Table 2).

However, when compared to other *F. verticillioides* assemblies at different levels—such as scaffold level (S1123A, NRRL20984), whole genome level (7600), chromosome level (HN2,

Table 2
Statistics of the completeness of the *de novo* assembly and summary of the BUSCO approach in the *F. verticillioides*.

Statistic	Value
N50	462380
N75	263483
L50	30
L75	75
Largest contig	1769758
Total length	42866930
GC (%)	48.26
# contigs (≥ 1000 bp)	42796516
# contigs (≥ 5000 bp)	42674077
# contigs (≥ 10,000 bp)	42549820
# contigs (≥ 25,000 bp)	42224523
# contigs (≥ 50,000 bp)	41728895
# N's per 100 kbp	2.51
Complete BUSCOs	4384
Complete and single-copy BUSCOs	4377
Complete and duplicated BUSCOs	7
Fragmented BUSCOs	27
Missing BUSCOs	83

BRIP53590, BRIP53263, and BRIP14953), and contig level (Fv10027_t1)—the assembly we obtained for REC01 had an N50 of 462 kb (Table 3).

Additionally, the assemblage has 97.55 % complete BUSCOs (C) (S: 97.4 % + D: 0.16 %), similar to the BRIP53590 type with 97.60 % C (S: 97.49 % + D:0.11 %) and the NRRL 20984 type with 97.64 % C (S: 97.33 % + D:0.31 %) (Fig. 4). The genome has been deposited in GenBank with the accession number JAWRVJ000000000.1.

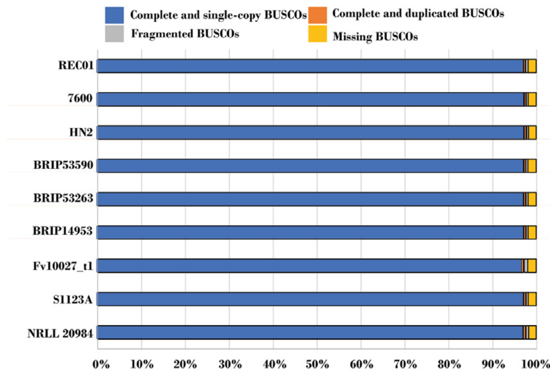

Fig. 4. Comparison of the BUSCO analysis of *F. verticillioides* REC01 with other types.

3.6. SSR data mining

The most abundant microsatellite motif type of REC01 was mononucleotide repeats, which accounted for 47.25 % of total Simple Sequence Repeats (SSRs), followed by dinucleotide repeats 18.98 %, trinucleotide 18.09 %, tetranucleotide 8.03 %, of pentanucleotides 4.34 % and hexanucleotides 3.31 %. Similar to the distribution of microsatellite motifs of other types 7600, HN2, BRIP53590, BRIP53263, BRIP14953, Fv10027_t1, S1123A and NRRL 20984 (Fig. 5).

A total of 3113 microsatellite loci were identified based on the assembled REC01 draft genome sequence, with a frequency of 72.62 SSR/Mb, which is almost the same as BRIP53263

Table 3
Comparison of the assembly of *F. verticillioides* REC01 with other types.

Type	*F. verticillioides*								
	7600	HN2	BRIP53590	BRIP53263	BRIP14953	Fv10027_t1	S1123A	NRRL 20984	REC01
Level Assembly	Genome complete	Chromosome	Chromosome	Chromosome	Chromosome	Contig	Scaffold	Scaffold	Scaffold
Total sequence length	41994356	42814391	42294396	42398840	4235067	44652197	43183040	41924634	42866930
Gaps per 100 kbp	0	0	177.56	183.5	189.26	0.02	6.09	1.88	2.51
Number of scaffolds	11	12	258	153	255	21	94	857	404
Scaffold N50	4 mb	4 mb	4 mb	4 mb	4 mb	2 mb	1 mb	104 kb	462 kb
Scaffold L50	5	5	5	5	5	5	10	118	30
Number of contigs	11	12	1009	931	1060	21	147	866	456
Contig N50	4 mb	4 mb	105 kb	101 kb	97 kb	2 mb	856 kb	104 kb	347 kb

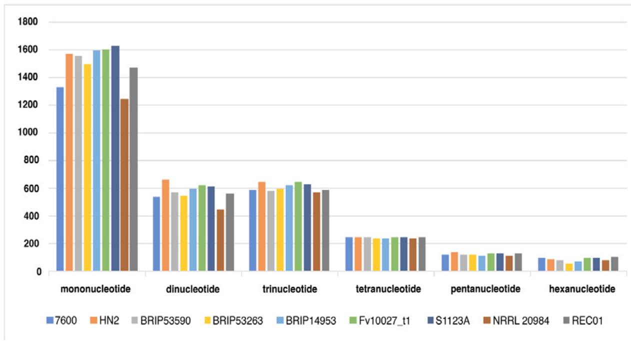

Fig. 5. Distribution of SSRs in types. Percentage of RSS by reason in REC01 compared to other types.

(72.17 SSR/Mb), lower than HN2 (3355 SSR/Mb), BRIP5359 (3146 SSR/Mb), BRIP14953 (3244 SSR/Mb), Fv100257_t1 (3339 SSR/Mb) and S1123A (3344 SSR/Mb) but higher than 7600 (2925 SSR/Mb), BRIP53263 (3060 SSR/Mb) and NRRL 20984 (2700 SSR/Mb) (Table 4). Furthermore, the number of SSRs present in the composite matrix of REC01 (202) is similar to the type of HN2 (203), greater than 7600 (162), Fv10027_t1 (194) and NRRL 20984 (165).

Table 4
Summary of RSS distribution in REC01 and other types.

F. verticillioides genomes									
Type	7600	HN2	BRIP 53590	BRIP 53263	BRIP 14953	Fv10027_t1	S1123A	NRRL 20984	REC01
Total number of identified SSRs	2925	3355	3146	3060	3244	3339	3344	2700	3113
Frequency (SSR/kb)	69.65	78.36	74.38	72.17	76.6	74.78	77.44	64.4	72.62
Number of SSRs present in compound formation	162	203	274	283	278	194	233	165	202

4. Experimental Design, Materials, and Methods

4.1. Microbiological identification

The microbiological analysis focused on the identification and characterization of the phytopathogen *F. verticillioides* from collected tissues of Peruvian maize stems that showed symptoms of rot. These fragments were subjected to an isolation and culture process in a PDA medium incubated at 28°C, allowing controlled growth and development between 6 and 10 days. In addition, a thorough microscopic examination was carried out to observe the reproductive structures, specifically macroconidia and microconidia.

4.2. Molecular identification

To confirm the presence of *F. verticillioides*, PCR was performed to amplify ITS1-5.8S-ITS2 region using the forward ITS-Fu-f (5′-CAACTCCCAAACCCCTGTGA-3′) and reverse ITS-Fu-r (5′) primers. -GCGACGATTACCAGTAACGA-3′) [5]. The sequence was first compared with the NCBI database using the blastn program [6], ensuring the identity of Fusarium verticillioides.

Subsequently, taxonomic classification was performed using massBLASTer (https://github.com/TU-NHM/massblaster_plutof_pub) available on the UNITE website (https://unite.ut.ee/analysis.php) v. 10.0, which offers a curated set of fungal ITS sequences [7]. The ITS1 region was extracted using tools from UNITE to ensure accurate classification, following methods similar to those described in previous studies [8,9]. This approach, independent of sequence alignment, is effective in reducing redundancy and ensuring consistency in taxonomic information [9].

4.3. Pathogenicity evaluation

4.3.1. Inoculation

The increase in inoculum of *F. verticillioides* REC01 was through the development of mycelium on wheat grains (previously cooked and autoclaved in polypropylene bags). The inoculated wheat was incubated at 25°C for 3 weeks; this was the methodology used at the Clínica de Diagnosis de Fitopatologia – Universidad Nacional Agraria La Molina [10].. The inoculation method was by incorporating the inoculated wheat grains into the substrate according to the doses established in treatments 2 and 3.

4.3.2. Statistical design

CRD was established with an absolute control (T1, substrate without inoculum), and 2 treatments with different levels of inoculum concentration: T2, 1 g/kg of substrate; and T3, 5 g/kg of substrate. 7 repetitions were established, each repetition with 2 pots; and each pot containing 3 seeds of hard yellow maize (Dekalb-7500). Analysis of variance and Tukey's comparison of means test were performed at an alpha of 0.05; using the statistical program Statistix version 9.0.

4.3.3. Parameters evaluated

After 15 days from sowing, the number of corn seedlings was quantified for each experiment as a percentage value. Additionally, the incidence of the characteristic sign of the pathogen in root medullary rot was measured 60 days after sowing. The development of necrosis in the internal part of the root was assessed as a percentage for each repetition.

4.4. Sample collection and DNA extraction

The strain isolated on the PDA medium was selected for the extraction of its genomic DNA; we used the E.Z.N.A. bacterial DNA isolation kit (Omega Bio-tek, USA) following the manufacturer's protocol. The genomic DNA was subjected to Illumina technology, specifically 150-bp paired-end (PE) sequencing, utilising the Illumina Nextera DNA Flex library preparation kit. The resulting PE Illumina library was loaded onto the NovaSeq 6000 instrument for cluster generation and subsequent sequencing, conducted by Novogene Co. Ltd. (CA, USA).

4.5. Genome sequencing and genomic survey

DNA sequencing of pair ends was performed using the Illumina HiSeq 2500 platform. Raw reads were verified using FastQC v.0.11.9 software [11]. In addition, read quality trimming (phred $Q > 25$) and adapter removal were performed using the Trimmomatic v0.36 [12] and TrimGalore v.0.6.7 [13] programs, respectively. For the purpose of genomic investigation, we employed the Jellyfish v.2.0 software [14]. Furthermore, the Genome Scope v1.0.0 tool [15] (Cold Spring Harbor Laboratory, Laurel Hollow, US) was utilised to derive estimations regarding key genome features, including genome size, repetitive content, and heterozygosity rate. These estimations were based on the output generated by Jellyfish, in conjunction with the utilisation of 17-mer for K-mer

analysis. To identify a uniform pattern characterised by a single peak in the K-mer frequency distribution analysis, the depth of K was estimated.

4.6. De novo *assembly and validation*

De novo assembly process was carried out using MaSuRCA v.4.0.6 [16]. Subsequently, assembly statistics were assessed using QUAST v.5.2.0 [17]. Furthermore, QUAST was employed with the scaffold output for further analysis. To evaluate the completeness of the genome assembly and identify any gene gaps, we employed the BUSCO v.5.4.2 [18] strategy utilising the *F. verticillioides*-specific profile.

4.7. Repeat annotation

To identify repetitive elements, we employed *de novo* and peer-based methods. The MISA Perl script [19] (http://pgrc.ipk-gatersleben.de/misa/) was used to identify SSRs within the REC01 genome. In addition, for SSR analysis, we incorporated genomes from various strains of *F. verticillioides*, namely 7600 (GenBank: GCA_027571605.1), HN2 (GenBank: GCA_026119585.1), BRIP53590 (GenBank: GCA_003316995.2), BRIP53263 (GenBank: GCA_003317015.2), BRIP14953 (GenBank: GCA_003316975.2), Fv10027_t1 (GenBank: GCA_020882315.1), S1123A (GenBank: GCA_025503005.1), and NRRL 20984 (GenBank: GCA_013759275.1). The BUSCO tool was employed to assess the assembly quality. To conduct a comparative analysis of BUSCO, we retrieved eight types (7600, HN2, BRIP53590, BRIP53263, BRIP14953, Fv10027_t1, S1123A, and NRRL 20984) from the NCBI database at the assembly level.

Limitations

Not applicable.

Ethics Statement

Work did not include animal experiments or data collected from social media platforms or human subjects.

CRediT Author Statement

Conceptualization, R.E., W.P., Y.R., L.A. and G.M.; methodology, R.E., G.M., E.V., L.A. and C.A.; software, R.E., and Y.R.; validation, C.A.; formal analysis, R.E. and Y.R.; research, R.E., W.P., Z.A. and C.A.; resources, Y.R. and C.A.; data curation, R.E., and Y.R.; writing: preparation of original draft, R.E., Y.R., W.P. and G.M.; writing: review and editing, W.P., L.A. and K.G.; visualisation, G.M., K.G. and E.V.; supervision, Z.A. and E.V.; project management, R.A. and J.C.; acquisition funding, C.A. and J.C. All authors have read and accepted the published version of the manuscript.

Data Availability

WGS of Fusarium verticillioides isolated from peruvian root corn (Original data) (NCBI).

Acknowledgments

We thank Ivan Ucharima for image editing and Linda Arteaga for her collaboration in the laboratory. C.I.A. thanks Vicerrectorado de Investigación of UNTRM.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A.A. Blacutt, S.E. Gold, K.A. Voss, M. Gao, A.E. Glenn, Fusarium verticillioides: advancements in understanding the toxicity, virulence, and niche adaptations of a model mycotoxigenic pathogen of maize, Phytopathology 108 (2018) 312–326, doi:10.1094/PHYTO-06-17-0203-RVW.

[2] N. Devanna, P.N. Achar, M.Y. Sreenivasa, Current perspectives of biocontrol agents for management of fusarium verticillioides and its fumonisin in cereals- a review, J. Fungi Basel Switz. 7 (2021) 776, doi:10.3390/jof7090776.

[3] K.E. Duncan, R.J. Howard, Biology of maize kernel infection by Fusarium verticillioides, Mol. Plant-Microbe Interactions®. 23 (2010) 6–16, doi:10.1094/MPMI-23-1-0006.

[4] M. Fallahi, H. Saremi, M. Javan-Nikkhah, S. Somma, M. Haidukowski, A.F. Logrieco, et al., Isolation, molecular identification and mycotoxin profile of fusarium species isolated from maize kernels in Iran, Toxins 11 (2019) 297, doi:10.3390/toxins11050297.

[5] T.J. White, T.D. Bruns, S.B. Lee, J.W. Taylor, M.A. Innis, D.H. Gelfand, J.J.W. Sninsky, in: PCR Protocols: A Guide to Methods and Applications, Academic Press, PCR Protocols, New York, 1990, pp. 315–322.

[6] C. Camacho, G. Coulouris, V. Avagyan, N. Ma, J. Papadopoulos, K. Bealer, et al., BLAST+: architecture and applications, BMC Bioinformatics 10 (2009) 421, doi:10.1186/1471-2105-10-421.

[7] E.R. French, Métodos de Investigación Fitopatológica, IICA, 1982.

[8] K. Abarenkov, R.H. Nilsson, K.-H. Larsson, A.F.S. Taylor, T.W. May, T.G. Frøslev, J. Pawlowska, B. Lindahl, K. Põldmaa, C. Truong, D. Vu, T. Hosoya, T. Niskanen, T. Piirmann, F. Ivanov, A. Zirk, M. Peterson, T.E. Cheeke, Y. Ishigami, A.T. Jansson, T.S. Jeppesen, E. Kristiansson, V. Mikryukov, J.T. Miller, R. Oono, F.J. Ossandon, J. Paupério, I. Saar, D. Schigel, A. Suija, L. Tedersoo, U. Kõljalg, The UNITE database for molecular identification and taxonomic communication of fungi and other eukaryotes: sequences, taxa and classifications reconsidered, Nucleic Acids Res. (2023), doi:10.1093/nar/gkad1039.

[9] V. Deshpande, Q. Wang, P. Greenfield, M. Charleston, A. Porras-Alfaro, C.R. Kuske, N. Tran-Dinh, Fungal identification using a Bayesian classifier and the Warcup training set of internal transcribed spacer sequences, Mycologia 108 (2016) 1–5, doi:10.3852/14-293.

[10] C. Xue, Y. Hao, X. Pu, C.R. Penton, Q. Wang, M. Zhao, J.M. Tiedje, Effect of LSU and ITS genetic markers and reference databases on analyses of fungal communities, Biol. Fertil. Soils. 55 (2019) 79–88, doi:10.1007/s00374-018-1331-4.

[11] J. Jurka, V.V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany, J. Walichiewicz, Repbase update, a database of eukaryotic repetitive elements, Cytogenet. Genome Res. 110 (2005) 462–467, doi:10.1159/000084979.

[12] A.M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data, Bioinformatics 30 (2014) 2114–2120, doi:10.1093/bioinformatics/btu170.

[13] M.S. Campbell, C. Holt, B. Moore, M. Yandell, Genome annotation and curation using MAKER and MAKER-P, Curr. Protoc. Bioinformatics. 48 (2014) 4.11.1–4.11.39, doi:10.1002/0471250953.bi0411s48.

[14] G. Marçais, C. Kingsford, A Fast, Lock-free approach for efficient parallel counting of occurrences of K-MERS, Bioinformatics 27 (2011) 764–770, doi:10.1093/bioinformatics/btr011.

[15] I. Korf, Gene finding in novel genomes, BMC Bioinformatics 5 (2004) 59.

[16] A.V. Zimin, G. Marçais, D. Puiu, M. Roberts, S.L. Salzberg, J.A. Yorke, The MaSuRCA genome assembler, Bioinformatics 29 (2013) 2669–2677, doi:10.1093/bioinformatics/btt476.

[17] A. Gurevich, V. Saveliev, N. Vyahhi, G. Tesler, QUAST: quality assessment tool for genome assemblies, Bioinformatics 29 (2013) 1072–1075, doi:10.1093/bioinformatics/btt086.

[18] F.A. Simão, R.M. Waterhouse, P. Ioannidis, E.V. Kriventseva, E.M. Zdobnov, BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs, Bioinformatics 31 (2015) 3210–3212, doi:10.1093/bioinformatics/btv351.

[19] S. Beier, T. Thiel, T. Münch, U. Scholz, M. Mascher, MISA-web: a web server for microsatellite prediction, Bioinformatics 33 (2017) 2583–2585, doi:10.1093/bioinformatics/btx198.