Title:

Genome sequence data reveal at least two distinct incursions of the tropical race 4 (TR4) variant of *Fusarium* wilt into South America

Authors:

Paula H. Reyes-Herrera 1†, Eliana Torres-Bedoya 1,2†, Diana Lopez-Alvarez 3†, Diana Burbano-David 1, Sandra L. Carmona 1, Daniel P. Bebber 2, David J. Studholme 2, Monica Betancourt 1, Mauricio Soto-Suarez 1,*

Affiliations:

1 Corporación Colombiana de Investigación Agropecuaria - Agrosavia. C.I Tibaitatá. Km 14 vía Mosquera-Bogotá, Cundinamarca, Colombia.

2 Biosciences, University of Exeter, Geoffrey Pope Building, Exeter, United Kingdom.

3 Universidad Nacional de Colombia, Sede Palmira. Facultad de Ciencias Agropecuarias, Departamento de Ciencias Biológicas.

* Correspondence: msoto@agrosavia.co

† These authors contributed equally to this work.

Abstract

The global banana industry is threatened by one of the most devastating diseases: Fusarium wilt (FWB). FWB is caused by the soil-borne fungus *Fusarium oxysporum* f. sp. *cubense* (*Foc*), which almost annihilated the banana production in the late 1950s. A new strain of *Foc*, known as tropical race 4 (TR4), attacks a wide range of banana varieties including Cavendish clones which are the source of 99% of banana exports. In 2019, *Foc* TR4 was reported in Colombia, and more recently (2021) in Peru. In this study, we sequenced three fungal isolates identified as *Foc* TR4 from La Guajira (Colombia) and compared them against 19 whole-genome sequences of *Foc* TR4 publicly available, including four genome sequences recently released from Peru. To understand the genetic relatedness of the Colombian *Foc* TR4 isolates and those from Peru, we conducted a phylogenetic analysis based on a

genome-wide set of single nucleotide polymorphisms (SNPs). Additionally, we compared the genomes of the 22 available *Foc* TR4 isolates looking for the presence-absence of gene polymorphisms and genomic regions. Our results reveal that (i) the Colombian and Peruvian isolates are genetically distant, which could be better explained by independent incursions of the pathogen to the continent, and (ii) there is a high correspondence between the genetic relatedness and geographic origin of *Foc* TR4. The profile of present/absent genes and the distribution of missing genomic regions showed a high correspondence to the clades recovered in the phylogenetic analysis, supporting the results obtained by SNP-based phylogeny.

Introduction

The global banana export industry generates around 14.5 billion USD per year (FAOSTAT 2020). Approximately 51% of global production is of the Cavendish cultivar (Lescot 2018). Latin America and the Caribbean (LAC) constitutes the world's most important exporting region for bananas. During 2018, the total production volume of bananas in LAC was estimated at 30 million tonnes and 13 million tonnes were exported to developed countries, mainly the United States of America and the European Union. In Colombia, banana is the country's third most important export crop after coffee and flowers. Approximately 91% of the national banana production is destined for export. Five departments concentrate almost 75% of the national production. In 2019, Antioquia was the department with the largest production of banana, with a cultivated area of 40,000 ha (42% of the total cultivated area of the nation), followed by Magdalena (14%), Nariño (9%), Valle del Cauca (7%) and La Guajira (3%) (Statistics home 2014).

Fusarium wilt of banana (FWB), one of the most devastating diseases of bananas, is threatening the global export industry. The disease, caused by the soil-borne fungus *Fusarium oxysporum* f. sp. *cubense* (*Foc*), was first reported in Australia in 1876 (Bancroft 1876). The devastation caused by F*oc* race 1 (R1) was mitigated in the late 1950s by the substituting Gros Michel with resistant Cavendish cultivars, which now account for nearly all banana exports. However, a newly emerging lineage of *Foc*, known as tropical race 4 (TR4) corresponding to vegetative compatibility group VCG 01213/16, attacks Cavendish clones and a wide range of other banana varieties. *Foc* TR4 was first detected in Taiwan in 1967 (Hwang and Ko 2004; Su et al. 1977), after most likely being introduced on infected plants from Sumatra, Indonesia, and subsequently spread widely in banana-producing countries (Ploetz et al. 2015; Ploetz 2015). *Foc* TR4 was restricted to Australasia from 1990 when it was formally recognized until 2013 when it was first reported in Jordan, and Mozambique (Butler 2013). In 2015 it emerged in Lebanon, Oman, India and Pakistan (Viljoen et al. 2020; Ordoñez et al. 2016; Thangavelu et al. 2019;

Ploetz et al. 2015). Between 2017 and 2019, *Foc* TR4 was found in Laos, Vietnam, Myanmar and Thailand (Ordoñez et al. 2016; García-Bastidas et al. 2014; Acuña et al. 2021; Chittarath et al. 2018; Hung et al. 2018; Zheng et al. 2018; Latest Pest Reports 2019). According to official information, *Foc* TR4 is currently confirmed in 22 countries (CABI ISC 2021), predominantly in South and Southeast Asia. In 2019, the pathogen was reported for the first time in Colombia, reaching out to the American continent (García-Bastidas et al. 2020; Viljoen et al. 2020), and more recently in Peru (Acuña et al. 2021). *Foc* TR4 was detected on a banana plantation in the northeastern region of La Guajira, Colombia. Currently, eleven farms in La Guajira and one in Magdalena are confirmed for the presence of *Foc* TR4. Consequently the 0.32 % of banana producing region in the country is under quarantine due to the presence *Foc* TR4.

Understanding the phylogenetic relationships among geographically disparate isolates can provide clues about a pathogen's chains of transmission. An understanding of the genetic diversity and relationships with other organisms is important for rational design of molecular assays for detection and identification. Taxonomy is important for implementation of control measures such as notification and quarantine.

Recently, the use of high-throughput genome-sequencing technologies has made important contributions to studies on *Foc* TR4, particularly on genetic diversity (Maymon et al. 2020; Ordonez et al. 2015) and phylogeographical analysis (Zheng et al. 2018). Ordoñez and colleagues (Ordoñez 2018; Ordonez et al. 2015) performed a hierarchical clustering analysis based on 4,298 DArTseq markers showing a limited genetic diversity between multiple *Foc* TR4 isolates from countries in the Middle East, Asia and Oceania (China, Indonesia, Jordan, Lebanon, Malaysia, Pakistan, Philippines and Australia). Genomic comparison of *Foc* TR4 isolates from the Greater Mekong subregion (Zheng et al. 2018) identified three geographically distinct clusters, one of which constituted isolates from Laos, Vietnam, Myanmar and China; this suggested that the source of infection in the Greater Mekong subregion probably originated from China. Most recently, a phylogeographical study conducted by Maymon and colleagues (Maymon et al. 2020), using SNPs across the whole genome, principal component analyses (PCA) and hierarchical clustering, claimed a strong similarity between the Colombian isolates and the Indonesian isolate II-5. The authors argued that this suggests that the pathogen most likely spread to Colombia from Indonesia (Maymon et al. 2020). A recent genomic comparison of Indian *Foc* isolates belonging to races 1, 2 and 4 revealed differences in the repertoire of *SIX* genes in Indian *Foc* TR4 compared with *Foc* TR4 isolated elsewhere (Raman et al. 2021).

Previous work (Maymon et al. 2020) suggested that the source of the Colombian *Foc* TR4 outbreak might be Indonesia. It is likely that the *Foc* TR4 lineage first arose in that part of the world (Bentley et

al. 1998; Vézina 2014) and in that sense, Indonesia might ultimately be the origin of all *Foc* TR4. However, it remains unclear as to what routes *Foc* TR4 has taken to disseminate across and between continents. Furthermore, misattribution of the pathogen's source may have important economic and political consequences; so, any such claims require careful scrutiny. Therefore, the aims of this study were (i) to evaluate the SNP genetic diversity between Colombian isolates and those from elsewhere, including genome sequences for three additional Colombian isolates, four from Peru and other sequence data that were not available at the time of the previous study (Maymon et al. 2020), and (ii) to identify new genomic differences/similarities between TR4 isolates able to support the SNP-based phylogeny. For this, we sequenced the genomes of the three additional *Foc* TR4 isolates using Oxford Nanopore Technologies' MinION platform. Our genomic comparison analyses revealed that outbreaks in Peru and Colombia are genetically distinct and likely have different origins. We generate a catalogue of genes/regions whose presence is variable among *Foc* TR4 pathogen individuals that will be useful resource for future study.

Results and Discussion

The *Foc* TR4 variant of Fusarium wilt has been detected in two South American countries, namely Peru and Colombia. This lineage of the pathogen likely emerged initially in Indonesia or Malaysia (Bentley et al. 1998; Vézina 2014) and has subsequently undergone a number of intercontinental transmission events. Each of these events presumably involved a founder population that represents a tiny sample of the pathogen population's genetic diversity, leading to a genetic bottleneck at each introduction to a new geographical location. Until recently, *Foc* TR4 was unknown in Latin America. This raises the question of whether recent outbreaks in the two South American countries are linked in a direct chain of transmission or whether they represent separate introductions. The 'bottleneck' effect of introduction from the *Foc* TR4 population in its centre of origin predicts that directly linked outbreaks would involve genetically similar pathogen individuals, whereas independent samples from the that original population would be relatively divergent from each other. We compared isolates from Colombia with previously sequenced isolates from elsewhere, using genome-wide sequence data to maximise the resolution of the genetic relationships.

Genome sequencing of Colombian *Foc* TR4 isolates

We generated approximately 17 Gb of long-read data for each Colombian *Foc* TR4 isolate. The sequence reads have been deposited in the Sequence Read Archive under the BioProject accession number PRJNA774343 (BioSample accession numbers SAMN22562322, SAMN22562323 and SAMN22562324). These genomic reads were aligned against the UK0001 *Foc* TR4 reference genome

sequence and the resulting alignments were used for calling SNPs for phylogenetic analysis and for surveying differentially present/absent genes.

Presence/absence of *SIX* genes in the Colombian *Foc* TR4 genomes

During the last ten years, several studies have developed molecular markers for detection of *Foc* TR4 (Matthews et al. 2020; Magdama et al. 2019; Ndayihanzamaso et al. 2020; Dita et al. 2010; Aguayo et al. 2017; Lin et al. 2013; Li et al. 2013; Carvalhais et al. 2019). This raises an important question: Are the existing molecular markers present in the Colombian *Foc* TR4 genomes? We investigated whether the secreted in xylem *SIX* genes (*SIX1* – S*IX13*) were present/absent in Colombian *Foc* TR4 isolates. For this, a BLAST analysis was carried out using *SIX* gene sequences against the 190098, 190203 and 03242 Colombian genome assemblies. All sequences were present in all three Colombian *Foc* TR4 genomes. *SIX*-gene homologues from *SIX1* to *SIX13* showed different levels of similarity, ranging from 98 to 100%, to those of our sequenced isolates (Supplementary Table S1). Interestingly, *SIX6* and *SIX9* genes presented low percentage similarities of 91 and 88%, respectively.

Colombian isolates constitute a distinct clade distinct from Peruvian isolates

We identified 671 single-nucleotide sites in the *Foc* TR4 genome that showed variation and for which the allele could be unambiguously determined in every one of the examined genomes. On the basis of these 671 SNPs (Supplementary File S2), we constructed the maximum-likelihood phylogeny shown in Figure 1. This phylogenetic tree displayed a clear correspondence between genetic relatedness and geographic origin. For example, there is a clade comprising isolates from the Middle East. Notably, isolates from South America are distributed among two distinct clades. The six isolates from Colombia comprise a single clade; similarly, the four isolates from Peru comprise another distinct clade. Colombian isolates are genetically much more distant from Peruvian isolates than they are from isolates collected in the Middle East and the United Kingdom. This phylogeographic pattern is not consistent with a single introduction of *Foc* TR4 from its centre of origin into South America and subsequent spread within the continent. Rather, it is better explained by separate, independent incursions into Colombia and Peru.

Our phylogenetic analysis revealed that the six Colombian isolates are much more closely related to each other than they are to any previously sequenced isolates from other geographic locations. In other words, there is greater genetic differentiation between than within geographical regions. Furthermore, there is no close relationship between Colombian isolates and isolates from elsewhere; Colombian isolates are genetically approximately equidistant to each of the other isolates. Therefore,

there is no genetic evidence of a direct source of the incursion of *Foc* TR4 into the Americas; the lineage comprising the Colombian isolates probably diverged from the lineages isolated elsewhere prior to the emergence of *Foc* TR4 outside of its centre of origin (whose location is unknown but likely to lie in Indonesia and/or Malaysia). The phylogenetic tree is consistent with several independent intercontinental transmissions of *Foc* TR4 from its origin. Similarly, within China and SE Asia, all sequenced isolates are closer to each other than to isolates from elsewhere, suggesting a single egress of *Foc* TR4 into that region. Similarly, most of the isolates from countries in the Middle East are genetically close and may have arisen from a single inoculum. In conclusion, the phylogenetic analysis is consistent with a single source of *Foc* TR4 into the Americas and shows no evidence that this is derived from the *Foc* TR4 populations seen in other regions where it has emerged.

Patterns of gene content are broadly consistent with phylogeny

Based on alignments of genomic reads from *Foc* TR4 isolates against the UK0001 reference genome sequence, we identified 615 gene presence-absence polymorphisms. These are tabulated as an Excel spreadsheet in Supplementary File S3. The distributions of these variable genes across the sequenced isolates were broadly consistent with phylogeny. Clustering of the genomes based on their profile of present/absent genes showed clusters that corresponded to the clades recovered in the phylogenetic analysis (Figure 2), revealing dozens of genes whose presence or absence is characteristic of specific *Foc* TR4 clades. This opens the future possibility of developing molecular typing assays to assign isolates to lineages without the need for whole-genome sequencing. Furthermore, it is notable that several genes show presence-absence polymorphism within the Colombian clade, suggesting that gene deletions have taken place very recently during the epidemic, some of which might be adaptive as the fungus finds itself in a new environment with new host genotypes. The biological significance of these differentially present genes is an avenue for future investigation.

Comparative genomic analysis reveals sequence gaps shared by *Foc* TR4 isolates from specific geographic locations

A comprehensive genomic analysis can identify components of the *Foc* TR4 genome that might complement the information provided by the SNP-based phylogeny. In this study, the whole genome assemblies were compared to the published genomes of *Foc* TR4 (Supplementary Table S4). We compared the genomes of 22 publicly available *Foc* TR4 isolates with the reference genome assembly of UK0001 (Figure S1). The comparison between the *de novo* assembled Colombian *Foc* TR4 shows

that they do not differ from each other. However, when comparing all *Foc* TR4 genomes, we observed several major gapped regions located on contigs VMNF0100005.1, VMNF0100007.1, VMNF01000013.1 and VMNF01000014.1 (Figure 3). Specifically, genomic regions that are present on the UK0001 reference genome are missing in (i) the Colombian *Foc* TR4 isolates (contigs VMNF0100005.1 andVMNF0100007.1); (ii) Middle Eastern isolates from Jordan, Israel and Lebanon (contigs VMNF01000013.1 and VMNF01000014.1); (iii) isolates from China, Vietnam, Myanmar, Laos and Peru (VMNF01000014.1). Thus, this comparative genomic analysis also supports the SNP-based phylogeny.

Conclusion

Phylogeographic relationships among *Foc* TR4 worldwide isolates are required to infer the routes of TR4's transmission between and within continents. In this work, we sequence and obtained a nearly complete genome assembly of three isolated of *Foc* TR4, obtained from banana plantations of la Guajira in Colombia. Our analysis suggests that Colombian isolates are more closely related to those from the UK and the Middle East and less to those from Perú. However, the divergences between those three lineages likely occurred prior to the global emergence of *Foc* TR4 outside its centre of origin, based on the genetic distances seen within geographically distinct lineages. Comparative genomic analysis revealed missing genomic regions that were shared between *Foc* TR4 isolates belonging to the same clade, also confirming a correspondence between genetic relatedness and geographic origin.

Materials and methods

Genome sequence data from public repositories

We used the previously published UK0001 genome (Warmington et al. 2019) as the reference sequence for alignment-based analyses as that genome sequence was assembled using long reads and is therefore of high quality. We used genome sequencing reads for previously reported *Foc* TR4 genome sequencing projects from the Sequence Read Archive (SRA) accession numbers SRR10103605, SRR10125423, SRR10747097, SRR9733598, SRR7226880, SRR10054450, SRR10054449, SRR7226881, SRR10054448, SRR10054446, SRR10054447, SRR7226882, SRR7226883, SRR7226879, SRR15514269, SRR15514270, SRR15514271, SRR15514272, SRR550155, SRR550152, SRR7226878 and SRR7226877 (Guo et al. 2014; Leinonen et al. 2011; Kodama et al. 2012; Warmington et al. 2019; Maymon et al. 2020; Acuña et al. 2021; Zheng et al. 2018).

Colombian *Foc* TR4 isolates

The *Fusarium oxysporum* f. sp. *cubense* TR4 isolates 190098, 03242 and 190203, isolated from symptomatic Cavendish banana from Dibulla, La Guajira state (Colombia) were provided by Instituto Agropecuario Colombiano (ICA). The isolates were maintained in potato dextrose agar (PDA) at 27 °C.

Extraction of genomic DNA

For high molecular weight (HMW) DNA extraction, the *Foc* TR4 isolates were grown on PDA plates for seven days at 28°C. Then, propagules were harvested and transferred to Czapek dox medium and incubated for six additional days to produce fungal mycelia. Fungal mycelium was obtained by filtering through two layers of Miracloth and washed twice with 10 mL sterile distilled water. Then, fungal mycelium was freeze-dried overnight and ground in a mortar with a pestle. Five hundred milligrams of ground mycelium were incubated for one hour at 65°C with 800 µL fresh DNA extraction buffer and 15 µl de RNAse (10mg/µl). DNA extraction buffer was prepared by mixing 2.5 volumes of solution A (0.35 M Sorbitol, 0.1 M Tris-base, 5 mM EDTA pH 7.5), 2.5 volumes of solution B (0.2 M Tris, 0.05 M EDTA, 2 M NaCl, 2% CTAB), 1 volume of Sarkosyl (10% w/v) and 1% β-mercaptoethanol. To separate the organic phase, 400 µL of phenol/chloroform/isoamyl alcohol (25:24:1) was added, vortexed for 5 minutes and incubated at room temperature (RT) for 5 minutes before centrifugation at 16,000 g for 15 min. Two chloroform extractions were used on the aqueous phase by adding 0.5 volumes and centrifuge at 16,000 g for 5 min at RT each time. The aqueous phase was mixed with 10 volumes of 100% ice-cold ethanol, incubated for 30 min at RT, and the precipitated DNA was collected in a new tube using a disposable inoculation loop. Collected HMW DNA was washed twice with 1 mL 70% ice-cold ethanol and the air-dried DNA was resuspended in nuclease-free water and conserved at 4°C. The DNA quality, size and quantity were assessed by spectrometry in a NanoDrop 2000 Spectrophotometer (Thermo Fisher Scientific, Wilmington, USA), electrophoresis in agarose gel and fluorometry in a Qubit Fluorometer v2.0 (Life Technologies, Thermo Fisher Scientific Inc.)

DNA sequencing

Sequencing library was prepared with the Ligation Sequencing Kit (SQK-LSK109) according to the manufacturer instructions (Oxford Nanopore Technologies, Oxford, UK) using 22 ng HMW DNA. An R9.4.1 flow cell (Oxford Nanopore Technologies, Oxford, UK) was loaded and run for 48 hours. Base calling was performed using Guppy from MinKNOW (version 4.0.21; Oxford Nanopore Technologies).

Genome assembly

For the *de novo* assembly of Oxford Nanopore Technologies data from three newly sequenced Colombian isolates, we used Canu 2.0 (Koren et al. 2017) and Flye 2.8.2 (Kolmogorov et al. 2019) followed by several iterations of Racon v1.4.3 (Vaser et al. 2017) and medaka 1.2.1 (Oxford Nanopore Technologies Ltd. 2018) to obtain a consensus sequence. The best assembly for each isolate was

selected according to most of the metrics such as BUSCO score (Simão et al. 2015), QUAST genome statistics (Gurevich et al. 2013) and Qualimap (García-Alcalde et al. 2012). In addition, BlobTools2 (Challis et al. 2020) as used to integrate to include coverage, BUSCO, Blast, and DIAMOND v2.0.11 (Buchfink et al. 2015) and do contaminant screening and genome assessment.

For most of the previously sequenced genomes, assemblies were available in the public databases. However, for several isolates, only unassembled sequence reads were available. Assemblies of isolates SRR10125423, SRR10747097, SRR550152, SRR550155, SRR7226877, SRR7226878, SRR7226879, SRR7226880, SRR7226881, SRR7226882, and SRR7226883 were performed using SPAdes and evaluated with Qualimap (García-Alcalde et al. 2012). For genomes with low numbers of reads (SRR10054447, SRR10054448, SRR10054449, SRR10054450, SRR10103605, SRR15514270, SRR15514271, and SRR15514272) we aligned the reads to the reference genome using Bowtie2 (Langmead and Salzberg 2012) and used the mpileup tool in Samtools (Li et al. 2009) to obtain a consensus sequence from the alignment. MUMmer4 was used to align the whole genome assemblies to the reference genome UK0001. Then, we used Circos to visualize the alignments in a circular representation.

Aligning sequence reads against reference genome sequence

To mitigate problems arising from incompleteness and errors in *de novo* assembly of short sequence reads, we used a genome-comparison strategy based on aligning sequencing reads against a high-quality reference genome sequence. For this assembly-free genomic comparison, we acquired short-read whole-genome Illumina sequence data as FastQ files (Cock et al. 2010) for *Fusarium oxysporum* f. sp. *cubense* from the SRA database (Kodama et al. 2012). The quality of the sequencing data was evaluated using FASTQC (Andrews n.d.). Reads with low quality or containing adaptor sequences were trimmed using Trim Galore (Babraham Bioinformatics - Trim Galore! 2022) or Canu (Koren et al. 2017) as appropriate to the sequencing method that generated the data. The sequences were aligned against the reference genome of isolate UK0001 (GenBank:GCA_007994515) using the Burrows Wheeler Aligner (BWA) (Li and Durbin 2010, 2009) and Minimap2 (Li 2018). The alignments were evaluated with Qualimap (García-Alcalde et al. 2012).

SNP-calling and phylogeny reconstruction

We used a genome-wide survey of SNPs towards understanding the relationship between Colombian *Foc* TR4 isolates and those from other geographical locations.

Single-nucleotide sites that showed sequence variability between *Foc* TR4 isolates (i.e. candidate SNPs) were identified using Pilon (Walker et al. 2014). There was some variation in the level of confidence in the nucleotide sequences at these candidate SNPs. For example, at some sites, there

was not a consensus among the multiple sequence reads aligned at that site. Therefore, candidate SNPs were filtered to retain only high-confidence SNPs with read-consensus above 95%, using a Perl script available at https://github.com/davidjstudholme/SNPsFromPileups. Full details of the command lines are given in Appendix 1 at the end of this document. Genomes were assigned to genetic types (haplotypes) according to the combination of nucleotide-states found at each of the high-confidence SNPs. These were represented as files in FastA and Nexus formats. A PhyML tree was generated in IQ-TREE (Nguyen et al. 2015) using the GTR model (Tavare 1986; Gouy et al. 2010; Guindon et al. 2010). The robustness of the phylogeny was assessed using 1000 bootstrap replicates.

Comparison of gene content

We identified genes in the UK0001 reference genome that were absent from one or more sequenced isolates (i.e. presence-absence polymorphisms) from the BWA alignments using the *coverageBed* tool in BEDtools (Quinlan and Hall 2010). This tool reports the breadth of coverage by aligned sequence reads for each gene. This approach, based on aligning reads against a reference genome, avoids problems arising from incompleteness of *de-novo* genome assemblies.

Author contributions

The study was conceived by MS-S and MB. The experiments were supervised by MS-S and MB. DNA extraction, library preparation and sequencing were conducted by SLC and DB-D. The analyses performed in the study were conceived by MS-S, DL-A, ET-B, DJS and PHR-H. Data analyses were completed by PHR-H, DL-A and ET-B. DPB and DJS supervised ET-B's bioinformatic analyses. PHR-H, DL-A, ET-B, MB, DPB, DJS and MS-S interpreted results. MS-S and DJS drafted a first version of this manuscript, edited by all other co-authors. All authors contributed to the article and approved the submitted version.

Funding

This work was supported by the earmarked fund to Agrosavia from Colombian Ministry of Agriculture, grant number (1000734: An update on Musaceae family diseases).

The Colombian *Foc* TR4 isolates were registered in the National Collections Registry (RNC129) and was collected under the AGROSAVIA´s permit framework No.1466 from 2014, updated by the 04039 resolution on July 19th, 2018.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement

The data that supports the findings of this study are available in the supplementary information of this article. Any additional data will be available on request to the corresponding author (msoto@agrosavia.co). Genome sequence data have been deposited in the Sequence Read Archive and GenBank and are available via BioProject accession number PRJNA774343 (BioSamples SAMN22562322, SAMN22562323 and SAMN22562324) and PRJNA731180 (BioSamples SAMN19572426, SAMN19275239 and SAMN19275177).

Conflicts of Interest

The authors declare no conflicts of interest.

References

Acuña, R., Rouard, M., Leiva, A. M., Marques, C., Olortegui, A., Ureta, C., et al. 2021. First report of Fusarium oxysporum f. sp. cubense Tropical Race 4, causing Fusarium wilt in Cavendish bananas in Peru. Plant Dis. 105:219.

Aguayo, J., Mostert, D., Fourrier-Jeandel, C., Cerf-Wendling, I., Hostachy, B., Viljoen, A., et al. 2017. Development of a hydrolysis probe-based real-time assay for the detection of tropical strains of Fusarium oxysporum f. sp. cubense race 4. PLoS One. 12:e0171767.

Andrews, S. FastQC A Quality Control tool for High Throughput Sequence Data.

Babraham Bioinformatics - Trim Galore! Available at: https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/ [Accessed January 8, 2022].

Bancroft, J. 1876. Report of the board appointed to enquire into the cause of disease affecting livestock and plants. Votes Proc. 3:1011–1038.

Bentley, S., Pegg, K. G., Moore, N. Y., Davis, R. D., and Buddenhagen, I. W. 1998. Genetic Variation Among Vegetative Compatibility Groups of Fusarium oxysporum f. sp. cubense Analyzed by DNA Fingerprinting. Phytopathology®. 88:1283–1293 Available at: https://apsjournals.apsnet.org/doi/10.1094/PHYTO.1998.88.12.1283.

Buchfink, B., Xie, C., and Huson, D. H. 2015. Fast and sensitive protein alignment using DIAMOND. Nat. Methods. 12:59–60.

Butler, D. 2013. Fungus threatens top banana. Nature. 504:195–196 Available at: http://www.nature.com/articles/504195a.

354 CABI ISC. 2021. Fusarium oxysporum f.sp. cubense tropical race 4 (Foc TR4). Invasive species
355 Compend. database, Retrieved from https//www.cabi.org/isc/tr4. :accessed 7 January 2022.

356 Carvalhais, L. C., Henderson, J., Rincon-Florez, V. A., O'Dwyer, C., Czislowski, E., Aitken, E. A. B., et al.
357 2019. Molecular diagnostics of banana Fusarium Wilt targeting secreted-in-xylem genes. Front. Plant
358 Sci. 10:547.

359 Challis, R., Richards, E., Rajan, J., Cochrane, G., and Blaxter, M. 2020. BlobToolKit–Interactive quality
360 assessment of genome assemblies. G3 Genes, Genomes, Genet. 10:1361–1374.

361 Chittarath, K., Mostert, D., Crew, K. S., Viljoen, A., Kong, G., Molina, A. B., et al. 2018. First Report of
362 Fusarium oxysporum f. sp. cubense Tropical Race 4 (VCG 01213/16) Associated with Cavendish
363 Bananas in Laos. Plant Dis. 102:449.

364 Cock, P. J. a, Fields, C. J., Goto, N., Heuer, M. L., and Rice, P. M. 2010. The Sanger FASTQ file format
365 for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res.
366 38:1767–71.

367 Dita, M. A., Waalwijk, C., Buddenhagen, I. W., Souza Jr, M. T., and Kema, G. H. J. 2010. A molecular
368 diagnostic for tropical race 4 of the banana fusarium wilt pathogen. Plant Pathol. 59:348–357.

369 FAOSTAT. Available at: https://www.fao.org/faostat/en/#data/QCL [Accessed January 8, 2022].

370 García-Alcalde, F., Okonechnikov, K., Carbonell, J., Cruz, L. M., Götz, S., Tarazona, S., et al. 2012.
371 Qualimap: evaluating next-generation sequencing alignment data. Bioinformatics. 28:2678–2679.

372 García-Bastidas, F. A., Quintero-Vargas, J. C., Ayala-Vasquez, M., Schermer, T., Seidl, M. F., Santos-
373 Paiva, M., et al. 2020. First report of Fusarium wilt Tropical Race 4 in Cavendish bananas caused by
374 Fusarium odoratissimum in Colombia. Plant Dis. 104:994.

375 García-Bastidas, F., Ordóñez, N., Konkol, J., Al-Qasim, M., Naser, Z., Abdelwali, M., et al. 2014. First
376 report of Fusarium oxysporum f. sp. cubense tropical race 4 associated with panama disease of
377 banana outside Southeast Asia. Plant Dis. 98:694.

378 Gouy, M., Guindon, S., and Gascuel, O. 2010. SeaView Version 4: A Multiplatform Graphical User
379 Interface for Sequence Alignment and Phylogenetic Tree Building. Mol. Biol. Evol. 27:221–224.

380 Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., and Gascuel, O. 2010. New
381 algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of
382 PhyML 3.0. Syst. Biol. 59:307–321.

383 Guo, L., Han, L., Yang, L., Zeng, H., Fan, D., Zhu, Y., et al. 2014. Genome and transcriptome analysis of

384 the fungal pathogen fusarium oxysporum f. Sp. Cubense causing banana vascular wilt disease. PLoS
385 One. 9.

386 Gurevich, A., Saveliev, V., Vyahhi, N., and Tesler, G. 2013. QUAST: quality assessment tool for
387 genome assemblies. Bioinformatics. 29:1072–1075.

388 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., and Vinh, L. S. 2018. UFBoot2: Improving
389 the Ultrafast Bootstrap Approximation. Mol. Biol. Evol. 35:518–522.

390 Hung, T. N., Hung, N. Q., Mostert, D., Viljoen, A., Chao, C. P., and Molina, A. B. 2018. First report of
391 Fusarium wilt on Cavendish bananas, caused by Fusarium oxysporum f. sp. cubense tropical race 4
392 (VCG 01213/16), in Vietnam. Plant Dis. 102:448.

393 Hwang, S.-C., and Ko, W.-H. 2004. Cavendish banana cultivars resistant to Fusarium wilt acquired
394 through somaclonal variation in Taiwan. Plant Dis. 88:580–588.

395 Kodama, Y., Shumway, M., and Leinonen, R. 2012. The sequence read archive: explosive growth of
396 sequencing data. Nucleic Acids Res. 40:D54–D56.

397 Kolmogorov, M., Yuan, J., Lin, Y., and Pevzner, P. A. 2019. Assembly of long, error-prone reads using
398 repeat graphs. Nat. Biotechnol. 37:540–546.

399 Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. 2017. Canu:
400 scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.
401 Genome Res. 27:722–736.

402 Langmead, B., and Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods.
403 9:357–359.

404 Latest Pest Reports - International Plant Protection Convention.

405 Leinonen, R., Sugawara, H., and Shumway, M. 2011. The Sequence Read Archive. Nucleic Acids Res.
406 39:D19–D21.

407 Lescot, T. 2018. Banane: Le fruit qui ne connait pas la crise? 256:92–96.

408 Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 34:3094–3100.

409 Li, H., and Durbin, R. 2010. Fast and accurate long-read alignment with Burrows-Wheeler transform.
410 Bioinformatics. 26:589–95.

411 Li, H., and Durbin, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform.
412 Bioinformatics. 25:1754–60.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. 2009. The sequence alignment/map format and SAMtools. Bioinformatics. 25:2078–2079.

Li, M., Shi, J., Xie, X., Leng, Y., Wang, H., Xi, P., et al. 2013. Identification and application of a unique genetic locus in diagnosis of Fusarium oxysporum f. sp. cubense tropical race 4. Can. J. Plant Pathol. 35:482–493.

Lin, Y.-H., Su, C.-C., Chao, C.-P., Chen, C.-Y., Chang, C.-J., Huang, J.-W., et al. 2013. A molecular diagnosis method using real-time PCR for quantification and detection of Fusarium oxysporum f. sp. cubense race 4. Eur. J. Plant Pathol. 135:395–405.

Magdama, F., Monserrate-Maggi, L., Serrano, L., Sosa, D., Geiser, D. M., and Jiménez-Gasco, M. del M. 2019. Comparative analysis uncovers the limitations of current molecular detection methods for Fusarium oxysporum f. sp. cubense race 4 strains. PLoS One. 14:e0222727.

Matthews, M. C., Mostert, D., Ndayihanzamaso, P., Rose, L. J., and Viljoen, A. 2020. Quantitative detection of economically important Fusarium oxysporum f. sp. cubense strains in Africa in plants, soil and water. PLoS One. 15:e0236110.

Maymon, M., Sela, N., Shpatz, U., Galpaz, N., and Freeman, S. 2020. The origin and current situation of Fusarium oxysporum f. sp. cubense tropical race 4 in Israel and the Middle East. Sci. Rep. 10:1–11.

Ndayihanzamaso, P., Karangwa, P., Mostert, D., Mahuku, G., Blomme, G., Beed, F., et al. 2020. The development of a multiplex PCR assay for the detection of Fusarium oxysporum f. sp. cubense lineage VI strains in East and Central Africa. Eur. J. Plant Pathol. 158:495–509.

Nguyen, L.-T., Schmidt, H. A., Von Haeseler, A., and Minh, B. Q. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol. Biol. Evol. 32:268–274.

Ordoñez, N. 2018. A global genetic diversity analysis of Fusarium oxysporum f.sp. cubense.

Ordoñez, N., García-Bastidas, F., Laghari, H. B., Akkary, M. Y., Harfouche, E. N., al Awar, B. N., et al. 2016. First Report of Fusarium oxysporum f. sp. cubense Tropical Race 4 Causing Panama Disease in Cavendish Bananas in Pakistan and Lebanon. Plant Dis. 100:209 Available at: https://apsjournals.apsnet.org/doi/10.1094/PDIS-12-14-1356-PDN.

Ordonez, N., Seidl, M. F., Waalwijk, C., Drenth, A., Kilian, A., Thomma, B. P. H. J., et al. 2015. Worse comes to worst: bananas and Panama disease—when plant and pathogen clones meet. PLoS Pathog. 11:e1005197.

Oxford Nanopore Technologies Ltd. 2018. Medaka. https://github.com/nanoporetech/medaka.

443 Ploetz, R. C. 2015. Fusarium Wilt of Banana. Phytopathology. 105:1512–1521 Available at:
444 http://apsjournals.apsnet.org/doi/10.1094/PHYTO-04-15-0101-RVW.

445 Ploetz, R., Freeman, S., Konkol, J., Al-Abed, A., Naser, Z., Shalan, K., et al. 2015. Tropical race 4 of
446 Panama disease in the Middle East. Phytoparasitica. 43:283–293.

447 Quinlan, A. R., and Hall, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic
448 features. Bioinformatics. 26:841–2.

449 Raman, T., Edwin Raj, E., Muthukathan, G., Loganathan, M., Periyasamy, P., Natesh, M., et al. 2021.
450 Comparative Whole-Genome Sequence Analyses of Fusarium Wilt Pathogen (Foc R1, STR4 and TR4)
451 Infecting Cavendish (AAA) Bananas in India, with a Special Emphasis on Pathogenicity Mechanisms. J.
452 Fungi. 7:717.

453 Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V, and Zdobnov, E. M. 2015. BUSCO:
454 assessing genome assembly and annotation completeness with single-copy orthologs.
455 Bioinformatics. 31:3210–3212.

456 Statistics home. 2014.

457 Su, H. J., Chuang, T. Y., and Kong, W. S. 1977. Physiological race of fusarial wilt fungus attacking
458 Cavendish banana of Taiwan. Taiwan Banan. Res. Inst. Spec. Publ. 2:1–21.

459 Tavare, S. 1986. Some probabilistic and statistical problems in the analysis of DNA sequences. Some
460 Math. Quest. Biol. / DNA Seq. Anal. Ed. by Robert M. Miura.

461 Thangavelu, R., Mostert, D., Gopi, M., Devi, P. G., Padmanaban, B., Molina, A. B., et al. 2019. First
462 detection of Fusarium oxysporum f. sp. cubense tropical race 4 (TR4) on Cavendish banana in India.
463 Eur. J. Plant Pathol. 154:777–786.

464 Vaser, R., Sović, I., Nagarajan, N., and Šikić, M. 2017. Fast and accurate de novo genome assembly
465 from long uncorrected reads. Genome Res. 27:737–746.

466 Vézina, A. 2014. The origin of TR4 : Under the peel | Improving the understanding of banana.
467 Available at: https://www.promusa.org/blogpost375-The-origin-of-TR4 [Accessed January 8, 2022].

468 Viljoen, A., Ma, L.-J., and Molina, A. B. 2020. CHAPTER 8: Fusarium Wilt (Panama Disease) and
469 Monoculture in Banana Production: Resurgence of a Century-Old Disease. In *Emerging Plant*
470 *Diseases and Global Food Security*, The American Phytopathological Society, p. 159–184. Available
471 at: https://apsjournals.apsnet.org/doi/10.1094/9780890546383.008.

472 Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., et al. 2014. Pilon: An

Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement ed. Junwen Wang. PLoS One. 9:e112963.

Warmington, R. J., Kay, W., Jeffries, A., O'Neill, P., Farbos, A., Moore, K., et al. 2019. High-quality draft genome sequence of the causal agent of the current Panama disease epidemic. Microbiol. Resour. Announc. 8:e00904-19.

Zheng, S.-J., García-Bastidas, F. A., Li, X., Zeng, L., Bai, T., Xu, S., et al. 2018. New Geographical Insights of the Latest Expansion of Fusarium oxysporum f.sp. cubense Tropical Race 4 Into the Greater Mekong Subregion. Front. Plant Sci. 9:1–9.

Figure Legends

Figure 1. Maximum-likelihood phylogeny of genome-sequenced Foc TR4 isolates from diverse geographical sources. The tree is based on 671 single-nucleotide polymorphisms and built using IQ-Tree (Nguyen et al. 2015). We obtained branch supports with the ultrafast bootstrap (Hoang et al. 2018).

Figure 2. Heatmap showing a comparison of gene content of Foc TR4 genomes. The rows represent 615 predicted protein-coding genes in the UK0001 reference genome that are absent from at least one TR4 genome; that is, differentially present/absent genes. The presence or absence of each gene was assessed in each genome based on breadth of coverage by genomic sequencing reads, using the coverageBed tool (Quinlan and Hall 2010); this generates a value between zero (no reads, i.e. absent) and one (completely covered by reads, i.e. present). The columns (i.e. genomes) are ordered according to by complete linkage clustering. Rows (i.e. genes) are ordered according to genomic location.

Figure 3. Comparative genomic analysis of 22 *Foc* TR4 genomes. Circle map showing contigs VMNF0100005.1, VMNF0100007.1, VMNF01000013.1 and VMNF01000014.1. The dotted rectangles indicate the missing genomic regions.

Figures

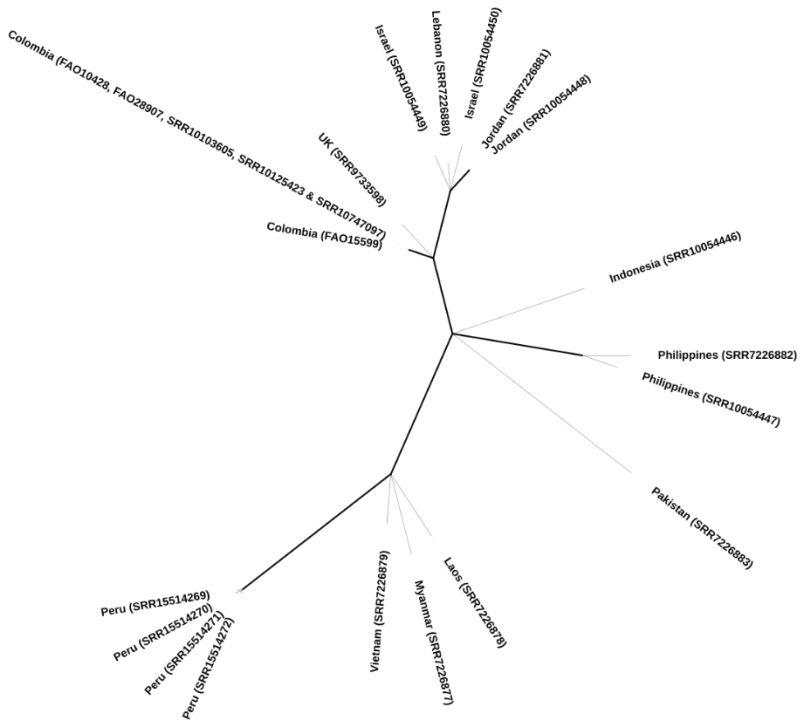


Figure 1.

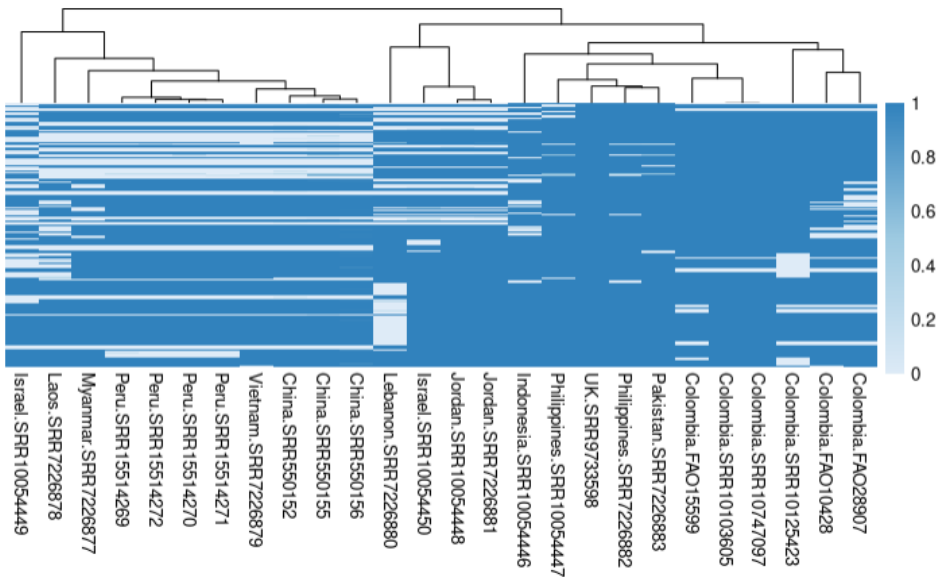
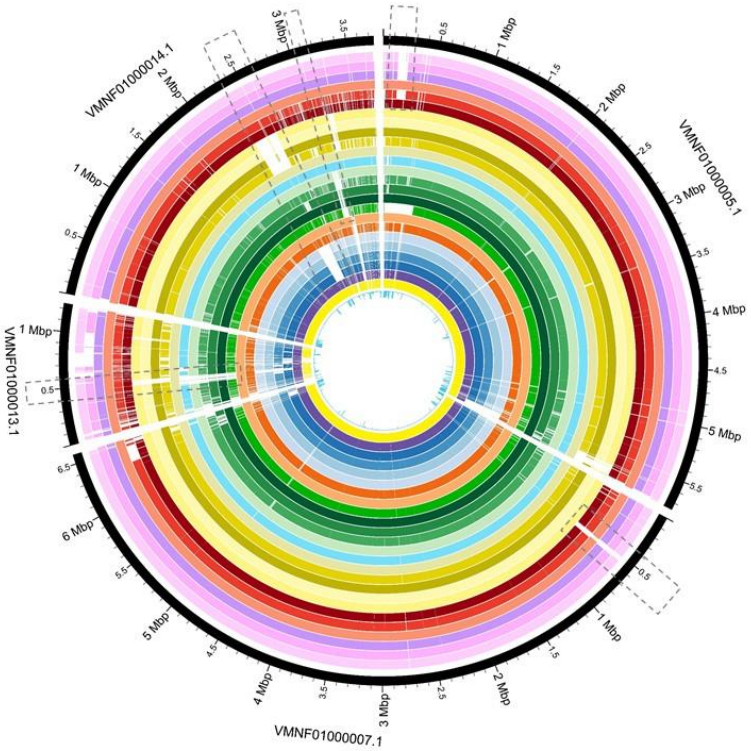


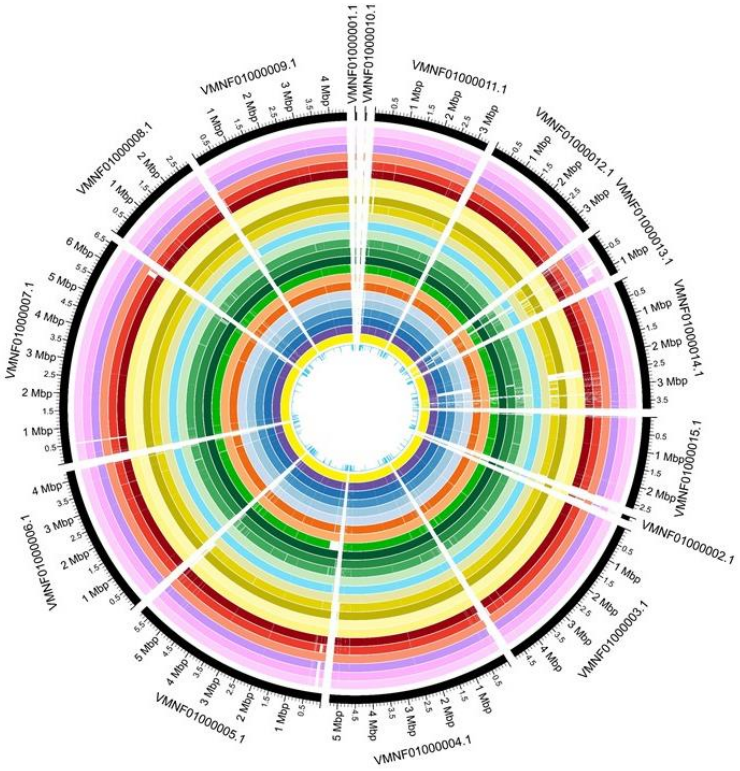
Figure 2.



Genome	Sample name	Country
FAO10428	FocR4T_190098	Colombia
FAO28907	FocR4T_03242	Colombia
FAO15599	FocR4T_190203	Colombia
SRR10103605	TR4Col	Colombia
SRR10125423	TR4Col4	Colombia
SRR10747097	TR4Col17	Colombia
SRR15514270	PerS3	Peru
SRR15514271	PerS2	Peru
SRR15514272	PerS1	Peru
SRR15514269	PerS4	Peru
SRR9733598	UK_0001	UK
CRR049239	FocTR4_58	China
SRR10054448	JV14	Jordan
SRR7226881	JV11	Jordan
SRR10054449	TR4-5 IL	Israel
SRR10054450	TR4-1 IL	Israel
SRR7226880	Leb1-2C	Lebanon
SRR10054447	S1B8	Philippines
SRR7226882	Phi2-6C	Philippines
SRR550152	Foc TR4	China
SRR7226879	Vn-2	Vietnam
SRR7226877	My-1	Myanmar
SRR7226878	La-2	Laos
SRR7226883	Pak1.1A	Pakistan
SRR10054446	II-5	Indonesia

Figure 3.

Supplementary materials



Genome	Sample name	Country
FAO10428	FocR4T_190098	Colombia
FAO28907	FocR4T_03242	Colombia
FAO15599	FocR4T_190203	Colombia
SRR10103605	TR4Col	Colombia
SRR10125423	TR4Col4	Colombia
SRR10747097	TR4Col17	Colombia
SRR15514270	PerS3	Peru
SRR15514271	PerS2	Peru
SRR15514272	PerS1	Peru
SRR15514269	PerS4	Peru
SRR9733598	UK_0001	UK
CRR049239	FocTR4_58	China
SRR10054448	JV14	Jordan
SRR7226881	JV11	Jordan
SRR10054449	TR4-5 IL	Israel
SRR10054450	TR4-1 IL	Israel
SRR7226880	Leb1-2C	Lebanon
SRR10054447	S1B8	Philippines
SRR7226882	Phi2-6C	Philippines
SRR550152	Foc TR4	China
SRR7226879	Vn-2	Vietnam
SRR7226877	My-1	Myanmar
SRR7226878	La-2	Laos
SRR7226883	Pak1.1A	Pakistan
SRR10054446	II-5	Indonesia

Supplementary Figure 1. Circos plot for all the contigs alignment between the 22 *Foc* TR4 genomes and the Reference isolate UK0001 used as reference.

Supplementary Table S1. Summary of BLAST analysis using *SIX* genes against assembled *Foc* TR4 isolates of Colombian origin.

Supplementary File S2. Tabulated Excel spreadsheet containing the 671 SNPs 671 single-nucleotide sites in the *Foc* TR4 genome that showed variation.

Supplementary File S3. Tabulated Excel spreadsheet containing the 615 gene presence-absence polymorphisms identified after comparing genomic reads from *Foc* TR4 isolates against the UK0001 reference genome sequence.

Supplementary Table S4. Information on three newly assembled and 19 publicly available *Foc* TR4 genomes analysed in this study.

Appendix 1: Command lines used for bioinformatics analysis

Gene presence/absence polymorphisms, SNPs and phylogeny of TR4 genomes

Download the reference genome sequence

```
wget --no-clobber
https://ftp.ncbi.nlm.nih.gov/genomes/all/GCA/007/994/515/GCA_007994515.1_ASM799451v1/GCA_00799
4515.1_ASM799451v1_genomic.fna.gz

gunzip GCA_007994515.1_ASM799451v1_genomic.fna.gz
```

Generate the pileup files from BAM files

```
for alignmentFile in SRR*.bam
  do
    echo $alignmentFile
    samtools mpileup -q 1 -f GCA_007994515.1_ASM799451v1_genomic.fna $alignmentFile >
$alignmentFile.pileup

for alignmentFile in FAO*.bam
  do
    echo $alignmentFile
    samtools mpileup -f GCA_007994515.1_ASM799451v1_genomic.fna $alignmentFile >
$alignmentFile.pileup
  done
```

Index the BAM files

```
for alignmentFile in FAO*.bam
  do
    echo $i
    samtools index $alignmentFile
  done
```

Call SNPs

Identify candidate SNP sites using Pilon

```
pilon --version

 for alignmentFile in *.bam
  do
    echo $alignmentFile
    pilon --genome GCA_007994515.1_ASM799451v1_genomic.fna --bam $alignmentFile --output
pilon_$alignmentFile --vcf
    done
```

Filter the candidate SNPs

```
for alignmentFile in *.bam
  do
    echo $alignmentFile
    bcftools filter --include '(REF="A" | REF="C" | REF="G" | REF="T") & (ALT="A" | ALT="C" |
ALT="G" | ALT="T")' pilon_$alignmentFile.vcf > $alignmentFile.filtered.vcf
    done
```

Get the SNP-calling scripts from GitHub

```
git clone https://github.com/davidjstudholme/SNPsFromPileups.git
```

Perform SNP-calling from pileup files.

To minimise memory usage, we only consider candidate sites previously identified using Pilon.

```
rm snps.csv*

perl SNPsFromPileups/get_snps_from_pileups_small_genome.pl 10 *.filtered.vcf FAO10428
SRR10054449 SRR15514270  SRR7226877 SRR7226883 FAO15599     SRR10054450 SRR15514271
```

```
SRR7226878 SRR9733598 FAO28907     SRR10103605 SRR15514272  SRR7226879 SRR10054446
SRR10125423 SRR7226880 SRR10054447  SRR10747097 SRR7226881 SRR10054448 SRR15514269 SRR7226882
> snps.csv

perl SNPsFromPileups/get_snps_from_pileups_small_genome.pl 10 *.filtered.vcf *.pileup  > snps-
all-pileups.csv
```

Convert the SNPs into Nexus format for input into IQ-Tree

```
perl SNPsFromPileups/get_haplotypes_and_aligned_fasta_from_csv.pl snps.csv
```

Perform phylogenetic analysis using IQ-Tree

```
/mnt/bio-tarako-home/djs217/iqtree-2.0.6-Linux/bin/iqtree2 --version
/mnt/bio-tarako-home/djs217/iqtree-2.0.6-Linux/bin/iqtree2 -s snps.csv.haplotype.nex -m
GTR+ASC
```

Perform bootstrapping

```
/mnt/bio-tarako-home/djs217/iqtree-2.0.6-Linux/bin/iqtree2 -nt AUTO -s
snps.csv.haplotype.nex.uniqueseq.phy -m TIM2+I+G -bb 1000
```

Examine gene content using coverageBed from Bedtools

```
coverageBed --help
for alignmentFile in *.bam
  do echo $alignmentFile
  if [ -s $alignmentFile.coverageBed.csv ]
    then
      echo $alignmentFile.coverageBed.csv already exists
    else
    coverageBed -a GCA_007994515.1_ASM799451v1_genomic.gff -b $alignmentFile >
$alignmentFile.coverageBed.csv
    fi
  done

./SNPsFromPileups/compare_coverages.pl GCA_007994515.1_ASM799451v1_genomic.gff
*.coverageBed.csv > comparison.csv
```

Plot the variable genes as a heatmap

Install the packages

```
install.packages('pheatmap')
```

Load the packages

```
library('pheatmap')
library(RColorBrewer)
library(tidyverse)
```

Make the plot

```
### Read the tab-delimited file that tabulates the variable genes/proteins
x <- read.table("variable_proteins_05_01_22.txt", header = T, stringsAsFactors = FALSE, quote
= "", sep = "\t")
row.names(x)<-paste(x$Gene)

### How many rowsand how many columns?
ncol<-ncol(x)
nrow<-nrow(x)

### Discard some columns
cols_remove <- c("Location", "Protein.domains.features", "Gene", "SignalP.v5", "X")
x <- x[, !(colnames(x) %in% cols_remove)]

### Format the data to be acceptable to pheatmap
```

```
x<-data.matrix(x)

### Plot the heatmap
hmcol<-colorRampPalette(brewer.pal(1,"Blues"))(256)

pheatmap(x,
col=hmcol,
display_numbers=F,
#clustering_distance_rows = "correlation",
#clustering_distance_cols = "correlation",
#clustering_method = "complete",
cluster_cols = T,
cluster_rows = F,
show_rownames=F
)
```