Analysis of telomeric and subtelomeric regions in the *Fusarium oxysporum* species complex.

Rikke Breemer
1426281.
Wageningen University and research.
Plant science group
Laboratory of Phytopathology.

Abstract.

Telomeres and subtelomeres are key genomic regions that play a critical role in chromosomal stability, integrity, and adaptability. These regions are often characterized by high sequence variability and are enriched in repetitive and transposable elements. In some pathogenic fungi, subtelomeric regions have been shown to contribute to niche adaptation and host specialization. Due to the repetitive nature of both telomeres and subtelomeres, they remain challenging to assemble, particularly with short-read sequencing technologies. In this study, we analysed 64 high-quality Fusarium oxysporum genome assemblies to systematically identify telomeric regions using well-established tandem repeat detection algorithms. We identified certain repetitive elements overrepresented in subtelomeric regions, which we term subtelomere-associated repeats. We assessed their conservation within and between different *F. oxysporum* strains, revealing that most strains contain conserved versions among all their subtelomeric regions. Most identified subtelomere-associated repeats share little similarity between strains, however some strains belonging to the same *formae specialis* exhibit striking similarities in these sequences. Additionally, these STARs can be used to identify putative chromosome ends and aid in assembly. These findings provide a foundation for further characterization of subtelomeric regions and their potential role in *Fusarium oxysporum* genome evolution.

Introduction.

Telomeres: Alternative methods of chromosomal rearrangement and potential source of adaptability.

Every eukaryotic chromosome consists of a linear DNA molecule capped by chromosomal termini known as telomeres. In most organisms they are characterized by short tandem repetitive elements which are conserved across most eukaryotes. The repeat consist of $(TTAGGG)_n$ in most fungi and vertebrates[1] and $(TTTAGG)_n$ in *Arabidopsis thaliana*[2]. In cell divisions, normally these sequences shorten as a result of incomplete telomere elongation. A process counteracted by the ribonucleoprotein complex called telomerase. Aside from using telomerase, a variety of other mechanisms have been proposed for telomere maintenance. All of these revolve around structural rearrangements, namely hair pin structures, T-loops, G-quadruplexes, extra chromosomal DNA break induced replication (BIR) and retrotransposons[3–5]. These mechanisms potentially serve a role as well in fungal genome structure and adaptation.

Subtelomeres are regions separating telomeric sequences from chromosome specific sequences and have been shown to vary greatly in length between different organisms ranging from 500kb in humans, 100kb in fission yeast and 10kb in budding yeast[6]. Additionally, they are shown to be the highest variable region of the genome[7]. Because these regions are highly dynamic, often undergoing spontaneous rearrangement and experiencing accelerated mutation, they seem to harbor sets of genes that are involved in niche adaptation. For example, in several pathogens like *Plasmodium spp*[8,9]*, Trypanosoma brucei*[10] and *Pneumocytes carinii*[11] subtelomeric clustering of large families of glycoprotein-encoding genes has been observed which aid in host infection[12].

Transposable elements (TEs) are highly diverse mobile genetic elements. The abundance of TEs is a primary determinant of genome size, and they are responsible for promoting genome change[13–16]. One mechanism by which this is achieved is gene inactivation by transposon insertion[17–19]. Rahnama *et al*. (2020) previously reported that the telomeres of the pathogenic fungi *Magnaporthe oryzea,* responsible for blast disease and leaf spot disease in a variety of crops has unusually polymorphic telomeres in strains infecting perennial ryegrass when compared to strains infecting rice[17]. Further research reveals that this high degree of polymorphism is associated with the presence of non-LTR retrotransposons embedded within the telomere repeats. These retrotransposons termed MoTeRs (*Magnaporthe oryzae* Telomeric Retrotransposons) have been shown to generate fragile sites at the ends of chromosomes which promote a variety of repair driven translocation/duplication events[20]. They hypothesize a mechanism by which the MoTeR invasion of telomeres causes genes in these regions to experience accelerated evolution and have increased potential for neo-functionalization. In addition, these newly formed genes have a way to be integrated into more stable sections of the genome through terminalization events[21]. This highlights the importance of studying telomeres and comparisons of telomeres both within and between strains to understand their role in genome evolution and adaptation.

Improvement of genome assemblies through long-read technology allows an unprecedented view on telomeres and subtelomeres.

Despite their role in genome stability and the fact that genes near chromosome ends have been shown to be implicated in host adaptation in different pathogens, telomeres and subtelomeres are not well-studied outside model organisms. Due to their repetitive nature both telomeric and subtelomeric regions are difficult to assemble and therefore often missing from genome assemblies[22]. These challenges are partly overcome by the use of long-read sequencing technology.

In Pacific Biosciences (PacBio) High-Fidelity (HiFi) sequencing each DNA molecule is sequenced multiple times in a circular manner creating a highly accurate consensus sequence ranging from one to 25Kb in length[23]. In contrast Oxford Nanopore sequencing relies on a single-stranded DNA molecule passing through a nanopore channel, during which the difference in electrical current between the inner and outer sections of the channel is used as a measurement and compared to a collection of known ionic currents using a deep neural network[24]. Oxford Nanopore reads are known to have less coverage at chromosomal ends and contain more sequencing errors than PacBio HiFi reads, especially with older flow cell

chemistry[25]. Tan *et al.* (2022)[26] report in their analysis of the telomere to telomere assembled human X chromosome, that the canonical telomere sequence: $(TTAGGG)_n$ is frequently recorded as $(TTAAAA)_n$. In a similar manner, the reverse complement sequence $(CCCTAA)_n$ is frequently recorded as either $(CTTCTT)_n$ or $(CCCTGG)_n$. This is hypothesized to be due to similarity in current profiles causing telomeric hexamers to be systematically wrongly called in raw nanopore reads. Despite the disadvantages mentioned before, nanopore sequencing is a more cost-effective method. This means that further research into the telomeric and sub telomeric regions becomes more accessible and would allow for a thus far unparalleled ability to explore the diversity of these regions, both within the genome and between different species.

Identifying subtelomeres in *Fusarium oxysporum* might aid in detecting horizontal chromosome transfer events.

The *Fusarium oxysporum (Fo)* species complex is comprised of soil-borne fungi found in cultivated and uncultivated soils worldwide under various climates. The complex consists of both pathogenic as well as many non-pathogenic strains which are morphologically indistinguishable from each other[27]. Plant-pathogenic strains can cause both wilt and root/crown rot on many economically important crops. They infect monocots and dicots, perennial and annual plants, land-based and aquatic[28]. A few *Fo* strains are known opportunistic pathogens in humans that can cause severe systemic infection, especially in immunocompromised patients[29]. Infections in mice[30] and caterpillars[31] are also reported. Individual *Fo* strains display pathogenicity towards a narrow host range of plants[32]. Strains with the same host range are grouped together in *formae speciales (f. sp.)*[27]. The fact that strains that infect the same host occur in different phylogenetic clades suggest that preference for a specific host has arisen multiple times[33].

The *Fo* genome is divided into a set of 11 gene-rich, indispensable core chromosomes that are generally conserved between all *Fusarium* strains/isolates, and one or more accessory chromosomes that are present in a subset of strains/isolates. Accessory chromosomes can into lineage specific (LS) chromosomes that are specific to clonal lines, and pathogenicity chromosomes that are present in all strains infecting a certain host and absent in others. In tomato-infecting *Fo* strains *(Fo f. sp. lycopersici (Fol)*) some of these accessory chromosomes have been shown to be sufficient for pathogenicity: horizontal chromosomal (HCT) transfer to a non-pathogenic Fo47 strain induces pathogenicity in tomato[34]. Similarly, in *F. oxysporum f. sp. melonis* (*Fom)* transfer of the Secreted in xylem 6 (SIX6) effector containing chromosome to a non-pathogenic strain causes pathogenicity to melon[35], and in *Fusarium oxysporum f. sp. conglutinans (Foc)* the transfer of a pathogenicity chromosome containing effector genes to a non-pathogenic strain has been shown to induce virulence in Arabidopsis and cabbage[36]. The presence of transposon potentially leads to genetic instability, driving evolution at an accelerated rate. This is made more likely by the LS regions in *Fol* genome where 74% of TEs, including 95% of all DNA transposons are clustered[37]. Additional studies have proposed the theory that transposon presence on pathogenicity chromosomes can play a role in fast adaptation to host resistance[38,39] most likely due to the variety of ways they promote chromosomal rearrangements. Alternative methods of transposon-driven genomic alterations are translocations, deletions, segmental duplications and inversions[40–43]. These can be a result of aberrant transposition, or through ectopic recombination between dispersed transposon copies and thus lead to a source of instability and a driving force in evolution and adaptation.

Due to their high repeat content and the presence of large segmental duplications[44] *de novo* Illumina assembly of *Fo* accessory chromosomes as well as telomeres remains challenging. Typically, they are dispersed over many contigs or scaffolds. This problem can partially be solved by the usage of long-read sequencing technology which would allow more accurate assembly of these regions. As seen in *Magnaporthe oryzea* variation in subtelomeric regions exists between different isolates. If this is the case in *Fo* comparing subtelomeric regions within and between strains might aid in the detection of HCT events or improve genome assemblies.

With the advance in long-read sequencing technology, multiple high quality *Fo* genomes have been assembled which show a wide range of transposable elements present[45]. Some families of transposons are potentially overrepresented in subtelomeric regions of the genome (L. Fokkens, Unpublished data). Although much research focus has been focused on the role of transposons in effector turnover, their presence in subtelomeric regions could have an effect on genomic organization. This research will focus on the following aspects.

- Identifying telomeres and sub telomeres across multiple *fusarium oxysporum* lineages.
- Identifying transposon families or repetitive elements that are overrepresented in subtelomeric regions across multiple *fusarium oxysporum* lineages.
- Compare this association both in high quality and fragmented genome assemblies.
- Identifying the presence of wrongly base-called telomeric repeats in the analyzed genome assemblies.

Together these research questions will provide the basis for further analysis of subtelomeric regions in *Fusarium oxysporum*.

Materials and Methods.

Compiling a dataset of high-quality, contiguous *Fo* genome assemblies.

A total of 50 high quality *Fusarium oxysporum* genome assemblies, with a N50 > 1.5 Mb were downloaded from Genbank (release 255.0)[46] between May and August 2023. In addition, we added 14 high-quality genome assemblies that were sequenced and assembled in house, arriving at a total of 64 assemblies in the dataset. An overview of each assembly can be found in supplementary file 1. An overview of the analysis performed is shown in figure 1.

Identification of telomeres and overrepresented repeat families.

In each assembly, *de novo* transposable element (TE) families were identified with RepeatModeler2[47] (version -2.0.3) with default settings and -LTRStruct to identify long terminal repeats separately. The consensus sequences for each predicted TE family were then used as input in the Earl Grey[48] pipeline (version -4.0.3-0). This pipeline applies a modified implementation of the "BLAST, extend, extract" process described by Platt *et al*[49]. termed "Blast, Extract, Align, Trim" (BEAT) to each *de novo* TE library. Telomeres are tandem repeats: multiple, adjacent copies of (approximately) the same nucleotide sequence. To identify telomeres in the genome assemblies, we use output of tandem repeat finder[50] (TRF), integrated into the EarlGrey pipeline, which identifies tandem repeats by comparing similarities (percent identity and indels) between adjacent pattern copies within a window with those predicted by a stochastic model. For each assembly the EarlGrey GFF output is used and for each identified repeat or putative TE family in this output the following information is stored: Contig id, repeat info, starting position, end position, attributes and unique repeat identifier. Because repeats or putative TE families can be present multiple times in each assembly the results are aggregated per unique repeat. We search for tandem repeats identified by TRF that correspond to any of the circular permutations of the *Fo* telomeric repeat and its reverse complement (CCCTAA, CCTAAC, CTAACC, etc.) in the first and last 100bp of each contig. When found these were considered telomeres. The regions 1000bp up or downstream were considered subtelomeric. Potentially wrongly basecalled telomeres are identified in the same manner as telomeres, changing the *Fo* telomeric repeat sequence to the frequently identified sequences reported by *Tan et al*. (2022)[26].

For each assembly we report the percentage of contigs with telomeres at one or both ends. Because the total contig number ranges from 12 to 243, this percentage may be skewed. Therefore, when the total amount of contigs was more than 15, we repeated this analysis using only the L90 largest contigs in the assembly, where L90 is the number of largest contigs that together contain 90% of the assembly.

To determine if a tandem repeat or TE family is overrepresented in subtelomeric regions, we calculate for each repeat family given a unique repeat ID by RepeatModeler[51], how many base pairs fall within the previously identified subtelomeric ranges, and use SciPy's[52](version -1.13.1) hypergeom.sf function to test for significant overlap between repeat families and the 1 kb up or downstream of a telomeric repeat. Because multiple comparisons are performed per assembly (1 for each repeat family) Benjamini Hochberg (BH) multiple testing correction is performed using statsmodels[53](version -0.14) multitest function with the fdr bh method. Only repeat families with a corrected P-value below 0.05 and a length larger than 3000 bp were used for subsequent analysis.

Comparison of telomere-associated TE families within assemblies.

When more than one putative TE family is found to be overrepresented in the subtelomeres, we compare these to each other with megablast[54] (version -2.12.0) using the following settings: '–evalue 1e-5, -dust no -soft_masking false' to ensure low complexity regions are not masked. We calculated the relative Smith-Waterman (SW) score for each query-subject pair by dividing the SW score of their alignment against the maximum self-alignment score of the query or subject. Next, sequences with a relative SW score above 0.5 were aligned using MAFFT[55] (version -7.526) in the 'adjust directionality mode'. When the only gaps in the MSA were at one or both ends of the MSA, ends were clipped to the length of the shortest sequence. Indels were removed and a consensus sequence was made using consensus function in ClAlign[56] (version -1.1.4) based on majority presence. After manual curation the fraction of assemblies with a single overrepresented repeat family increased from 25% to 63%, the rest of the assemblies contain up to 9 overrepresented repeat families.

In order to exclude sequences not present near all identified telomeres we searched for copies of the consensus sequence(s) in the assembly they were identified in using megablast with the settings: '–evalue 1e-5, -dust no -soft_masking false'. We used bedtools[57] (version -2.31.1) intersect to identify BLAST hits of consensus sequences that fall within the first and last 2000 bases of each contig. Subsequently we compared these intersect results to our previously found telomere locations using a custom python script. Sub-telomeric Associated Repeats (STARS) are congregated into a single fasta file containing 52 sequences. If a STAR was present at the start or end of a contig while no telomere was identified, we search for *Fo* telomeric di-repeats 100 bases from the start or end of a contig using a custom python script.

Comparison of telomere-associated TE families between assemblies.

In order to compare the sequences of different STARS, every consensus sequence that passes our filters is analyzed using BLAST as a query against a database of all overrepresented sequences. We calculated the relative SW score for each query-subject. These scores are used to visualize a network with a minimum relative SW score of 0.5. Network visualization is done in Gephi[58] (version -0.10.1) and Inkscape[59] (version -1.3.1). In order to verify if the observed relationship between clustering samples and f. sp. is nonrandom, a permutation test is performed using a custom python script. In this test 10000 permutations of f. sp. are performed. The original network serves as a reference, when matching hosts and clusters in the permutation exceed the reference they get scored, p values are calculated by dividing the accumulated permuted scores by the number of permutations performed.

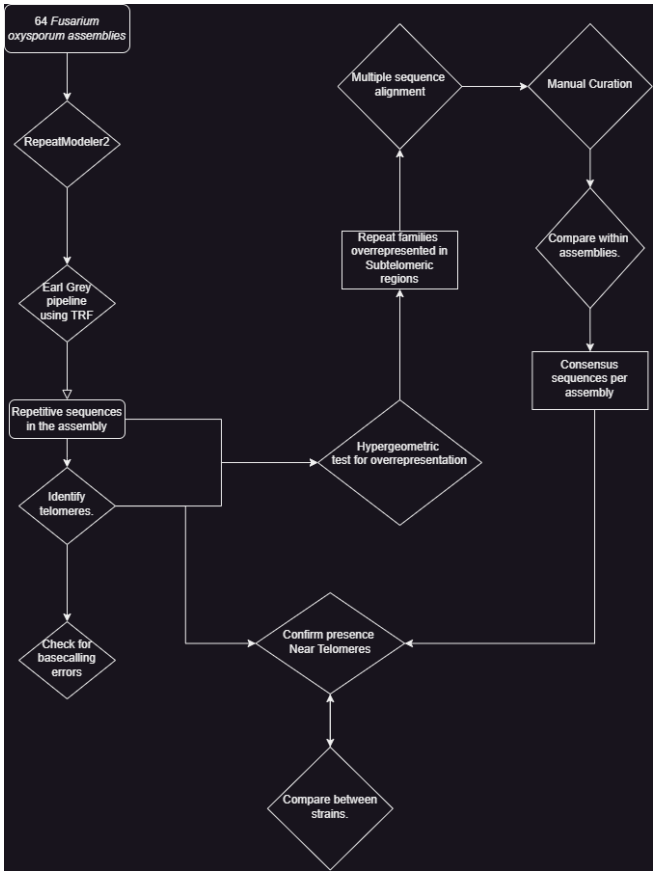

Figure 1: Overview of the analysis workflow. RepeatModeler 2 is used to identify transposable elements in 64 Fusarium oxysporum genome assemblies. The identified elements are processed through the Earl Grey pipeline to detect repetitive elements, assess the presence of telomeric repeats, and identify potential base-calling errors. When telomeres are detected, we analyse repetitive elements that are overrepresented in a defined subtelomeric domain. These elements are manually curated and compared within assemblies to generate consensus sequences. Only consensus sequences consistently found near all identified telomeres within an assembly are used for further comparisons across assemblies.

Results.

Public datasets include high quality Fo assemblies with a diverse host range.

To identify transposon families that are overrepresented in subtelomeric regions across multiple *Fusarium oxysporum* lineages, we compiled a dataset consisting of 64 high-quality genome assemblies, most of which are publicly available. This dataset includes genome sequences from strains with different host ranges (*forma specialis)* including banana (f. sp. *cubense*), cabbage (f. sp. *conglutinans*), cucumber (f. sp. *cucumerinum*), celery (f. sp. *apii)*, cilantro (f. sp. *coriandrii*), cotton (f. sp. *vasinfectum*),chili (f. sp. *capsicum*), strawberry (f. sp. *fragariae*), tomato (f. sp. *lycopersici*), flax (f. sp. *lini*), melon (f. sp. *melonis*), watermelon (f. sp. *niveum*), peanut (f. sp. *plukenetiae*), sesame (f. sp. *sesami*), onion (f. sp. *cepae*), stock (f. sp. *matthiolae*), date palm (f. sp. *albedinis*) and radish (f. sp. *raphani*), a strain that is used as a biocontrol (Fo47), a potential biocontrol strain isolated from a healthy plant (CH-0212), a strain isolated from a beetle (FCALT), and strain Forc016 that causes root rot on cucurbits (f. sp. *radicis-cucumerinum*). Of these 64 selected assemblies, 19 assemblies have been sequenced using Oxford Nanopore™ technology, 35 using PacBio™ technology, 6 using PacBio with Illumina™ polishing and 4 using Oxford Nanopore with Illumina polishing (see supplementary file 1 for an overview of the sequencing technologies, strain host, *f. sp.* and software used for assembly). The number of contigs ranges from 12 to 243 (Figure 2), indicating that at least some assemblies include chromosome fragments in addition to potential telomere-to-telomere chromosomes. All assemblies have a N50 > 1,5 Mb, indicating that more than 50% of the assembly is in contigs that span at least 1,5 Mb – the size of a small *Fo* chromosome. Based on this, we expect that most assemblies will include multiple (sub)telomeric regions.

Using tandem repeats identified by TRF results in reliable identification of telomeric regions.

To identify telomeres, we searched for tandem repeats corresponding to the *Fo* telomeric repeat sequence (CCCTAA) within the first or last 100 base pairs of each contig. All analyzed assemblies, except for the assembly of Fol59, contained at least one telomere. The number of contigs with telomeric repeats on both ends, ranges from 0 to 83,3% and the number of contigs with at least one telomeric repeat ranges from 0 to 45% (Figure 2). However, some assemblies come very close to telomere-to-telomere, with the highest percentage of contigs with telomeric repeats on both ends, at 83,3% of contigs, observed for strains FCALT, and 36102. Strain FCALT lacks telomeric repeats at the end of contig 1 and the beginning of contig 9, similar to what is reported by Berasategui et al (2022)[60]. In the Fo47 assembly, 75% of contigs contain telomeric repeats on both ends. We found that contig 11 lacks telomeric repeats at the 3" end, and contigs 4 and 6 lack telomeres at the 5" end, in concordance with the results published by *Wang* et al (2020)[61]. In the assembly of strain 160527, 66% of the contigs contain telomeric repeats on both ends: we found contigs 4, 5, 6 and 11 lack telomeric repeats on their terminal end, in concordance with the results obtained by *Asai* et al. (2019)[62]. Strain Forc016 has five contigs with telomeric repeats on both ends and 15 with telomeric repeats on one end, in concordance with the results published by *van Dam* et al. (2017)[63]. Together these results indicate that the used method can reliably identify telomeres in our dataset. Out of a total of 3287 contigs in 64 analyzed assemblies, we identified 184 contigs (6%) with telomeric repeats on both sides, and 470 contigs (14%) with a telomeric repeat on either side. Telomeric data for all strains can be accessed in supplementary file 2.

Out of the 64 assemblies in the dataset, 54 contain more than 15 contigs. When only considering the L90 contigs, i.e. the largest contigs that together make up 90% of the total assembly length, we find that out of a total of 1053 contigs 175 contigs (17%) have telomeric repeats on both sides and 194 contigs (18%) have a telomeric repeat on either side (Figure 3). The fact that a majority of contigs with telomeric repeats on both sides is present in the L90 dataset seems to indicate that these represent full chromosomes. 9 contigs however contain telomeric sequences on both ends but are not present in the L90 dataset. These can be viewed in supplementary file 3. Telomeric results for the L90 dataset can be accessed in supplementary file 4.

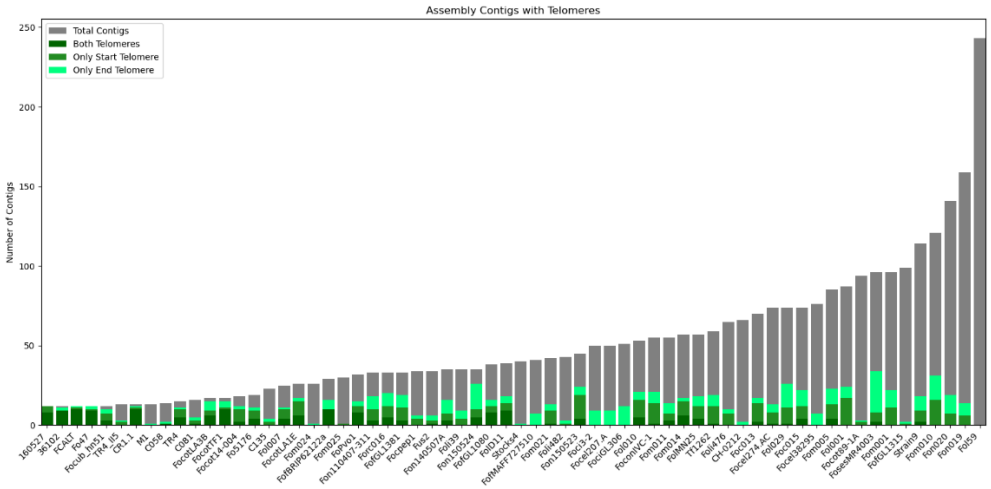

Figure 2: All analysed assemblies with their total contig numbers (grey) and the presence of telomeric repeats on both ends or end of a contig (green).

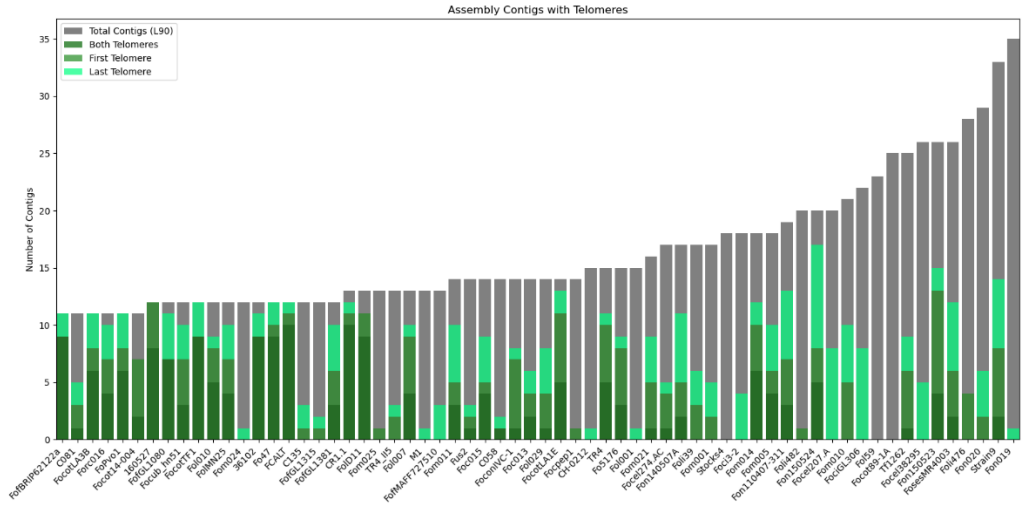

Figure 3: All analysed L90 assemblies with their total contig numbers (grey) and the presence of telomeric repeats on both ends or end of a contig (green).

Putative transposon families are overrepresented in subtelomeric regions.

In order to determine whether certain transposon families are associated with subtelomeric regions, we searched for repeats that are overrepresented in regions directly adjacent to telomeres (hypergeometric test, BH significance level (a = 0.05)) and have a minimal length of 3000 bp. In total we identified 144 Sub Telomeric Associated Repeats (STARS) that pass these filters. 16 strains have one STAR across all their contigs. 15 strains have 2 STARs, 16 strains have three STARS, and 12 strains have more than three identified STARs (Figure 4). The size of identified STARS ranges from 3003 bp to 40604 bp (Figure 5). For the following strains no STAR(s) are identified: CH-0212, TR4, TR4-II5, and M1, additionally, Fol59, since its assembly does not contain identified telomeres. These results indicate the presence of statistically significant association between putative transposon families or repetitive elements and subtelomeric region across most of the analyzed *Fo* strains. Hypergeometric test results for all repeat families and strains can be accessed in supplementary file 5.

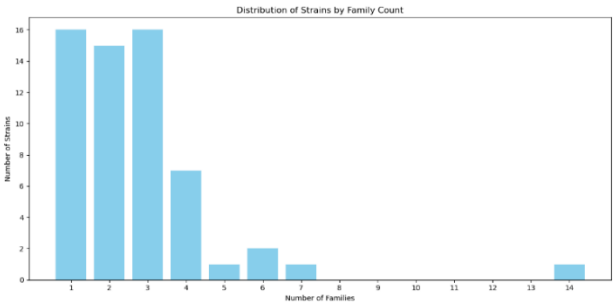

Figure 4: Number of overrepresented subtelomere associated repeats (STARs) per strain identified by EarlGrey and filtered on BH significance level (a = 0.05) and minimum length of 3000 bp in subtelomeric regions.

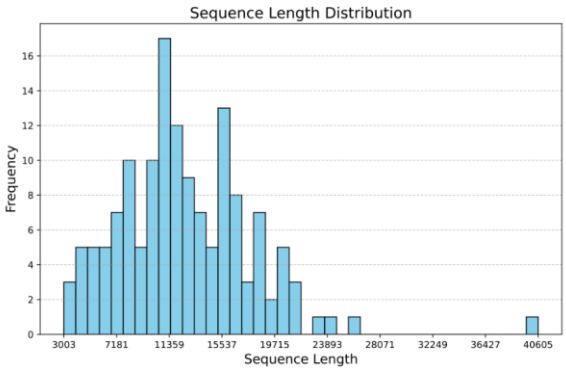

Figure 5: Sequence length distribution of identified STARS.

Subtelomeric regions in most analysed strains contain conserved sequences.

In order to further assess the diversity of STAR sequences per strain, and remove any redundancy between multiple, highly similar STARs, we compared STAR sequences within each assembly. Similar STARs were merged into one family with a consensus sequence derived from a manually curated multiple sequence alignment of consensus sequences of similar STARs. After this step we reduced the original 170, partially redundant STARs to 107 unique STARs, where 41 strains have a single conserved STAR, six strains contain two STARs and 17 have three or more STARs (Figure 6). The lengths of the curated STARs are also reduced and now range from 3003 to 27246 (Figure 7). In conclusion, most strains have a single conserved repeat family in their subtelomeric region which can vary in size on different contigs.

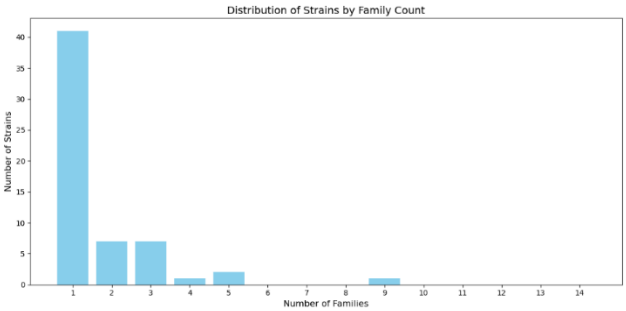

Figure 6: Number of overrepresented subtelomere associated repeats (STARs) per strain identified by EarlGrey and filtered on BH significance level (a = 0.05), minimum length of 3000bp and manually curated.

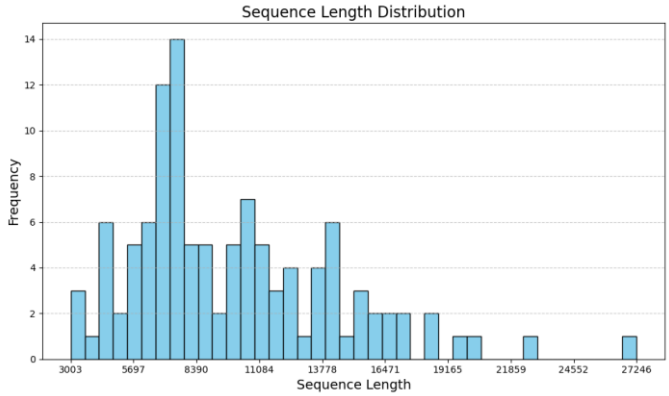

Figure 7 Sequence length distribution of identified and curated STARS.

Created STAR consensus sequences can be used to search for putative telomeres.

In order to verify if STAR consensus sequences are always present in the presence of telomeres we analyzed them using BLAST. The consensus sequences created is used as a query against its assembly. Blast hits are filtered at a minimum of 30% of consensus sequence length and presence within the first or last 2000 bases of a contig with identified telomeres. These results are in supplementary file 6. Out of the 107 STARs 45 are always present where a telomere has been previously identified. Interestingly however, all STARs are present contig extremities where no telomere had been identified. These regions could represent bona fide subtelomeres where the telomeric repeat had not been assembled correctly. In order to verify the presence of putative telomeres we searched for telomeric di-repeats in these regions. Strains FocotLA1E, Focpep1, Focub_hn51, FofGL1315, FofMAFF727510, Fom011 and Fom014 have extra telomeres when analyzed in this manner which likely have been missed before due to there not being enough instances of the repeat to be identified by RepeatModeler. These results can be found in supplementary file 7.

Non telomeric patterns are present across both PacBio and Nanopore assemblies.

When looking for the wrongly base-called telomeric patterns reported by Tan *et al.* (2022)[26] In the same manner as we identified telomeres, they are present in eight strains which are not limited to assemblies sequenced with Oxford Nanopore technology. Nanopore sequenced strain Foli39 has a normal telomeric repeat on contig 5, ranging from bases 1-68 directly followed by the (CCCTGG) hexamer ranging from bases 69-99. Similarly strain Foli476 has this same hexamer on the start of contigs 41 and 23 both ranging from bases 1-27. PacBio sequenced strain Fom001 however shows a similar instance where on contigs 9,15,18 and 86 telomeric sequences are directly followed by the (CCCTGG) hexamer. Strain Fom005 shows this pattern on contig 15. Fom010 shows this pattern on contig 90. Fom014 shows the CCCTGG hexamer on the first 33 bases of contig 9, however no telomere was identified in this instance. Strain Fon110407-311 shows this hexamer on positions 54-106 of contig 8, although no telomere was identified here. Finally strain FosesMR4003 shows this hexamer on the first 37 bases of contig 2. These results are in supplementary file 7.

Identified STARs have high similarities within host range.

In order to compare the different STARs of *Fo* strains they are compared using BLAST as a query against a database of all STARs that match our previous filtering criteria. Subsequently these results are converted into a network format with a relative SW score threshold of 0.5. The following network is obtained (Figure 8). Nodes are colored according to strain host. Edge thickness is relative to the calculated SW score. The network contains 27 nodes and 42 edges. 21 STAR sequences are not connected because they do not pass the similar threshold of 0,5, these are not visualized. The network contains seven connected components, four of which contain more than two strains (Additional table 1). The association between *f. sp.* and connected components is significant ($p<0.05$). Additional figure 1 shows the same network, but for all STARs analyzed, including the ones that are not connected.

Cluster one contains a variety of f. spp. including *apii, melonis, coriandrii, cucumerum, melonis* and *albedinis*. Out of the six analyzed f. sp. *melonis* strains three are present in this cluster. Notably strain

Fom021 has two STARs that pass our criteria and cluster together. Out of the eight sequences in this cluster four are categorized as a transposable element: Fom025: Line/CRE, Fom021: hAT/restless, Fom011: DNA kolobok-H and strain9: DNA kolobok-H. In this cluster the similarities range from 0.51 to 0.80. The STAR belonging to Strain9 is only connected to the STAR belonging to Fom011. Cluster two contains five *Fo* f . sp. *lycopersici* strains out of a total of seven in the starting dataset. Fol059 is missing because it has no identified telomeres. Fol010 has 2 contigs with telomeres where no STAR passes the filtering criteria. The STAR of each tomato infecting strain that passes the filtering criteria shows a high degree of similarity with a relative SW score ranging from 0.87 to 0.99. Only the sequences of Fol001 is categorized as a transposable element belonging to the hAT/restless family. Cluster three contains three strains of f. sp. *melonis* with a relative SW score ranging from 0.55 to 0.57. Two of the sequences are unidentified. One is identified as a Line. Cluster four contains three strains of f. sp. *cubense* and one strain of f. sp. *fragariae*. In this cluster the relative SW score ranges from 0.62 to 0.96. When looking at only the f. sp. *cubense* strains the relative SW scores are above 0.90. Three of the four sequences are identified as DNA Academ 2. One of the sequences is identified as a Kolobok-H. Cluster five contains two f. sp. *cubense* strains which barely pass the SW score threshold. One of these identified as a DNA transposon, the other as a line. Cluster six contains two f. sp. *niveum* strains with a relative SW score of 0.81. Finally, cluster seven contains two strains belonging to f. sp. *conglutinans* and f. sp. *lini* with a relative SW score of 0.55.

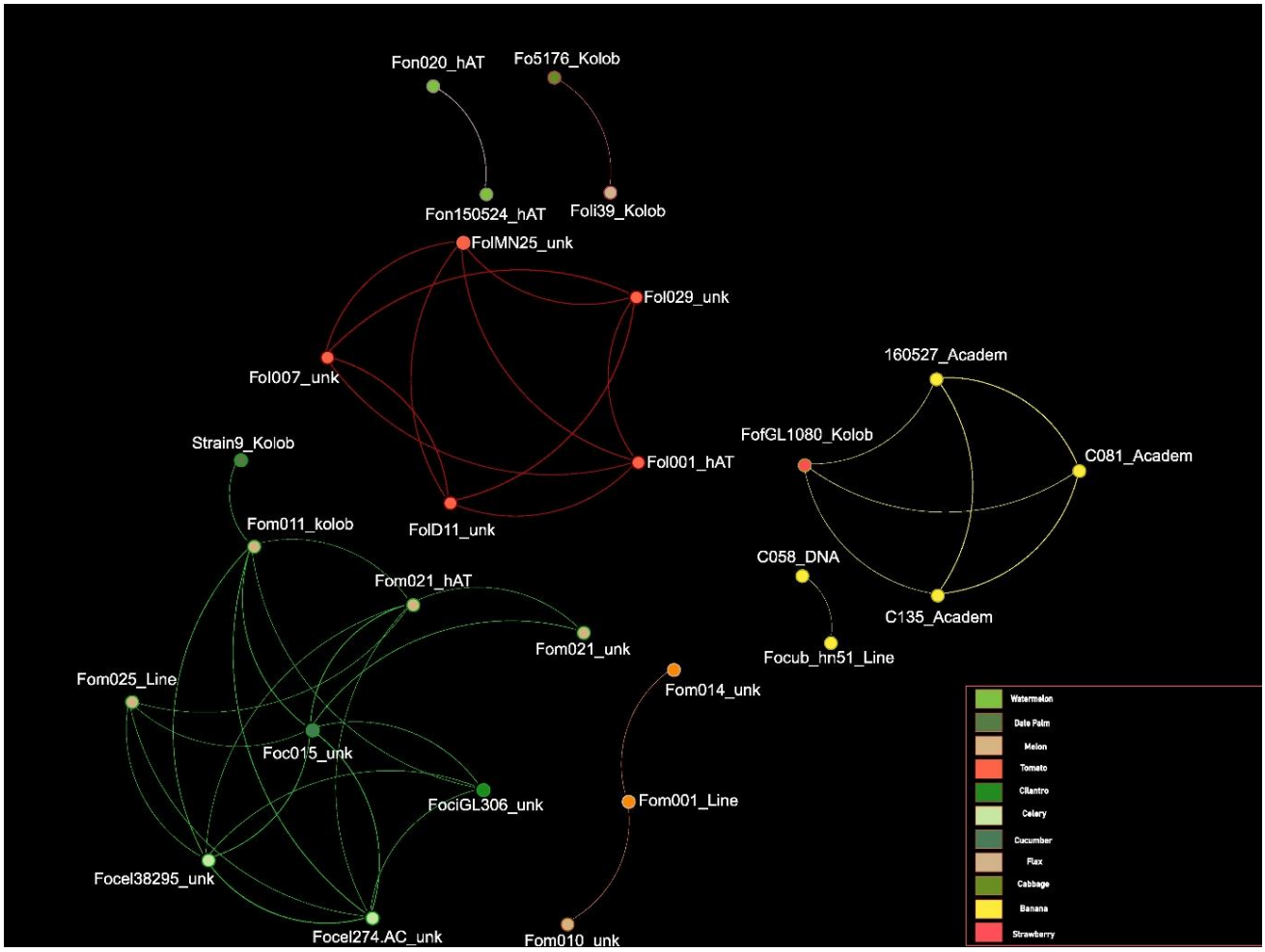

Figure 8: Visualisation of the similarity between different stars with a relative SW score above 0.5. Nodes and edges are coloured according to their host. This network contains 27 nodes and 42 edges. Partial clustering of STARs according to host partition is visible. Relation between connected component and *f. sp.* is significant ($p < 0.05$). Unconnected components are not visualized.

Discussion and Conclusions.

Telomeres in the *FOSC*.

A diverse range of methods exists for identifying and reporting on the presence of telomeres, which complicates comparisons between studies. Out of 64 analysed *Fo* genome assemblies four methods of telomere identification have been reported. When publishing the genome assembly of *Fo* f. sp. *cubense* strain 160527 telomeres are identified by looking at over 25 copies of a telomeric repeat within 50kb of contig ends[62]. Berasategui *et al*. *(2022)*[60] reported looking for stretches of telomeric repeat hexamers at contig ends, without specifying a copy number or defining contig ends. Wang *et al*. *(2020)*[61] reports looking for more than three copies of the telomeric repeat on contig ends. Other studies simply specified that they identified enrichment of telomeric repeats on contig ends[64,65]. Several tools exist specifically to identify telomeric sequences, however these are mostly used for studying tumors and are incapable of identifying alternative telomeric sequences[66]. For example: $(TTTAGGG)_n$ in *Arabidopsis thaliana*[2], or $(AATGGGGGG)_n$ in *Cyanidioschyzon merolae*[67]. Due to the wide-ranging application telomeres have in chromosomal structure, it is important to have a uniform method to both identify and quantify their presence. In this research we have attempted to resolve this by using well tested algorithms for tandem repetitive sequence identification in identifying telomeric repeats.

Assembly software may cause a loss of telomeric data.

The dataset analysed compromises a wide range of f. sp. representing the diversity of the *Fo* genus. Identified telomeres match previous studies, and we report their presence in strains where no telomeres are previously identified. Strain Fol59 does not contain any telomeres and is the most fragmented assembly in our dataset. One potential explanation is the assembly software used. This is partially concordance with results obtained by Saud *et al*. (2021)[68] , who report that the long read assemblers Raven[69], Shasta[70] and wtdbg2[71] suffer from a loss of telomeric sequences. No assemblies that were made by Raven and wtdbg2 were present in our dataset, but strain Fol59 is the only strain assembled by Shasta (v0.1.0). We recommend assembling this strain with different assembly software.

Assessing basecalling accuracy of telomeres in nanopore assemblies.

When looking at the alternative telomere hexamers reported *Tan et al*. (2022)[26] in a similar manner telomeres are identified, they are present in eight strains. Only three of these strains are sequenced using Nanopore technology, leading to the question of these sequences being wrongly basecalled, or genuine errors caused by malfunction in the elongation of telomeres. This however seems unlikely since the hexamers are found after properly identified telomeres while one would expect to be more frequent in the extremities. One way to potentially see if these are wrongly based-called is by looking at the raw sequence reads and identifying the occurrence of these sequences. We have, however, successfully identified telomeric sequences in 22 of our 23 nanopore sequenced assemblies indicating a reliable, uniform and robust method for telomere identification.

Identifying *Fusarium oxysporum* subtelomeric domains.

Subtelomeres have been shown to play an extensive role in genome maintenance[6] but in *Fo* they are not well characterized. With our telomeric regions identified we decided to look for distinct repetitive sequences identified by Earl grey and test their overrepresentation in the subtelomeric region. Nearly all strains seem to possess these sequences termed subtelomere associated repeats, and we show that they are usually conserved within strains. Our results do represent conservative estimates. During the hypergeometric test all repetitive elements have been analysed leading to an inflation of p-values and therefore a more stringent multiple testing correction. The test can be repeated after filtering the Earl Grey output for sequences identified as simple repeats and removing them from the analysis. Additionally, we searched for overrepresentation in the regions 1000bp up and downstream of identified telomeres. Considering the large size of the STARs identified (Figures 5 and 6) it is likely that the subtelomeric domain in *Fo* is much larger and contains more STARs.

Subtelomeric associated repeats vary in size across contigs.
To identify the most conserved version of STARs, we performed manual curation and proceeded only with the conserved version. This leads to a quite striking size reduction for some STARs (Figures 4 and 6). If the identified STARs are indeed transposable elements this reduction in size is hardly surprising but rather indicative of the duplication events or recombination taking place on some chromosome ends. Some of these STARs have indeed been identified as transposable elements, however careful curation of their structural and functional aspects should be performed to verify this. Alternatively, these regions do not represent TEs but simply parts of highly repetitive subtelomeres. In this research we have analysed only the STAR sequences conserved within strains by manually curating the sequences. It is recommended to further analyse the identified sequences, and what leads to their vastly different size within strains.

Identifying subtelomeres in strains without identified STARs.
We have identified sequences that are overrepresented and directly adjacent to telomeres which are also conserved within *Fo* strains. It is important to realize that these might not be the full subtelomeric domains, since they only represent one family identified by Earl Grey. Due to our stringent selection, we might miss subtelomeric regions compromised of multiple distinct repeat sequences. To identify full subtelomeric domains It is recommended to perform a pairwise alignment on contig ends with identified telomeres. This analysis would be especially interesting on strains CH-0212, TR4, TR4-II5, and M1 where we have failed to identify any STARs. The same analysis should be performed on the rest of the strains in our dataset to verify the extend of the subtelomeric domain. Once this domain is identified, gene families present near them should be further investigated since they potentially have a role in pathogenicity. Failure to identify overrepresented families does not seem related to sequencing technology or assembly software used, since these strains are sequenced using either PacBio or Nanopore sequencing and assembled using Falcon[72], HGAP 4[73] or Canu[74]. All of which are also used in assemblies were STARs have been identified.

STARs can be used to identify putative telomeric regions.
Due to the difficulty in assembling telomeric regions, we tried to identify putative chromosome ends by looking at the STARs identified. When analysing these sequences using BLAST as a query against their own assembly and subsequently checking their presence in contig extremities in the presence of telomeric di repeats. This has resulted in the identification of putative telomeric regions in seven strains. The telomeric regions in these strains themselves have possibly not been assembled properly. Most likely due to the difficulties mentioned before. One method to overcome this problem would be to employ the Telo-seq method[75] . This method was originally designed for studying cancer cells and uses Oxford Nanopore sequencing from the start of a telomere well into the subtelomeric region.

Analysed STARs represent only a fraction of those present.
After an additional filtering step, where we only consider STARs that are present within an assembly near all identified telomeres (removing sequences that have present near a region where no telomere is identified, or if they are more than 2000bp removed from that telomere) we are left with 48 out of 94 STARs. However further analysis is required. For example, STARs in strains CR1.1 and FocotLA3B and TF1262 fail this criterion because they are present on a contig that does not contain a telomere. This could potentially be an assembly error. The intersecting region of blast hits and telomeres could also be increased since for example strain Focel207.A fails to meet our thresholds because the STAR is around 2500bp removed from identified telomeres. There are also occasions where multiple families per strain pass all filtering criteria. This is due to the sequences being similar, yet not similar enough for them to be merged into a single consensus sequence. In conclusion further analysis of these STARs is recommended, since they represent the diversity of STARs found within strains. An analysis should be performed with less strict filtering criteria on the already created consensus STAR sequences.

Highly similar STARs belong to the same *formae specialis*.
Strikingly, 21 out of the 48 STARs have a relative SW score below 0.5 indicating low sequence similarity (unconnected components in additional figure 1). STARs in these strains are unique. The rest of the STARs

cluster together depending on the SW score threshold. When the SW score threshold is 0.7, all connected components belong to the same f. sp.. Below the threshold of 0.7 we get clusters of mixed f. spp.. This indicates that STARs within f. sp., namely the *Fol* and *Foc* strains analysed share significant similarities. All but two of our analysed *Fol* strains have highly similar STARs. The STAR of strain Fol010 is not present near every telomeric sequence and therefore excluded. It is recommended to assess what sequences are near the telomeres where this STAR is not present. And to calculate the similarity of strain Fol010s STAR sequence with the rest of the identified STARS in *Fol*.

Manual curation is required to confirm transposon sequences.

In the cluster containing exclusively *Fol* strains, only the sequence of *Fol001* is categorized by EarlGrey as a transposable element belonging to the hAT/restless family. The rest of the sequences are classified as unknown even though their relative SW score is >0.7. van Westerhoven *et al*. (2024)[44] have shown that accessory regions on chromosome 1 and contig 12 can be distinguished from the smaller accessory regions located near subtelomeres. This separation is based on GC content which can be caused by either repeat induced polymorphism (RIP) or fungal defence mechanisms that introduce C to T mutations in repetitive regions[76]. This leads to the hypothesis that the identified transposon in *Fol001* is less affected by RIP, causing an easier identification by Earl Grey. Alternatively, that the transposon insertion is more recent compared to the other sequences identified. Another possibility is that the Earl Grey pipeline overclassifies repetitive sequences into transposon families. Analysis of the structure and functional domains in STARs identified as transposons could confirm this.

Clustering of STARs seems unrelated to phylogeny.

When comparing the clusters obtained to a phylogenetic tree (additional figure 2) it becomes apparent that clustering STARs can be phylogenetically very distant from each other. For example, strains FolD11, FolMN25, Fol007 and Fol029 are highly similar but Fol001, whose STAR sequence clusters together with the STARs from these strains appears in a different branch. Similarly, in cluster 2 we have three highly similar f. sp. *cubense* strains which are genetically almost identical. The STAR sequence of more distant strain FofGL1080 also appears in this cluster although with a lower relative SW score. Strain Fom001 is genetically more similar to the *Fol* strains, yet its STAR sequence clusters with other *Fom* strains. This can potentially be explained by horizontal chromosome transfer between strains, which has been shown to be widespread in *Fo*[77]. When a horizontally transferred chromosome contains an active transposon insertion in its subtelomeric region, it possible that it has spread to subtelomeric regions present on other chromosomes. Alternatively, if a certain transposon family is overrepresented in the genome of multiple *Fo* strains it is not surprising that at some point it translocated to a telomeric region. In this case however, we would not expect it to be significant in our test of subtelomeric overrepresentation.

Similarity between subtelomeric regions can also be explained by physical interaction and DNA repair mechanisms. Chromosome conformation capture followed by high-throughput sequencing (HI-C) has revealed that most fungal genomes are globally organised in Rabl chromosome configurations[78]. This configuration is characterized by the clustering of centromeres on one side of the nuclear envelope and chromosomal arms extending outwards towards the opposing nuclear periphery on which the (sub)telomeres associate. The occurrence of a double stranded DNA break and its concomitant repair onto DNA strands spatially close to each other might over time result in subtelomeric regions homogenising within genomes.

Our results have shown a reliable method to identify telomeres, even in fragmented genome assemblies by relying on the identification of tandem repeats. Subsequently we show that these sequences are flanked by STARs that seem to be conserved within strains, even though the region they occupy is highly dynamic. These STARs can be used to identify putative chromosome ends and may aid in the assembly of *Fusarium oxysporum* genomes. Additionally, while most of these regions are unique to strains, some of these display high similarities and are identified as potential transposable elements. This will allow further research into the notoriously difficult to study telomeric and subtelomeric regions in *Fusarium oxysporum*.

Bibliography.

1. Diotti, R., Esposito, M. & Shen, C. H. Telomeric and sub-telomeric structure and implications in fungal opportunistic pathogens. *Microorganisms* vol. 9 Preprint at https://doi.org/10.3390/microorganisms9071405 (2021).

2. Tao, Y. *et al.* Atlas of telomeric repeat diversity in Arabidopsis thaliana. *Genome Biology 2024 25:1* **25**, (2024).

3. Blackburn, E. H. Structure and function of telomeres. *Nature* vol. 350 Preprint at https://doi.org/10.1038/350569a0 (1991).

4. Greider, C. W. & Blackburn, E. H. The telomere terminal transferase of tetrahymena is a ribonucleoprotein enzyme with two kinds of primer specificity. *Cell* **51**, (1987).

5. Morrish, T. A. *et al.* Multiple Mechanisms Contribute To Telomere Maintenance. *J Cancer Biol Res* **1**, (2013).

6. Kwapisz, M. & Morillon, A. Subtelomeric Transcription and its Regulation. *Journal of Molecular Biology* vol. 432 Preprint at https://doi.org/10.1016/j.jmb.2020.01.026 (2020).

7. Flint, J. *et al.* Sequence comparison of human and yeast telomeres identifies structurally distinct subtelomeric domains. *Hum Mol Genet* **6**, (1997).

8. Del Portillo, H. A. *et al.* A superfamily of variant genes encoded in the subtelomeric region of Plasmodium vivax. *Nature* **410**, (2001).

9. Cheng, Q. *et al.* stevor and rif are Plasmodium falciparum multicopy gene families which potentially encode variant antigens. *Mol Biochem Parasitol* **97**, (1998).

10. Bringaud, F. *et al.* A new, expressed multigene family containing a hot spot for insertion of retroelements is associated with polymorphic subtelomeric regions of Trypanosoma brucei. *Eukaryot Cell* **1**, (2002).

11. Keely, S. P. *et al.* Gene arrays at Pneumocystis carinii telomeres. *Genetics* **170**, (2005).

12. De Las Peñas, A. *et al.* Virulence-related surface glycoproteins in the yeast pathogen Candida glabrata are encoded in subtelomeric clusters and subject to RAP1- and SIR-dependent transcriptional silencing. *Genes Dev* **17**, (2003).

13. Sen, S. K. *et al.* Human genomic deletions mediated by recombination between Alu elements. *Am J Hum Genet* **79**, (2006).

14. Jiang, N., Bao, Z., Zhang, X., Eddy, S. R. & Wessler, S. R. Pack-MULE transposable elements mediate gene evolution in plants. *Nature* **431**, (2004).

15. Bennetzen, J. L. & Wang, H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual Review of Plant Biology* vol. 65 Preprint at https://doi.org/10.1146/annurev-arplant-050213-035811 (2014).

16. Fedoroff, N. V. Transposable elements, epigenetics, and genome evolution. in *Science* vol. 338 (2012).

17. Farman, M. L. & Kim, Y. S. Telomere hypervariability in Magnaporthe oryzae. *Mol Plant Pathol* **6**, (2005).

18. Kang, S., Lebrun, M. H., Farrall, L. & Valent, B. Gain of virulence caused by insertion of a Pot3 transposon in a Magnaporthe grisea avirulence gene. *Molecular Plant-Microbe Interactions* **14**, (2001).

19. Schmidt, S. M. *et al.* MITEs in the promoters of effector genes allow prediction of novel virulence genes in Fusarium oxysporum. *BMC Genomics* **14**, (2013).

20. Starnes, J. H., Thornbury, D. W., Novikova, O. S., Rehmeyer, C. J. & Farman, M. L. Telomere-targeted retrotransposons in the rice blast fungus magnaporthe oryzae: Agents of telomere instability. *Genetics* **191**, (2012).

21. Rahnama, M. *et al.* Transposon-mediated telomere destabilization: A driver of genome evolution in the blast fungus. *Nucleic Acids Res* **48**, (2020).

22. Eichler, E. E. Segmental duplications: What's missing, misassigned, and misassembled-and should we care? *Genome Res* **11**, (2001).

23. Hon, T. *et al.* Highly accurate long-read HiFi sequencing data for five complex genomes. *Sci Data* **7**, (2020).

24. Wang, Y., Zhao, Y., Bollas, A., Wang, Y. & Au, K. F. Nanopore sequencing technology, bioinformatics and applications. *Nature Biotechnology* vol. 39 Preprint at https://doi.org/10.1038/s41587-021-01108-x (2021).

25. Udaondo, Z. *et al.* Comparative analysis of pacbio and oxford nanopore sequencing technologies for transcriptomic landscape identification of penaeus monodon. *Life* **11**, (2021).

26. Tan, K. T., Slevin, M. K., Meyerson, M. & Li, H. Identifying and correcting repeat-calling errors in nanopore sequencing of telomeres. *Genome Biol* **23**, (2022).

27. Edel-Hermann, V. & Lecomte, C. Current status of fusarium oxysporum formae speciales and races. *Phytopathology* vol. 109 Preprint at https://doi.org/10.1094/PHYTO-08-18-0320-RVW (2019).

28. Watanabe, A., Miura, Y., Sakane, K., Ito, S. ichi & Sasaki, K. Identification and characterization of Fusarium commune, a causal agent of lotus rhizome rot. *Journal of General Plant Pathology* **89**, (2023).

29. Hof, H. The medical relevance of fusarium spp. *Journal of Fungi* vol. 6 Preprint at https://doi.org/10.3390/jof6030117 (2020).

30. Ortoneda, M. *et al.* Fusarium oxysporum As A Multihost Model for the Genetic Dissection of Fungal Virulence in Plants and Mammals. *Infect Immun* **72**, (2004).

31. Navarro-Velasco, G. Y., Prados-Rosales, R. C., Ortíz-Urquiza, A., Quesada-Moraga, E. & Di Pietro, A. Galleria mellonella as model host for the trans-kingdom pathogen Fusarium oxysporum. *Fungal Genetics and Biology* **48**, (2011).

32. Di Pietro, A., Madrid, M. P., Caracuel, Z., Delgado-Jarana, J. & Roncero, M. I. G. Fusarium oxysporum: Exploring the molecular arsenal of a vascular wilt fungus. *Molecular Plant Pathology* vol. 4 Preprint at https://doi.org/10.1046/j.1364-3703.2003.00180.x (2003).

33. Hao, Y. *et al*. The Genome of Fusarium oxysporum f. sp. phaseoli Provides Insight into the Evolution of Genomes and Effectors of Fusarium oxysporum Species. *Int J Mol Sci* **24**, (2023).

34. Fokkens, L. *et al*. The multi-speed genome of Fusarium oxysporum reveals association of histone modifications with sequence divergence and footprints of past horizontal chromosome transfer events. *bioRxiv* Preprint at https://doi.org/10.1101/465070 (2018).

35. Li, J., Fokkens, L., van Dam, P. & Rep, M. Related mobile pathogenicity chromosomes in Fusarium oxysporum determine host range on cucurbits. *Mol Plant Pathol* **21**, (2020).

36. Ayukawa, Y. *et al*. A pair of effectors encoded on a conditionally dispensable chromosome of Fusarium oxysporum suppress host-specific immunity. *Commun Biol* **4**, (2021).

37. Ma, L. J. *et al*. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature* **464**, (2010).

38. Grandaubert, J. *et al*. Transposable element-assisted evolution and adaptation to host plant within the Leptosphaeria maculans-Leptosphaeria biglobosa species complex of fungal pathogens. *BMC Genomics* **15**, (2014).

39. Biju, V. C., Fokkens, L., Houterman, P. M., Rep, M. & Cornelissen, B. J. C. Multiple evolutionary trajectories have led to the emergence of races in Fusarium oxysporum f. sp. lycopersici. *Appl Environ Microbiol* **83**, (2017).

40. Kolomietz, E., Meyn, M. S., Pandita, A. & Squire, J. A. The role of Alu repeat clusters as mediators of recurrent chromosomal aberrations in tumors. *Genes Chromosomes and Cancer* vol. 35 Preprint at https://doi.org/10.1002/gcc.10111 (2002).

41. Delprat, A., Negre, B., Puig, M. & Ruiz, A. The transposon Galileo generates natural chromosomal inversions in Drosophila by ectopic recombination. *PLoS One* **4**, (2009).

42. Tosa, Y. *et al*. Evolution of an avirulence gene, AVR1-CO39, concomitant with the evolution and differentiation of Magnaporthe oryzae. *Molecular Plant-Microbe Interactions* **18**, (2005).

43. Zhang, J., Zuo, T. & Peterson, T. Generation of Tandem Direct Duplications by Reversed-Ends Transposition of Maize Ac Elements. *PLoS Genet* **9**, (2013).

44. van Westerhoven, A. C. *et al*. Segmental duplications drive the evolution of accessory regions in a major crop pathogen. *New Phytologist* **242**, (2024).

45. Daboussi, M. J. & Langin, T. Transposable elements in the fungal plant pathogen Fusarium oxysporum. *Genetica* **93**, (1994).

46. Sayers, E. W. *et al*. GenBank. *Nucleic Acids Res* **50**, (2022).

47. Flynn, J. M. *et al*. RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**, (2020).

48. Baril, T., Galbraith, J. & Hayward, A. Earl Grey: A Fully Automated User-Friendly Transposable Element Annotation and Analysis Pipeline. *Mol Biol Evol* **41**, (2024).

49. Platt, R. N., Blanco-Berdugo, L. & Ray, D. A. Accurate transposable element annotation is vital when analyzing new genome assemblies. *Genome Biol Evol* **8**, (2016).

50. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* **27**, (1999).

51. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. *RepeatMasker Open-3.0* Preprint at (1996).

52. Virtanen, P. *et al.* SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nat Methods* **17**, (2020).

53. Seabold, S. & Perktold, J. Statsmodels: Econometric and Statistical Modeling with Python. in *Proceedings of the 9th Python in Science Conference* (2010). doi:10.25080/majora-92bf1922-011.

54. Chen, Y., Ye, W., Zhang, Y. & Xu, Y. High speed BLASTN: An accelerated MegaBLAST search tool. *Nucleic Acids Res* **43**, (2015).

55. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Mol Biol Evol* **30**, (2013).

56. Tumescheit, C., Firth, A. E. & Brown, K. CIAlign: A highly customisable command line tool to clean, interpret and visualise multiple sequence alignments. *PeerJ* (2022) doi:10.7717/peerj.12983.

57. Quinlan, A. R. & Hall, I. M. BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, (2010).

58. Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. in *Proceedings of the 3rd International AAAI Conference on Weblogs and Social Media, ICWSM 2009* (2009). doi:10.1609/icwsm.v3i1.13937.

59. Schöler, U. Inkscape. in *Inkscape* (2024). doi:10.3139/9783446479296.fm.

60. Berasategui, A. *et al.* The leaf beetle Chelymorpha alternans propagates a plant pathogen in exchange for pupal protection. *Current Biology* **32**, (2022).

61. Wang, B. *et al.* Chromosome-scale genome assembly of fusarium oxysporum strain Fo47, a fungal endophyte and biocontrol agent. *Molecular Plant-Microbe Interactions* **33**, (2020).

62. Asai, S. *et al.* High-Quality Draft Genome Sequence of Fusarium oxysporum f. sp. cubense Strain 160527, a Causal Agent of Panama Disease . *Microbiol Resour Announc* **8**, (2019).

63. Van Dam, P. *et al.* A mobile pathogenicity chromosome in Fusarium oxysporum for infection of multiple cucurbit species. *Sci Rep* **7**, (2017).

64. Jenner, B. N. & Henry, P. M. Pathotypes of Fusarium oxysporum f. sp. fragariae express discrete repertoires of accessory genes and induce distinct host transcriptional responses during root infection. *Environ Microbiol* **24**, (2022).

65. Fokkens, L. *et al.* A chromosome-Scale genome assembly for the fusarium oxysporum strain Fo5176 to establish a model arabidopsis-Fungal pathosystem. *G3: Genes, Genomes, Genetics* **10**, (2020).

66. Feuerbach, L. *et al.* TelomereHunter - In silico estimation of telomere content and composition from cancer genomes. *BMC Bioinformatics* **20**, (2019).

67. Nozaki, H. *et al.* A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. *BMC Biol* **5**, (2007).

68. Saud, Z., Kortsinoglou, A. M., Kouvelis, V. N. & Butt, T. M. Telomere length de novo assembly of all 7 chromosomes and mitogenome sequencing of the model entomopathogenic fungus, Metarhizium brunneum, by means of a novel assembly pipeline. *BMC Genomics* **22**, (2021).

69. Vaser, R. & Šikić, M. Raven: a de novo genome assembler for long reads. *bioRxiv* (2021).

70. Shafin, K. *et al.* Nanopore sequencing and the Shasta toolkit enable efficient de novo assembly of eleven human genomes. *Nat Biotechnol* **38**, (2020).

71. Ruan, J. & Li, H. Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, (2020).

72. Chin, C. S. *et al.* Phased diploid genome assembly with single-molecule real-time sequencing. *Nat Methods* **13**, (2016).

73. Chin, C. S. *et al.* Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods* **10**, (2013).

74. Suzuki, Y. & Myers, G. Accurate k-mer Classification Using Read Profiles. in *Leibniz International Proceedings in Informatics, LIPIcs* vol. 242 (2022).

75. Schmidt, T. T. *et al.* High resolution long-read telomere sequencing reveals dynamic mechanisms in aging and cancer. *Nat Commun* **15**, (2024).

76. Galagan, J. E. & Selker, E. U. RIP: The evolutionary cost of genome defense. *Trends in Genetics* vol. 20 Preprint at https://doi.org/10.1016/j.tig.2004.07.007 (2004).

77. Henry, P. M. *et al.* Horizontal chromosome transfer and independent evolution drive diversification in Fusarium oxysporum f. sp. fragariae. *New Phytologist* **230**, (2021).

78. Torres, D. E., Reckard, A. T., Klocko, A. D. & Seidl, M. F. Nuclear genome organization in fungi: From gene folding to Rabl chromosomes. *FEMS Microbiology Reviews* vol. 47 Preprint at https://doi.org/10.1093/femsre/fuad021 (2023).