 doi:10.20944/preprints202204.0298.v1

Article

Intra-Species Genomic Variation in the Pine Pathogen *Fusarium Circinatum*

Mkhululi Maphosa[1], Emma Steenkamp[1], Aquillah Kanzi[1], Stephanie van Wyk[1], Lieschen De Vos[1*], Quentin Santana[1], Tuan Duong[1], Brenda Wingfield[1]

[1] Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa
* Correspondence: lbahlman@up.ac.za; Tel.: +27124203939

Abstract: *Fusarium circinatum* is an important global pathogen of pine trees. Genome plasticity has been observed in different isolates of the fungus, but no genome comparisons are available. To address this gap, we sequenced and assembled to chromosome level five isolates of *F. circinatum*. These genomes were analysed together with previously published genomes of *F. circinatum* isolates FSP34 and KS17. Multi-sample variant calling identified a total of 461683 micro variants (SNPs and small indels) and a total of 1828 macro structural variants of which 1717 were copy number variants and 111 were inversions. Variant density was higher on sub-telomeric regions of chromosomes. Variant annotation revealed that genes involved in transcription, transport, metabolism and transmembrane proteins were overrepresented in gene sets affected by high impact variants. A core genome representing genomic elements conserved in all the isolates and a non-redundant pangenome representing all genomic elements is presented. Whole genome alignments showed that an average of 93% of the genomic elements are present in all isolates. The results of this study reveal that some genomic elements are not conserved within the isolates and some variants are high impact. The described genome-scale variations will help inform novel disease management strategies against the pathogen.

Keywords: genome; accessory; core genome; *Fusarium circinatum*; structural variants; inversions; indels; pangenome

1. Introduction

Unravelling the genetic basis of population-level phenotypic variation, such as pathogenicity, has been the underlying driving force in comparative genomics research [1,2]. For many fungal pathogens, we now know that genomes can be divided into sub-genomic compartments based on their evolutionary rates [3]. This is often referred to as the "two speed genome concept" [4], where some genomic regions seem to accumulate polymorphisms relatively quickly while others remain stable over extended periods of time. These more variable regions are often rich in repetitive elements and usually harbour genes that encode niche-defining phenotypes [5,6]. The latter includes effectors, which are secreted disease determinants of host infection and colonization, and are thus key to all plant-fungus interactions [7,8]. Variable regions are thus valuable resources for studying how fungal pathogens are adapted to their specific environments.

Although selection for disease phenotypes in plant breeding is dependent on knowledge about fungal pathogens inter- and intra-species variability [9,10], access to whole genome sequences can speed up the process. For example, virulence can be associated with specific micro and/or macro genomic polymorphisms located in specific sub-genomic compartments. The accuracy of such a process is therefore dependent on the quality of genome sequencing technologies, as well as genome assembly and variant calling platforms. Currently, long-read single-molecule sequencing technologies are useful

 doi:10.20944/preprints202204.0298.v1

for providing a broad scaffolding framework with which to assemble whole fungal genomes. This is because it can allow for the generation of long (up to 80kb) sequence reads [11,12] that can span complex structural variants including stretches of repetitive [12] and highly flexible regions which have been shown to be associative with pathogenic determinants [1]. To account for the higher error rate of long-read sequencing technologies, they are often combined with short reads from platforms such as the Illumina technology [13]. Whole genomes in which chromosomes are sequenced from telomere to telomere can thus be assembled, thereby allowing comprehensive analysis of genomic structural variation as well as more fine-scale analyses such as those involving single nucleotide polymorphisms (SNPs) [13].

In this study we were interested in genomic variation in *Fusarium circinatum*. This economically important pathogen of pines is responsible for major losses in pine-based forestry across the world. Although high levels of genetic diversity and phenotypic variation have been reported within populations of this fungus [14,15], this was mostly attributed to sexual reproduction and mutation [16-18]. Other common sources of variation that could play a role in *F. circinatum* includes transposable elements [19,20] and horizontal gene transfer (HGT)[21]. Transposable elements can bring about genomic variations in several ways; e.g., transposition into exons, introns, regulatory regions and the mediation of non-homologous recombination [20]. To counter their impact on genome integrity, many fungi employ mechanisms such as Repeat Induced Point mutation (RIP) [22,23] that may introduce more variations, as have been shown in *F. circinatum* [24]. HGT is the exchange of genetic material between organisms that are not in a parent-offspring relationship [25]. A notable example in fungal pathogens are dispensable chromosomes, often harbouring pathogenicity genes, that can be exchanged among isolates [2]. In *F. circinatum*, loss of a dispensable chromosome, in this case the 12th chromosome, reduces the pathogen's virulence [26,27].

When trying to link phenotype with genome-base polymorphism, a range of genomic features may be considered. These can range from micro variants like SNPs to macro variants. The latter is mostly represented by structural variants (SVs) such as copy number variations (CNVs) or indels, translocations, inversions and whole chromosome loss/gain [28]. Although SNPs are employed most commonly, SVs also have great value as their occurrence and distribution can vary within populations [29], where they can have profound effects on phenotype diversity [30]. This is particularly true when SVs occur within genes or in regulatory regions [31,32]. CNVs or indels represent replicated/deleted genomic regions that arise from nonallelic homologous recombination and retrotransposition [33], and has been shown capable of extensively impacting gene expression and phenotypic diversity [33,34]. Translocations result from recombination between nonhomologous chromosomes leading to reciprocal exchange or nonreciprocal transfer of genetic material between the chromosomes [35]. Inversions are inverted chromosomal segments in which the DNA content remains the same and only the linear order of the DNA bases is changed [36]. They arise when a chromosome breaks at two points and the segment is reinserted in an inverted orientation. Inversions suppress recombination in heterokaryotypes and can even cause genetic isolation between populations [36,37]. However, relatively few studies have investigated SVs at the whole-genome level in fungi and/or attempted to associate them to phenotypic diversity [29,38,39].

The overall goal of the current study was to determine the extent to which SVs might influence the genome of *F. circinatum*. Our aims were four-fold: (i) sequencing of whole genomes for a set of geographically diverse *F. circinatum* isolates by making use of the short-read sequencing technology from Illumina® (Illumina, San Diego, California, USA), together with long-read sequencing technology from either PACBIO® (Pacific Biosciences, Menlo Park, California, USA) or MinION (Oxford Nanopore Technologies, Oxford Science Park, Oxford, United Kingdom); (ii) compilation of a comprehensive catalogue of fully characterized SVs occurring in the *F. circinatum* genome; (iii) annotation of genes

 doi:10.20944/preprints202204.0298.v1

associated with SVs; and (iv) characterization of the *F. circinatum* pangenome for demarcating conserved and non-conserved genomic regions. This study will thus provide a valuable resource for future investigations on the functional effects of SVs on *F. circinatum* and contribute immensely towards the understanding of the biology of this pathogen leading to the development of effective control mechanisms and management strategies.

2. Materials and Methods

Isolates

In this study we used seven isolates (FSP34, CMWF1803, CMWF560, CMWF567, UG10, UG27 and KS17) of *F. circinatum*. FSP34 was isolated from pitch canker-affected *Pinus* sp. in California, USA [40]. Isolates CMWF1803, CMWF560 and CMWF567 all originate from pitch canker-affected *Pinus* species in Mexico [41]. Isolates UG10 and UG27 were isolated from pitch canker-affected *P. greggii* trees in a plantation near Ugie in the Eastern Cape Province of South Africa [17]. Isolate KS17 was isolated from the diseased roots of a *P. radiata* seedling that were obtained during a study of the pathogen in a commercial seedling production nursery in the Western Cape Province of South Africa [42].

Genome sequencing, data sets and assembly

High-quality DNAs were extracted from the respective fungi using lyophilised mycelia with the protocol described by Murray and Thompson [43]. For isolates CMWF560, CMWF567, CMWF1803, UG10 and UG27, Pacbio long reads (10kb and 20kb libraries) were generated by Macrogen (Seoul, South Korea), and for isolates KS17 and FSP34, MinION long reads were available in-house at the Forestry and Agriculture Biotechnology Institute (FABI, Pretoria, South Africa). For all the isolates, Illumina HiSeq2000 250bp paired-end sequencing was performed by Macrogen.

The Pacbio reads were filtered using the SMRT® (Single Molecule, Real-Time) portal (Pacific Biosciences, Menlo Park, California, USA), while MinION read correction and trimming were done using CANU [44]. Illumina reads were trimmed and filtered using Trimmomatic v 0.38 [45] and FASTQC v 0.11.5 was used to check the quality of trimmed reads. *De novo* genome assemblies were then generated with long reads using CANU [44]. The assemblies were initially polished with Quiver v 2.2.2 [46] and Nanopolish [47] using Pacbio and MinION raw reads respectively. Final polishing was done with Pilon [48] using the Illumina HiSeq reads.

The polished scaffolds were ordered and oriented into contiguous pseudomolecules based on the macrosynteny found within the *Fusarium fujikuroi* species complex (FFSC) [49] using the LASTZ [50] plugin of Geneious v 7.0.4 [51]. The latter employed as references, the chromosome-level genome assemblies of *F. fujikuroi* [52] and *F. temperatum* [53]. Following this, redundant contigs that showed high similarity with chromosome scaffolds or other longer contigs were discarded.

Quality checking on the final genome assemblies was done by mapping individual reads to the genomes of their respective genome assemblies, and then performing a variant calling analysis. For this purpose, BWA MEM version 0.7.17-r1188 [54] with the -M option was used for the Illumina reads, and for the Pacbio and MinION reads we used CoNvex Gap-cost alignMents for Long Reads (NGMLR) version 0.2.7 [55]. After mapping, the SAM files were converted to BAM files using samtools view [56]. Read groups were then added using bamaddrg (https://github.com/ekg/bamaddrg), after which BAM files were sorted using bamtools sort [57]. Duplicated reads were marked with samtools rmdup. Samtools depth was used to determine the depth of coverage for each BAM file. The breadth of coverage was determined using samtools mpileup. For each genome, SVs were detected using Sniffles [55] with default settings.

 doi:10.20944/preprints202204.0298.v1

Completeness of the various genomes was estimated using BUSCO (Benchmarking Universal Single-Copy Orthologs) v 2.0.1 [58]. and the "Sordariomyceta" database containing 3725 genes. WebAUGUSTUS [59] was used to annotate the genomes with *F. graminearum* as the reference.

Identification and annotation of SVs

To identify SVs, the genome assembly of isolate FSP34 was used as the reference. FSP34 was chosen for this purpose as it was the first *F. circinatum* isolate to have its genome sequenced [60], and has since undergone several re-sequencing and assembly improvements [61]. The version used here was further improved to an assembly of 12 chromosome-level scaffolds and has 15 unmapped contigs. Quality-filtered reads from all sequencing platforms were used for mapping against the FSP34 genome assembly using BWA MEM for Illumina reads and NGMLR for Pacbio/ MinION. As described above, samtools view was used to convert SAM files to BAM files, after which read groups were added with bamaddrg. The BAM files were sorted with bamtools and duplicated reads marked with samtools rmdup. SVs were then identified using Sniffles with default settings. To filter variants from low mapping quality regions we used samtools view to extract the low mapping quality reads (MQ < 5) from sorted BAM files. Samtools depth was then used to compute the base coverage from sorted BAM files with low mapping quality reads. We then used SURVIVOR bincov [28] to cluster the coverage track into a BED file for filtering. SURVIVOR filter was then used to filter the VCF files, which were sorted and merged using SURVIVOR. The merged VCF file was then used to force call SVs across all samples. The resultant VCF files were merged to obtain a final multiple sample VCF file. Identified SVs were viewed with the Integrative Genome Viewer (IGV) v 2.4.14 [62].

To annotate the identified SVs with SnpEff [63], we first compiled a SnpEff database for *F. circinatum*. This was done using the annotation files obtained from WebAUGUSTUS. After SVs in the VCF files were annotated, we used SnpSift [64] to extract genes with high or moderate impact variants and genes from chromosomal regions with high variant density for further analysis. These genes were subjected to gene ontology (GO) enrichment analysis using the Fisher test (P value < 0.05) in the Blast2GO [65] plugin in CLC Genomic Workbench (Aarhus, Denmark). The list of all predicted genes from the FSP34 reference genome were used as the reference set for this analysis.

All reads that did not map to the FSP34 assembly were filtered from the respective BWA-MEM and NGMLR alignments. These were then assembled using SPAdes version 3.7.1 [66]. The assembled contigs were annotated using WebAUGUSTUS as described above. BLASTn and BLASTp analysis of these predicted annotations were conducted using NCBI.

Identification and characterization of micro variants

For the micro variant analysis, we specifically targeted SNPs, multiple nucleotide polymorphisms (MNPs) and indels in the 1-29 base pair (bp) range. For this purpose, the BAM files generated above were subjected to analysis with Freebayes version v1.2.0-2-g29c4002 [67] using the options "—ploidy 1", "--min-mapping-quality 30", "—min-base-quality 20" and default settings for the rest of the parameters. Filtering of generated variants was done using vcffilter (https://biopet.github.io/vcffilter/0.2). Variants with quality scores (QUAL) greater than 30 and with a minimum depth of coverage (DP) greater than 10 were filtered and removed. The identified micro-variants were then annotated using SnpEff, SnpSift and Blast2GO as described above.

Synteny and pangenome analyses

 doi:10.20944/preprints202204.0298.v1

Whole genome alignments were done using LASTZ as described above. Pairwise alignments between the genomes were visualized in Geneious. Synteny analysis was done using Synchro [68] with the protein FASTA files derived from WebAUGUSTUS annotations. Visualization of SVs was also done as graphic outputs from Synchro.

Spine [69] was used to reconstruct the pangenome of *F. circinatum* from the seven genomes examined. This program uses NUCmer [70,71] to align whole genomic sequences to build a nonredundant pangenome and extracts the core and accessory genomic elements as defined in the input parameters [69]. The program can take annotated genomes and output protein information online, but this function currently only works for small data sets such as prokaryotic genomes. For our purposes we downloaded the program and ran the analysis on our local servers. For this study, we used whole genome FASTA files and then searched for genes in the resultant data sets using WebAUGUSTUS, as described above. Similarly, BUSCO analysis was done on all predicted proteins.

Genomic sequences that were present in all seven genomes, with >85% similarity and a minimum output fragment of 100bp and a gap of 10bp between fragments, were regarded as belonging to the core sub-genomic compartment. To determine the distribution of accessory genomic elements among isolates, we used the ClustAGE [72] package in Spine. From the ClustAGE bins we selected bins that had sequences > 4.5kb for further analysis of their distribution patterns within the respective genomes. We focused on two groups of accessory genomic elements, the uniquely absent (sequences that were present in 6 isolates and missing only in 1 isolate) and the uniquely present (sequences that were present only in a single isolate). We used BLASTp analysis in NCBI to determine the origins or possible sources of these sequences. Geneious was used to map the accessory genomic elements to the genomes to check their distribution across the twelve chromosomes.

3. Results

3.1. Genome sequencing, data sets and assembly

We used long reads and short reads to construct five new and nearly complete chromosome-scale genome assemblies for *F. circinatum* (Table 1). Long read sequences generated with either PacBio or MinION, yielded a total of 1 557 037 filtered, corrected and trimmed reads for the seven isolates. Illumina sequencing generated a total of 64 129 476 reads after filtering and trimming. Most of the Illumina reads had a base quality sequence score greater than 30. Together with the two previously sequenced isolates (FSP34 and KS17), the total genome sizes ranged from 45 008 552bp to 43 828 286bp (Table 2). All genomes were assembled to the expected 12 linear chromosome scaffolds and the single circular mitochondrion scaffold. BUSCO analyses indicated that genome completeness ranged from 98.2% to 99.1% (Table 1). The number of complete genes predicted for these genomes by WebAUGUSTUS ranged from 13 854 to 14 382.

Table 1. Summary of whole nuclear genome assembly statistics.

Isolate	CMWF560	CMWF567	CMW1803	UG10	UG27
Accession number	JAEHFI000000000	JADZLS000000000	JAEHFH000000000	JAGJRQ000000000	JAELVK000000000
Genome size (bp)	46 691 343	45 984 420	46 810 763	44 774 968	45 546 500
Genome coverage[1]	35 (58)	65 (56)	54 (56)	31 (60)	43 (43)
N50	4 436 154bp	4 431 017bp	4 492 802bp	4 380 615bp	4 358 900bp
N75	3 211 240bp	3 566 220bp	3 263 251bp	3 014 022bp	3 202 209bp
L50	5	5	5	5	5
L75	8	8	8	8	8
% G+C	46.78	46.87	47.05	47.50	46.83
BUSCO (%)	99.0	99.0	99.1	98.9	99.1
Number of	13	12	13	12	12

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1

chromosomes					
Uncharacterised contigs	36	12	6	16	35
ORF	14 170	14 116	14 382	14 094	13 987

[1] Genome coverage for Pacbio sequencing, with Illumina sequencing indicated in brackets ().

Table 2. Size of the different scaffolds in each of the assemblies for the seven *Fusarium circinatum* isolates used in this study.

Scaffold[1]	***F. circinatum* isolate**[2]						
	FSP34	**CMWF560**	**CMWF567**	**CMWF1803**	**UG10**	**UG27**	**KS17**
Chr01	6 407 689	6 519 590	6 423 820	6 503 389	6 400 822	6 430 758	6 397 914
Chr02	5 066 197	5 005 606	5 040 005	5 011 558	4 859 649	4 976 076	4 709 326
Chr03	5 081 888	5 037 826	5 141 289	5 085 556	4 913 444	5 257 461	5 148 568
Chr04	4 313 168	4 556 945	4 479 672	4 551 627	4 383 147	4 236 059	4 401 926
Chr05	4 432 553	4 436 323	4 431 017	4 492 934	4 408 860	4 425 559	4 304 443
Chr06	4 301 895	4 284 508	4 281 318	4 282 150	4 261 236	4 358 679	4 219 930
Chr07	3 541 054	3 578 508	3 565 977	3 555 373	3 412 693	3 591 987	3 312 103
Chr08	3 172 915	3 210 569	3 718 348	3 263 924	3 015 014	3 202 361	3 066 990
Chr09	2 981 544	2 920 641	2 844 992	2 857 338	2 289 925	2 952 169	2 282 005
Chr10	2 698 820	2 714 737	2 642 782	2 649 870	2 413 675	2 564 176	2 483 521
Chr11	2 228 420	2 291 757	2 249799	2 266 931	2 087 508	2 247 110	2 291 537
Chr12	525 065	969 164	857 395	771 183	680 337	978 035	870 680
Chr13	-	536 197	-	1 045 802	-	-	-
UC	257 344	631 872	308 006	473 128	560 738	328 470	339 343

[1] Chr01-Chr13 corresponds to chromosomes 1-13. Data for all unassembled contigs for each genome were combined and indicated as UC.
[2] Scaffold sizes are indicated in bp, and the absence of Chr13 is indicated with -.

The results of our read mapping experiments suggested that the assemblies generated were robust. Mapping reads to their corresponding genome assemblies revealed that sequencing depth for long read PACBIO and MinION sequencing ranged from 14X to 74X coverage, while for short read Illumina sequencing ranged from 40X to 60X coverage. Also, no variants were detected on all of the seven genome assemblies when SV calling was done with reads mapped to their corresponding assemblies. It is thus unlikely that the genome assemblies have errors that could result in false calls of variants based on the reference genome and whole genome comparisons.

For all isolates, sequencing data that could not be assembled to the 12 chromosome scaffolds and mitochondrion ranged from 114 529bp for isolate FSP34 to 1 142 366bp for isolate UG10. The latter corresponded to 42 contigs ranging in size from 1 992bp to 65 074bp. Overall, however, unmapped contigs from all assemblies showed high similarity with fragments from the 12 core chromosomes (Figure 1), but due to their smaller size, we could not accurately position them in their respective chromosomes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1

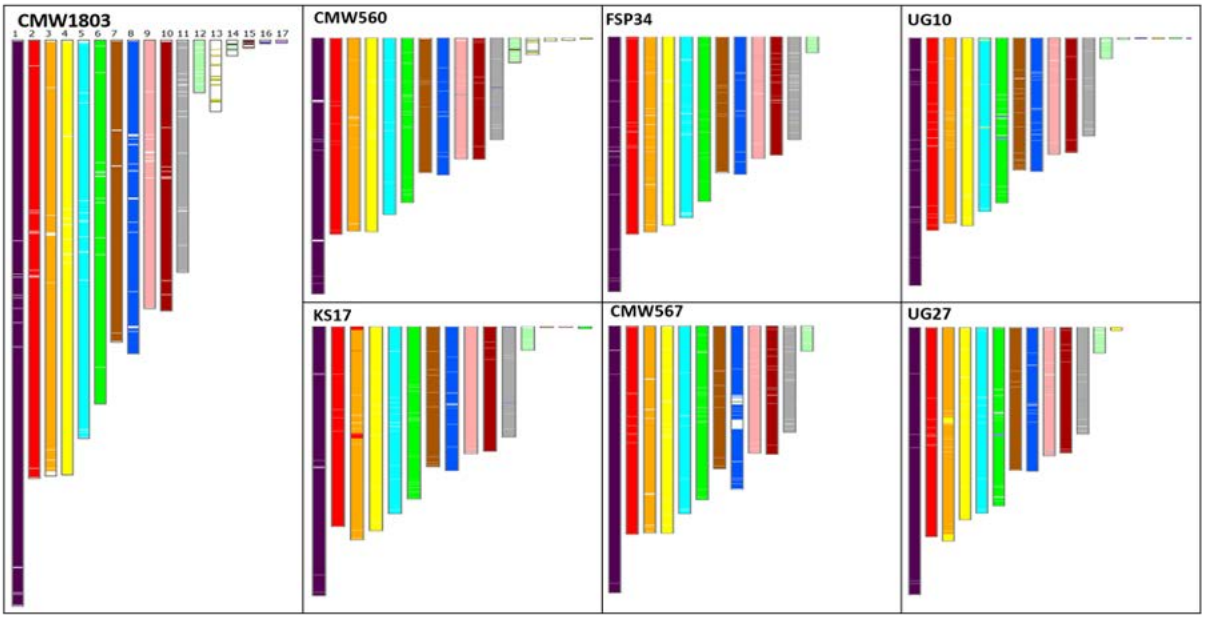


Figure 1. SynChro [68] generated chromosome painting, comparison of the isolates showing the 13th chromosome in CMWF1803 and CMWF560. No similar sequences were found in the other isolates.

Additionally, a scaffold that could not be assembled to any of the 12 chromosomes and that is larger in size (1045806bp) than chromosome 12 was observed for isolate CMWF1803. Comparisons with the other isolates revealed a similar scaffold with high similarity in isolate CMWF560 (Figure 1). These two isolates also had a higher overall genome size compared to the rest of the isolates. We therefore suggest that this scaffold represents an additional chromosome (Chr13) for *F. circinatum*.

3.2. Identification and annotation of SVs

Of the short reads generated for the different genomes, 90.41% to 99.0% mapped to the FSP34 assembly, while 93.7% to 99.2% of the long reads mapped to it (Table 3). Following construction of BAM files and multi-sample variant calling with Sniffles, we identified a total of 1 828 SVs, ranging in size from 30 bp to >10 000 bp (Table 4; Supplementary File S1). Of these, 1 717 were copy number variants (990 deletions, 719 insertions, 8 duplications) and 111 were inversions.

Table 3. Summary statistics for number of reads obtained and mapped to each genome for the different sequencing platforms.

Sequencing platform	Isolate	Total number of quality-filtered reads	Number reads mapped to the FSP34 reference assembly (%)
Illumina	FSP34	7 840 006	7 762 015 (99.0)
	CMWF560	9 471 099	8 931 113 (94.0)
	CMWF567	9 482 308	9 012 583 (95.1)
	CMWF1803	9 860 143	9 028 628 (91.6)
	UG10	9 737 181	9 394 412 (96.5)
	UG27	9 743 693	8 934 966 (91.7)
	KS17	7 995 046	7 228 461 (90.4)
PacBio	CMWF560	300 889	281 873 (93.7)
	CMWF567	187 327	181 372 (96.8)
	CMWF1803	194 164	183 210 (94.4)

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1

	UG10	256 808	247 772 (96.5))
	UG27	357 305	339 326 (95.0)
MinION	KS17	95 510	92 336 (96.7)
	FSP34	165 477	164 130 (99.2)

Table 4. Micro variant and structural variant (SV) rate details per chromosome.

Chromosome	Micro variants [1]		SVs [2]			
	Number of variants	Variant rate[3]	Number of deletions	Number of duplications	Number of inversions	Number of insertions
Chr01	41 809	153	99	1	10	83
Chr02	44 429	114	104	2	14	72
Chr03	35 947	141	97	0	18	69
Chr04	42 594	101	80	1	4	71
Chr05	36 398	121	81	0	8	67
Chr06	44 223	97	112	0	11	68
Chr07	36 258	97	75	3	8	63
Chr08	43 130	73	97	0	4	61
Chr09	31 971	93	60	0	6	44
Chr10	43 870	61	80	0	8	39
Chr11	36 755	60	68	1	10	49
Chr12	17 314	30	33	0	2	31
Total	***46 1683***	***97***	***986***	***8***	***103***	***717***

[1] Micro variants were determined using Freebayes.
[2] SVs were determined using Sniffles
[3] Variant rate determined by number of variants per base (total chromosome size/total number of variants).

Annotation with SnpEff predicted a total of 476 229 functional effects for the collection of SVs identified (Table 5; Supplementary File S2). SnpEff also provided an indication of the magnitude of the predicted effects. Variants with high impact are disruptive with regards to probable protein function, moderate impact variants are less disruptive would likely reduce functionality of the protein, low impact variants would be mostly harmless, and modifier variants often occur in non-coding genes, introns, intergenic regions, intragenic regions and downstream to genes where predictions of the effect are more difficult [63]. For the SVs identified here, a greater proportion (88.2%) were predicted to have high impact, 10.5% moderate impact, 0.0% (a single effect count) low impact and 1.3% have modifier impact. The high impact effect variant types included bidirectional gene fusions (46.6 %), chromosome number variations (0.002 %), exon loss variants (0.02 %), feature ablations (1.6 %), frame shift variants (0.03 %), gene fusions (37.9 %), inversions (10.5 %), splice acceptor variants (0.003 %), splice donor variants (0.004 %), start losses (0.01 %), stop losses (0.008 %) and transcript ablations (2.0 %). Moderate impact SV variant types included conservative in frame deletions (0.02 %), disruptive in frame deletions (0.01 %) and upstream gene variants (0.5 %). Modifier impact variant types predicted included downstream gene variants (0.5 %), exon regions (0.0 %), intergenic regions (0.13 %), intergenic variants (0.08 %), intron variants (0.01 %), non-coding transcripts variants (0.07 %) and splice region variants (0.01 %).

Table 5. SnpEff annotation summary for macro structural variants (SVs) indicating number of effects. The impact of the different types of effects are indicated in differing colours: High impact (red), moderate impact (orange), low impact (green) and modifier effect (yellow).

Effect classification	Fields	SnpEff count	Percentage (%)

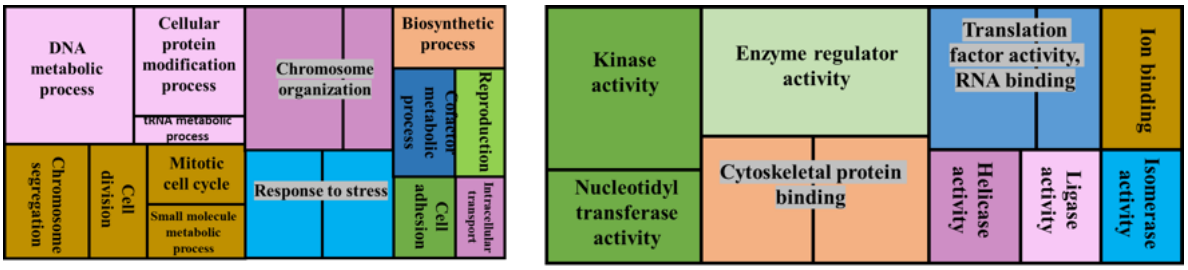
 doi:10.20944/preprints202204.0298.v1

Number of effects by impact	High	420 017	88.2
	Low	1	0.0
	Moderate	49 895	10.5
	Modifier	6 316	1.3
Number of effects by type	Bidirectional gene fusion	222 121	46.6
	Chromosome number variation	10	0.002
	Conservative in frame deletion	103	0.02
	Disruptive in frame deletion	57	0.01
	Downstream gene variant	2 499	0.5
	Duplication	1	0.0
	Exon loss variant	100	0.02
	Exon region	2	0.0
	Feature ablation	7 724	1.6
	Frame shift variant	149	0.03
	Gene fusion	180 468	37.9
	Intergenic region	608	0.13
	Intragenic variant	367	0.08
	Intron variant	49	0.01
	Inversion	49 863	10.5
	Non-coding transcript variant	310	0.07
	Splice acceptor variant	16	0.003
	Splice donor variant	19	0.004
	Splice region variant	67	0.01
	Splice site region	1	0.0
	Start lost	45	0.01
	Stop gained	1	0.0
	Stop lost	39	0.008
	Transcript ablation	9 316	2.0
	Upstream gene variant	2 515	0.5
Number of effects by region	Chromosome	49	0.01
	Downstream	2 499	0.53
	Exon	449	0.1
	Gene	438 343	92.1
	Intergenic	608	0.1
	Intron	15	0.003
	Splice site acceptor	5	0.001
	Splice site donor	6	0.001
	Splice site region	1	0.0
	Transcript	31 739	6.7
	Upstream	2 515	0.5

A total of 4 190 proteins were annotated in the regions that were predicted to have deletions, 345 in regions displaying insertions, 8 885 were in inversion areas and 1 protein in an area displaying duplications (Supplementary File S3). The following biological processes were overrepresented in the gene set of proteins annotated in regions associated with deletions (Figure 2): DNA metabolism, cellular protein modification process, chromosome organization, response to stress, signal transduction, macromolecular complex assembly, tRNA metabolism, chromosome segregation, cell division, mitotic cell cycle, small molecule metabolism, biosynthesis, cofactor metabolism, reproduction, cell adhesion and intracellular transport. The overrepresented cellular components with the deletions were macromolecular complex, nucleoplasm, chromosome, ribosome, endosome,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1

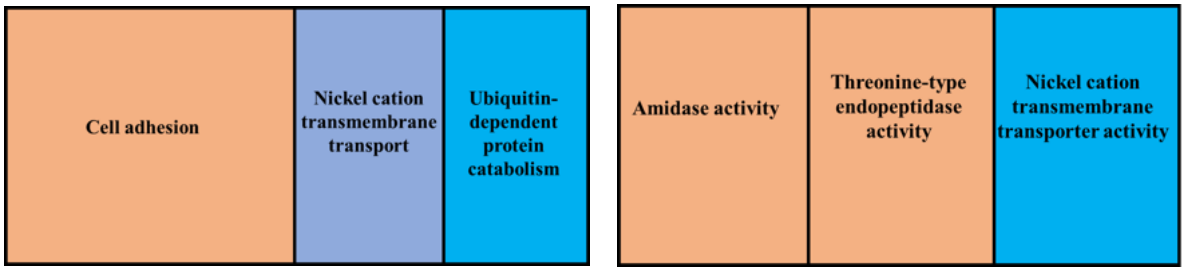
cytosol, cytoskeleton, mitochondrion and nuclear envelope. Molecular functions overrepresented within proteins displaying deletions were enzyme regulator activity, kinase activity, nucleotidyltransferase activity, cytoskeleton protein binding, enzyme binding, protein binding, ion binding, helicase activity, rRNA binding, ligase activity, isomerase activity, rRNA binding and translation factor binding.



(**a**) (**b**)

Figure 2. GO enrichment analysis for genes associated with deletions showing overrepresented GO terms, biological processes (**a**) and molecular functions (**b**). The whole genome annotations of the reference FSP34 was used as the reference set, genes associated with deletions were used as the test set.

Three biological processes were overrepresented in the predicted gene set that was predicted to be affected by insertions (Figure 3). These were cell adhesion, nickel cation transmembrane transport and ubiquitin-dependent protein catabolism. Cellular components overrepresented in insertions were integral component of membrane, proteasome core complex and alpha subunit. Dominant molecular functions in this gene set were threonine-type endopeptidase activity, amidase activity, nickel cation transmembrane and transporter activity.

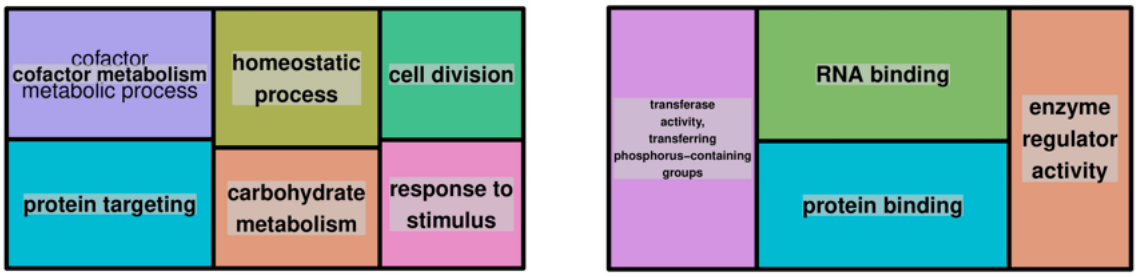


(**a**) (**b**)

Figure 3. GO enrichment analysis for genes affected by insertions showing overrepresented GO terms, biological processes (**a**) and molecular functions (**b**). The whole genome annotations of the reference FSP34 was used as the reference set, genes affected by insertions were used as the test set.

Six biological processes and four cellular components were overrepresented in the genes affected by inversions (Figure 4). The biological processes were cofactor metabolism, protein targeting, homeostatic process, carbohydrate metabolism, cell division and response to stimulus. Overrepresented molecular functions were transferase activity, transferring phosphorus containing groups, RNA binding, protein binding and enzyme regulator activity.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1



(**a**) (**b**)

Figure 4. GO enrichment analysis for genes associated with inversions showing overrepresented GO terms, biological processes (**a**) and molecular functions (**b**). The whole genome annotations of the reference FSP34 was used as the reference set, genes within inverted regions were used as the test set.

The single gene predicted to be affected by duplications encoded an RNA binding protein with helicase activity, which is a ubiquitous family of proteins that potentially play a role in cellular processes that involve RNA metabolism [73]. The unmapped reads represented loci that were absent in the reference strain and/or contaminants in our samples. BLASTp analysis retained top hits from *Fusarium* species. Genes annotated from these sequences were mainly for proteins involved in transport and metabolism (Figure 5).

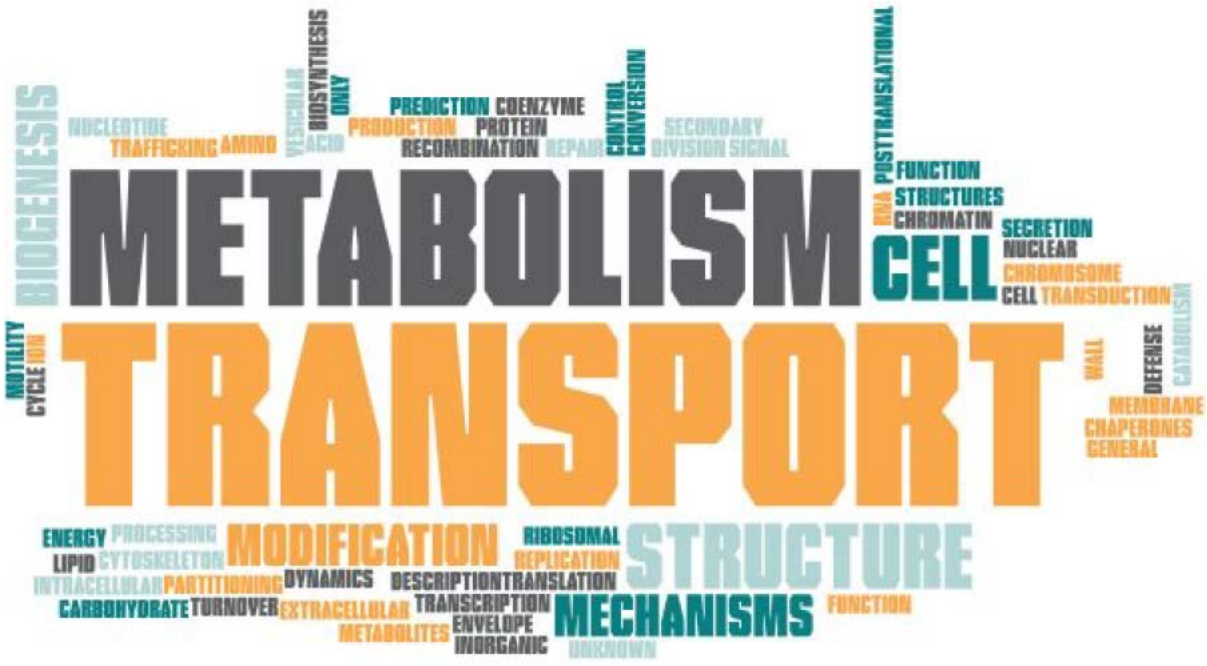


Figure 5. Protein classes coded by genes found in assembled sequences that did not map to the FSP34 genome assembly.

3.3. Identification and annotation of micro variants

Multi-sample variant calling using Freebayes generated a total of 397 704 polymorphic sites among the genomes examined (Supplementary File S4). Of these 353 117 (88.8%) were biallelic and 44 587 (11.2%) were multiallelic. From the collection of polymorphisms, a total of 461 683 micro variants were identified, of which 68.7% were SNPs, 23.1% were 1-29 MNPs, 3.2% were insertions, 2.4% were deletions and 2.7% were mixed. The observed genome-wide transitions/transversions (Ts/Tv) ratio was 2.9 and the overall variant rate

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1

was 1 variant for every 97 bases. The observed Ts/Tv ratio is consistent with the phenomenon that transitions generally occur at a higher frequency than transversions [74]. The whole genome SNP density showed an uneven distribution of SNPs across chromosome lengths. There were high SNP densities on chromosome arms while the middle regions showed relatively lower SNP densities (Figure 6). For chromosome 11 and 12 SNP densities were relatively high across full chromosome lengths.

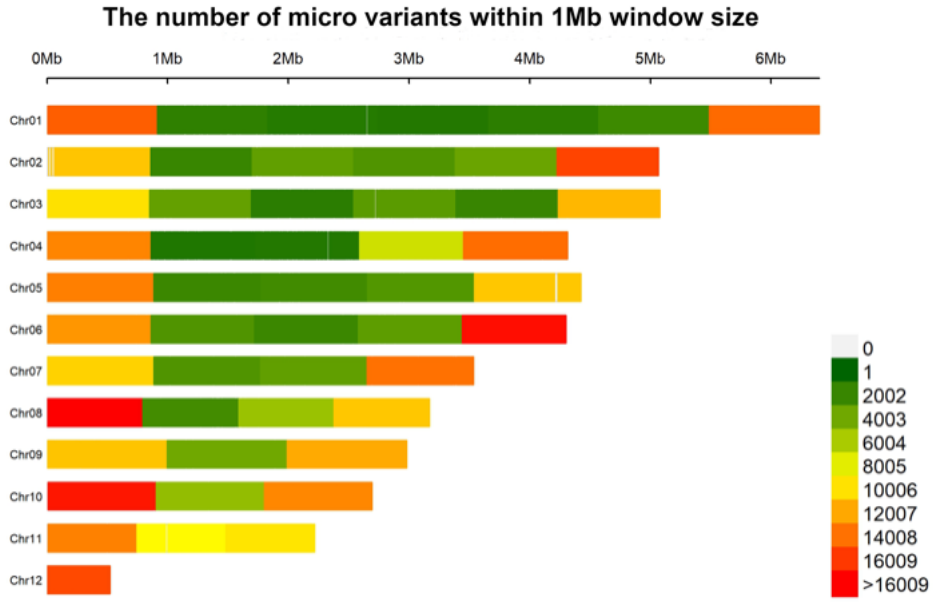


Figure 6. Micro variant density distribution across chromosome lengths of *F. circinatum* FSP34 relative to the other 6 genomes showing regions of higher variant density in chromosome arms and sparse distribution within the inner central parts of the chromosomes. The figure was generated using CM plot [75].

A total of 2 719 213 putative effects were assigned to a collection of micro variants. Of these, SnpEff predicted 0.7 % of the effects to have high impact, 2.3 % to have moderate impact, 2.8 % to have low impact and 94.3 % to have a modifier effect. In terms of functional classes 42.0 % of the effects were classified as missense, 0.9 % nonsense and 57.1 % to be silent variants (Table 6 and Supplementary File S5). High impact variants included frameshifts, gene fusions, splice acceptors, splice donors, start losses, stop gains and stop losses. Moderate impact variants predicted were conservative in frame deletions, conservative in frame insertions, disruptive in frame deletions, disruptive in frame insertions, missense variants and upstream gene variants. Low impact variants included initiator codon variants, splice region variants, stop retained variants and synonymous variants. Modifier variants included were downstream gene variants, intergenic region, intragenic variants, intron variants and non-coding transcript variants.

Table 6. SnpEff annotation summary of micro variants indicating number of effects (Freebayes). The impact of the different types of effects are indicated in differing colours: High impact (red), moderate impact (orange), low impact (green) and modifier effect (yellow).

Effect classification	Fields	SnpEff count	Percentage (%)
Number of effects by impact	High	18 047	0.7
	Moderate	63 033	2.3
	Low	74 845	2.8
	Modifier	2 563 288	94.3
Number of effects by functional class	Missense	46 372	42.0
	Nonsense	1 027	0.9

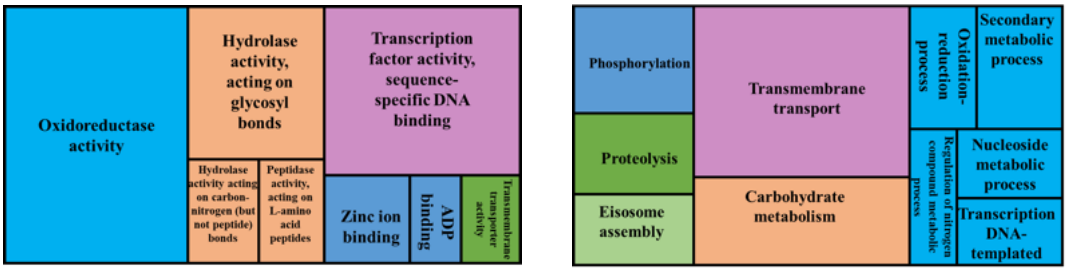
 doi:10.20944/preprints202204.0298.v1

	Silent	62 991	57.1
Number of effects by type	Conservative in frame deletion	404	0.02
	Conservative in frame insertion	396	0.02
	Disruptive in frame deletion	354	0.01
	Disruptive in frame insertion	216	0.01
	Downstream gene variant	977 631	35.9
	Frameshift variant	1 881	0.07
	Gene fusion	13 881	0.5
	Initiator codon variant	17	0.001
	Intergenic region	277 047	10.2
	Intragenic variant	168 282	6.2
	Intron variant	23 572	0.9
	Missense variant	62 233	2.3
	Non-coding transcript variant	133 146	4.9
	Splice acceptor variant	287	0.01
	Splice donor variant	325	0.01
	Splice region variant	5 689	0.2
	Start lost	110	0.004
	Stop gained	1 434	0.05
	Stop lost	269	0.01
	Stop retained variant	170	0.006
	Synonymous variant	71 054	2.6
	Upstream gene variant	988 027	36.2
Number of effects by region	Downstream	977 631	36.0
	Exon	136 959	5.0
	Gene	13 881	0.5
	Intergenic	277 047	10.2
	Intron	19 155	0.7
	Splice site acceptor	266	0.01
	Splice site donor	301	0.01
	Splice site region	4 518	0.2
	Transcript	301 428	0.0
	Upstream	988 027	36.3

The bulk of the effects predicted were outside coding regions, in downstream, intergenic, intron and upstream regions. A total of 136 959 effects were predicted to impact on exons, 13 881 on genes and 30 148 on coding and non-coding transcripts. Within the splice sites 266 effects were predicted on the splice site acceptor region, 301 splice site donor and 4 518 within the splice site region.

GO enrichment analysis showed that genes involved in the following biological processes were overrepresented (P>0.05) in chromosome regions with high micro variant densities (Figure 7): transmembrane transport, carbohydrate metabolism, phosphorylation, proteolysis, eisosome assembly, secondary metabolism, oxidation-reduction process, nucleoside metabolism, transcription, DNA-templated and regulation of nitrogen compound metabolism. These overrepresented genes are localised within the integral component of membrane, the extracellular space and the eisosome. Overrepresented molecular functions within this gene set were oxidoreductase activity, transcription factor activity, sequence-specific DNA binding, hydrolase activity, acting on glycosyl bonds, hydrolase activity, activity on carbon-nitrogen (but not peptide) bonds, peptidase activity, acting on L-amino acid peptides, zinc ion binding, ADP binding and transmembrane transporter.

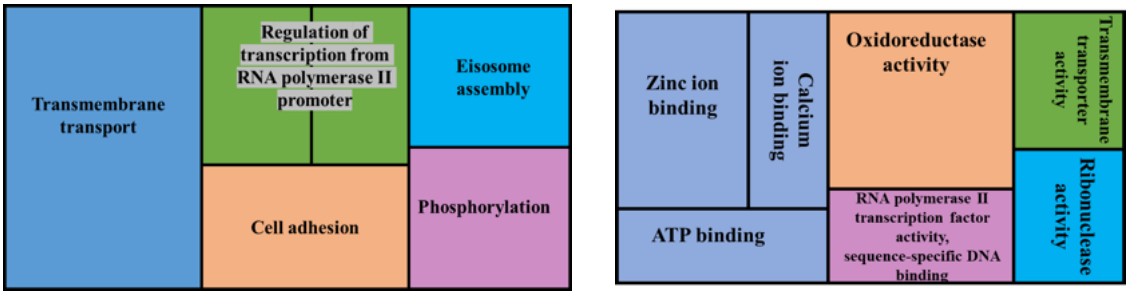
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1



(**a**) (**b**)

Figure 7. GO enrichment analysis for genes within high variant density chromosomal regions showing overrepresented GO terms, biological processes (**a**) and molecular functions (**b**). The whole genome annotations of the reference FSP34 was used as the reference set, genes within high variant density regions were used as the test set.

Genes that were overrepresented in gene sets affected by high impact micro variants included the following biological processes: transmembrane transport, regulation of transcription from RNA polymerase II promotor, RNA phosphodiester bond hydrolysis, eisosome assembly, cell adhesion and phosphorylation (Figure 8). The overrepresented cellular components were integral part of the membrane and eisosome. Overrepresented molecular functions included oxidoreductase activity, zinc ion binding, calcium ion binding, ATP binding, RNA polymerase II transcription factor activity, sequence-specific DNA binding, transmembrane transporter activity and ribonuclease activity.

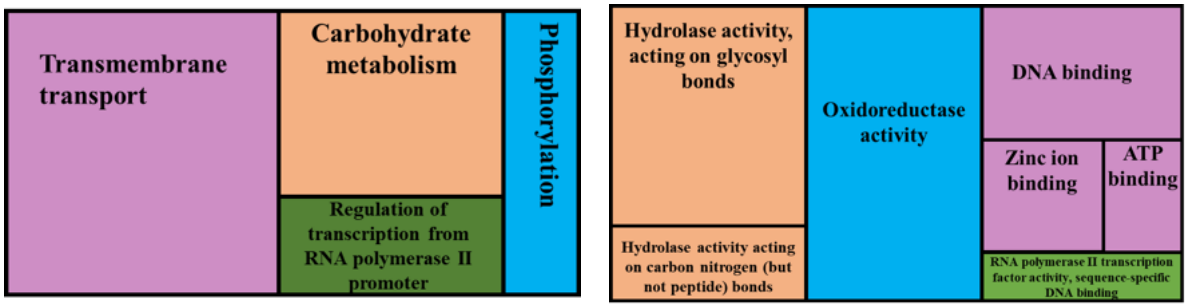


(**a**) (**b**)

Figure 8. GO enrichment analysis of genes predicted to have high impact variants showing overrepresented GO terms, biological processes (**a**) and molecular functions (**b**). The whole genome annotations of the reference FSP34 were used as the reference set and genes predicted to have high impact variants were used as the test set.

Proteins classes that were overrepresented in regions affected by moderate impact micro variants (Figure 9) included these biological processes, transmembrane transport, carbohydrate metabolism, regulation of transcription from RNA polymerase II promoter and phosphorylation. Molecular functions overrepresented in genes affected by moderate variants are, oxidoreductase activity, hydrolase activity, acting on glycosyl bonds, DNA binding, hydrolase activity, acting on carbo-nitrogen (but not peptide) bonds, zinc ion binding, ATP binding, RNA polymerase II transcription factor activity and sequence specific DNA binding.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 29 April 2022 doi:10.20944/preprints202204.0298.v1



(**a**) (**b**)

Figure 9. GO enrichment analysis for genes predicted to have moderate impact variants showing overrepresented GO terms, biological processes (**a**) and molecular functions (**b**). The whole genome annotations of the reference FSP34 were used as the test set and the genes with moderate impact variants were the test set.

3.4. Synteny and pangenome analyses

Using Spine, a nonredundant pangenome of 50 076 541bp was built from the seven genomes used in this study. WebAUGUSTUS predicted a total of 15 099 complete genes from the pangenome. It was 96.7% complete according to BUSCO with 76 fragmented BUSCOs and 47 BUSCO genes were missing. The percentage completeness of the pangenome is lower than the individual genomes of the seven isolates, which is indicative of the fragmentation of the pangenome. Analyses of the combined proteins annotated from the seven isolates indicated that there were 18 BUSCO genes missing based on the "Sordariomyceta" database. BLASTp analyses of the 18 missing BUSCOs revealed that only four of the proteins had top hits to *Fusarium* species, suggesting that these missing genes may not be common among *Fusarium* species. A total combined number of annotated proteins from the seven genomes was 98 484 of which 99.2% were clustered into orthogroups [76] with an average of 7.1 genes per orthogroup. Sequences from all isolates were present in 12 424 orthogroups, of which 11 712 represented single copy gene families. A combined total of 765 genes were not assigned to any orthogroup. These were the isolate-specific genes that could have been acquired through HGT.

Whole genome alignments using Spine resulted in a backbone (core) of 42 260 189 bp that was present with a greater than 85% nucleotide similarity, in all seven of the genomes (Table 7). This core comprised an average of 93% of all the seven genomes used in this study and had a GC content of 47.2%. A total of 13 282 complete genes were predicted by WebAUGUSTUS in the core genome. Comparison against BUSCO's "Sordariomyceta" databased indicated that the core genome was 95.6% complete. BLASTp analysis of the core proteins, excluding hits on *F. circinatum*, showed 13 182 proteins had top hits to *Fusarium* species, 48 proteins had top hits to non-*Fusarium* species and 52 predicted proteins had no blast hits to the NCBI database (Supplementary File S6). Non-*Fusarium* top hits included species from the genera *Neonectria*, *Trichoderma*, *Exophlala* and *Aspergillus*. Biological processes overrepresented within the set of 48 non-*Fusarium* proteins were regulation of transcription by RNA polymerase II, methylation and phosphorylation. Overrepresented molecular functions within the 48 proteins were DNA binding transcription factor activity, RNA polymerase II specific and zinc ion binding.

Table 7. Spine statistics showing the size of the accessory, core and pangenome elements of the *F. circinatum* isolates.

Isolate	Source	Total bp	GC %
FSP34	Accessory	2 669 713	44.19
	Core	42 260 189	47.20
CMWF560	Accessory	3 831 939	45.05
	Core	42 633 795	47.02

 doi:10.20944/preprints202204.0298.v1

CMWF567	Accessory	3 290 975	45.14
	Core	42 465 505	47.06
CMWF1803	Accessory	4 359 907	45.27
	Core	42 501 744	47.22
KS17	Accessory	2 369 668	44.57
	Core	42 008 433	47.42
UG10	Accessory	2 733 552	45.07
	Core	41 905 489	47.72
UG27	Accessory	2 919 797	44.25
	Core	42 449 440	47.06
	Backbone	42 260 189	47.20
	Pangenome	50 076 541	46.85

An average of 7% of the genome sequence across the seven isolates was predicted to constitute the accessory genomic elements (Supplementary File S7). These were distributed across chromosome lengths, with chromosomes 1 to 11 showing higher densities in sub-telomeric regions while chromosomes 12 and 13 had relatively higher densities across full chromosome lengths (Supplementary Figure 1). The number of complete genes in the accessory genomic elements of the isolates was 58 881 of which 64 were complete BUSCOs. The combined accessory genomic elements from all the isolates amounted to a total of 22 175 551 bp with an average GC content of 44.8%. ClustAGE grouped sequences greater than 200bp into 2 023 unique bins of similar sequences.

A total of 632 genes were annotated on sequences that was present only in one of the genomes examined. Of these, 596 had best BLAST hits in other *Fusarium* species, 15 had similarity in non-*Fusarium* species and 20 genes returned no significant result using BLAST. Isolate CMWF1803 had the highest number of such uniquely present genes (240) while none were found in isolate UG10. Isolates CMWF560, CMWF567, FSP34, KS17 and UG27 had 148, 143, 57, 40 and 3 uniquely present genes, respectively. From the sequences that were absent in only one of the genomes examined, a total of 466 genes were annotated. Of these, 445 had best BLAST hits to other *Fusarium* species and 21 had best BLAST hits to non-*Fusarium* species. Isolate FSP34 had the highest number of uniquely absent genes (150) followed by KS17 (140) while none were found in isolate UG27. Isolates CMWF560, CMWF567, CMWF1803 and UG10 had 72, 47, 21 and 36 uniquely absent genes, respectively.

A total of 18 biological processes were overrepresented in the uniquely present proteins. These included tRNA metabolism, chromosome segregation, cellular protein modifications, ribonucleoprotein complex biogenesis, anatomical structure formation involved in morphogenesis, nucleocytoplasmic transport, chromosome organization, cell differentiation, nucleocytoplasmic transport, translation, DNA metabolism, protein transport, sulphur compound metabolism, response to stress, cofactor metabolism and cellular catabolism. Dominant molecular functions in this set were cytoskeletal protein binding, structural constituent of ribosome, RNA binding, phosphatase activity, nuclease activity, GTPase activity, protein binding, bridging and ubiquitin like protein binding. Overrepresented cellular components in this set were macromolecular complex, cytoskeleton, nuclear chromosome, mitochondrion, Golgi apparatus, ribosome, nuclear envelope and extracellular matrix.

Among the uniquely absent proteins, the following biological processes were overrepresented: mitotic cell cycle, tRNA metabolism, reproduction, anatomical structure morphogenesis, chromosome segregation, cellular protein modification process and organelle organization. Molecular functions that were overrepresented in this set included peptidase activity, nuclease activity and phosphatase activity. The dominant cellular components were macromolecular complex, nuclear chromosome, extracellular matrix and endomembrane system.

 doi:10.20944/preprints202204.0298.v1

4. Discussion

In this study, we used multi-sample variant calling to develop a catalogue of intra-species SVs for the pine pitch canker fungus, *F. circinatum*. This type of resource is relatively uncommon for fungi, as most previous studies explored genomic variation at the level of inter-species [2,5]. These precious comparative genomics studies have provided insights into the biology of closely related species and improved our understanding of how they evolve [5,77]. By contrast, studies on the intra-species genetic variation provides insights into how individuals of the same species accumulate genetic differences and how these affect their adaptability [14,78,79]. As genomes of many isolates of the same species become available it will become routine for whole genome comparative studies to be conducted and enable phenotypic traits to be associated with genotypic variations.

To facilitate compilation of our catalogue of SVs, we first had to ensure that we had access to high-quality reference assemblies. Most variant calling platforms use a reference genome to identify polymorphisms in other genomes of interest [12,67]. However, considering the broad structural variation that exists within fungal genomes [80], reference genomes can be limiting in variant calling if the assemble is incomplete or if it has assembly errors, which can lead to mapping errors that distort variant calling [81]. Indeed, it is widely recognized that the availability of good reference genomes against which comparisons can be made represent one of the main limitations of efficient and accurate in variant calling [82]. The current genome assembly for *F. circinatum* isolate FSP34 is near-complete, containing 12 chromosome scaffolds and the mitochondrion scaffold [61]. This isolate has also been used in various previous studies [49,83-85] and has emerged as the "model" strain for *F. circinatum* genomics studies.

In addition to the FSP34 reference genome, we also report here the high-quality genome assemblies for five additional *F. circinatum* strains. They have different origins, with three (CMWF560, CMWF567 and CMWF1803) originating from pitch canker-affected *Pinus* spp. tissue collected in Mexico [14] and two (UG10 and UG27) that were isolated from pitch canker-affected *P. greggii* tissue in South Africa [17]. The strains also have different mating types. *Fusarium circinatum* is heterothallic [86], with three of the isolates (CMWF560, CMWF567 and CMWF1803) being MAT 1-1 and two (UG10 and UG27) MAT 1-2. The five additional genomes were also assembled to near completion, with all 12 of the expected chromosome scaffolds. The 12th chromosome is the smallest and apparently dispensable [2,26,87].

In two of our new genome assemblies, we detected a scaffold that likely represents a 13th chromosome. These two chromosome-sized scaffolds were found among the uncharacterized scaffolds in the CMWF560 and CMWF1803 sequence data. Subsequent analyses confirmed that it was found only in isolates CMWF560 and CMWF1803, but not in the other five *F. circinatum* isolates including FSP34. Chromosome number variation has been reported within species of the FFSC, where at least two dispensable or accessory chromosomes have been visualised [88]. We thus propose that the uncharacterised scaffolds identified here is an additional accessory chromosome, which we call chromosome 13. The 13th chromosome in CMWF1803 was the larger of the two at 1 045 806 bp, which is bigger than chromosome 12 assembled across all the isolates. The overall genome sizes of isolates CMWF560 and CMWF1803 were generally bigger than the rest of the isolates that did not have sequences resembling the 13th chromosome.

Our results showed that micro variants were unevenly distributed across the lengths of chromosomes 1-10. In these chromosomes, we observed higher densities over the first 1Mb of chromosome ends, while they were more sparsely distributed in the middle regions of the chromosomes. This distribution pattern is consistent with previous reports of high variant density on sub-telomeric regions [31,32], which are widely recognized as hot spots for genome plasticity in eukaryotes. This is because telomeres accumulate variants during replication as they replicate differently from the inner regions of chromosomes [89]. Also, sub-telomeric regions are prone to recombination and are thus likely to accumulate mutations [90]. Contrast to chromosomes 1-10, chromosomes 11 and 12 generally

 doi:10.20944/preprints202204.0298.v1

had high micro variant densities across full chromosome lengths. The sequence and size variability of chromosome 12 [61,87], and the presence of a large reciprocal translocation and an inversion in chromosome 11 [49,53], make chromosomes 11 and 12 the most variable of the chromosomes within the FFSC [53]. This is evidenced in our results as the higher variant density across the full chromosome lengths.

Copy number variants (insertions, deletions and duplications) ranging SVs as small as 30bps to large ones of >100kbs were detected in this study. These types of variants are known to play a major role in fungal adaptation [29,39]. Duplications can influence expression levels of target regions, e.g., duplication of regions containing biosynthetic genes can increase production of the relevant compounds [91-93], or duplication of certain genes can increase resistance to toxic compounds [94]. In this study, we identified a gene encoding an RNA binding protein with helicase activity arising by duplication. Furthermore, insertions can drive the accumulation of novel genetic traits, while deletions highlight genomic regions that might be dispensable as pathogenic fungi often shed genomic regions to avoid detection by hosts [6]. Therefore, the copy number variants identified in this study provide a robust starting point for further research into the biology of the pitch canker fungus.

Structural variants that impact genes can lead to functional variation among individuals and influence phenotypic traits. We used SnpEff [63] to access the significance of the identified variants based on gene models from WebAUGUSTUS annotations [59]. Although the accuracy of our gene models still needs to be verified, the identified putative effects provide a basis for future association studies using these isolates. The most prevalent high impact variant effect was gene fusion. Clustering of genes involved in the same metabolic pathway, biological process or structural complex is a well-known phenomenon in fungi [95]. This has a beneficial effect because it allows for the physical coupling of proteins that are functionally related. Inversion was also a common high impact macro structural variant effect. Inversions are indicative of genes whose transcriptional orientation along chromosomes is different [96]. Frameshift mutations are indels within the coding regions of genes that alter the reading frame resulting in the codons downstream being altered. In many cases the deletion or insertion results in truncated gene products [97]. Splice site mutations would result in abnormal splicing leading to base changes in the processed mRNA [98,99]. The phenotypic effect of all these mutations on the different isolates of *F. circinatum* needs to be investigated. GO enrichment analyses showed that genes involved in transcription, DNA binding, transmembrane transport, oxidoreductase activity, hydrolase activity, acting on glycosyl bonds, ion binding, metal binding and cation binding were overrepresented in regions affected by high and moderate impact variants. These classes of genes are mainly involved with proteins needed by the fungus when interacting with its environment. The presence of variants in these gene classes highlights the pathogen's propensity to adapt to its environment.

We took a sequence centric approach [100] in building the pangenome of *F. circinatum*. We characterized the core, accessory and pangenomes based on nucleotide sequence rather than only the protein coding portion of the genomes. This approach ensured the inclusion of noncoding genomic elements in the core, accessory and pangenomes. This methodology allowed us to avoid the bias and errors of gene prediction models in the identification of genomic elements that are conserved in *F. circinatum*. Our BLAST analysis showed that within the core genome there are some proteins which did not return significant hits to closely related *Fusarium* species, but rather to other distantly related species. These genomic elements may have been acquired by *F. circinatum* from other species through horizontal gene transfer [101] and have been subsequently maintained, possibly as they have provided the species some advantage. Gene gains can also be attributable to *de novo* gene emergence from non-coding DNA. There were also some proteins that did not return any BLAST hits. These were mainly proteins of unknown function.

These orphan genes which lack homologues in other lineages can also arise through duplications followed by diversification [101]. The proteins that are unique to *F. circinatum*

and are conserved in all the isolates could be involved in lineage specific adaptations and are potential targets for species-specific pathogen management and control strategies.

Within the accessory genome, isolate-specific open reading frames present in only one isolate and not the others (uniquely present) also showed different patterns of possible origins. Predicted proteins that had BLAST hits with closely related *Fusarium* species could indicate genomic elements that have been maintained in some isolates but have been dispensed in others. These genomic regions could have arisen through high impact variants which include deletions, loss of function mutations such as frame shifts and early stop codons resulting in truncated proteins. Even if truncated proteins are expressed in different isolates, they are less likely to have exactly the same function as full-length proteins. Therefore, the functional variation of truncated proteins could have phenotypic effects on different isolates. Isolate specific sequences with best BLAST hits to other non-*Fusarium* species could be indicative of recently gained genomic elements that have not been fixed in *F. circinatum* populations. The accessory genome elements also had a reduced GC content, in comparison to the core genome. There were more accessory elements mapping to chromosomal regions that also had high SNP density. Accessory genomic elements were associated with AT rich and highly variable genomic regions.

5. Conclusions

The genomic structural variants identified in *F. circinatum* range from SNPs to chromosome scale SVs, that include insertions, deletions and inversions. It is possible that some of these variations are neutral with little or no impact on the biology of this pine pathogen. However, it is also conceivable some of the SVs described in this study have profound effects on the biology, evolution and niche adaptation of this pine pathogen. The genetic variations could then impact the disease management and control strategies used in nurseries and plantations as these strategies need to be applicable to strains of the pathogen notwithstanding variations between the strains. This study therefore provides a resource for future association studies within *F. circinatum*. Progress in understanding the exact impact of these genetic variations that exist within *F. circinatum* will also benefit our understanding of the biology of the pitch canker fungus.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: Distribution of accessory genome elements across chromosomes of isolate CMW1803. Accessory genome elements are generally sparsely distributed within the middle parts of the chromosomes while the chromosome arms show a high density; File S1: VCF file showing structural variants (SVs) identified in *F. circinatum*. Multi-sample variant calling was done using Sniffles. A total of 1 828 SVs ranging in size from 30 bp to >10 000 bp were identified. These include 1 717 copy number variants (990 deletions, 719 insertions, 8 duplications) and 111 inversions; File S2: Structural variant annotation output file from SnpEff; File S3: A total of 476 229 functional effects for the collection of SVs were identified. List of proteins annotated in the regions that were predicted to have deletions, in regions displaying insertions, in inversion areas and in areas displaying duplications; File S4: VCF file showing micro variants identified in *F. circinatum* following a multi-sample variant calling using Freebayes. A total of 397 704 polymorphic sites were identified among the genomes examined; File S5: Micro variant annotation output file from SnpEff. A total of 2 719 213 putative effects were assigned to the identified micro variants; File S6: BLASTp analysis output of the core proteins, excluding hits on *F. circinatum*, 13 182 proteins had top hits to *Fusarium* species, 48 proteins had top hits to non-Fusarium species and 52 predicted proteins had no blast hits to the NCBI database; File S7: Accessory genomic elements identified in *F. circinatum*.

Author Contributions: Conceptualization, A.K., E.S., M.M. and B.; methodology, M.M., L.D.V. and A.K.; software, M.M., S.V.W., Q.S. and A.K.; validation, M.M. and A.K.; formal analysis, M.M..; investigation, M.M.; resources, B.W.; data curation, L.D.V and M.M.; writing—original draft preparation, M.M.; writing—review and editing, A.K., L.D.V., T.D., Q.C., S.V.W., M.M. and B.W.; visualization, M.M. and A.K.; supervision, E.S., A.K., L.D.V. and B.W.; project administration, M.M., E.S. and B.W.; funding acquisition, B.W. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the South African Department of Science and Innovation's South African Research Chair Initiative (SARChI grant number 98353), through the research chair of Prof. BD Wingfield. We would like to acknowledge the DSI-NRF Centre of Excellence in Plant Health Biotechnology (CPHB grant number 40945], at the Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The Whole Genome Shotgun project for *Fusarium circinatum* CMWF1803 has been deposited at DDBJ/ENA/GenBank under the accession JAEHFH000000000. The version described in this paper is version JAEHFH010000000. The Whole Genome Shotgun project for *Fusarium circinatum* CMWF560 has been deposited at DDBJ/ENA/GenBank under the accession JAEHFI000000000. The version described in this paper is version JAEHFI010000000. The Whole Genome Shotgun project for *Fusarium circinatum* CMWF567 has been deposited at DDBJ/ENA/GenBank under the accession JADZLS000000000. The version described in this paper is version JADZLS010000000. The Whole Genome Shotgun project for *Fusarium circinatum* UG27 has been deposited at DDBJ/ENA/GenBank under the accession JAELVK000000000. The version described in this paper is version JAELVK010000000. The Whole Genome Shotgun project for *Fusarium circinatum* UG10 has been deposited at DDBJ/ENA/GenBank under the accession JAGJRQ000000000. The version described in this paper is version JAGJRQ010000000.

Acknowledgments: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript, or in the decision to publish the results.

References

1. Schirawski, J.; Mannhaupt, G.; Münch, K.; Brefort, T.; Schipper, K.; Doehlemann, G.; Di Stasio, M.; Rössel, N.; Mendoza-Mendoza, A.; Pester, D. Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* **2010**, *330*, 1546-1548.
2. Ma, L.-J.; Van Der Does, H.C.; Borkovich, K.A.; Coleman, J.J.; Daboussi, M.-J.; Di Pietro, A.; Dufresne, M.; Freitag, M.; Grabherr, M.; Henrissat, B. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* **2010**, *464*, 367.
3. Schmidt, S.M.; Panstruga, R. Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis? *Current Opinion in Plant Biology* **2011**, *14*, 392-399.
4. Croll, D.; McDonald, B.A. The accessory genome as a cradle for adaptive evolution in pathogens. *PLoS Pathogens* **2012**, *8*, e1002608.
5. Gardiner, D.M.; McDonald, M.C.; Covarelli, L.; Solomon, P.S.; Rusu, A.G.; Marshall, M.; Kazan, K.; Chakraborty, S.; McDonald, B.A.; Manners, J.M. Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathogens* **2012**, *8*, e1002952.
6. Möller, M.; Stukenbrock, E.H. Evolution and genome architecture in fungal plant pathogens. *Nature Reviews Microbiology* **2017**, *15*, 756.
7. De Wit, P.J.; Mehrabi, R.; Van den Burg, H.A.; Stergiopoulos, I. Fungal effector proteins: past, present and future. *Molecular Plant Pathology* **2009**, *10*, 735-747.
8. Perez-Nadales, E.; Nogueira, M.F.A.; Baldin, C.; Castanheira, S.; El Ghalid, M.; Grund, E.; Lengeler, K.; Marchegiani, E.; Mehrotra, P.V.; Moretti, M. Fungal model systems and the elucidation of pathogenicity determinants. *Fungal Genetics and Biology* **2014**, *70*, 42-67.

9. Parfrey, L.W.; Lahr, D.J.; Katz, L.A. The dynamic nature of eukaryotic genomes. *Molecular Biology and Evolution* **2008**, *25*, 787-794.
10. McCarthy, C.G.; Fitzpatrick, D.A. Pan-genome analyses of model fungal species. *Microbial Genomics* **2019**, *5*.
11. Eid, J.; Fehr, A.; Gray, J.; Luong, K.; Lyle, J.; Otto, G.; Peluso, P.; Rank, D.; Baybayan, P.; Bettman, B. Real-time DNA sequencing from single polymerase molecules. *Science* **2009**, *323*, 133-138.
12. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Nattestad, M.; von Haeseler, A.; Schatz, M.C. Accurate detection of complex structural variations using single-molecule sequencing. *Nature Methods* **2018**, *15*, 461-468.
13. Quail, M.A.; Smith, M.; Coupland, P.; Otto, T.D.; Harris, S.R.; Connor, T.R.; Bertoni, A.; Swerdlow, H.P.; Gu, Y. A tale of three next generation sequencing platforms: comparison of Ion Torrent, Pacific Biosciences and Illumina MiSeq sequencers. *BMC Genomics* **2012**, *13*, 341.
14. Wikler, K.; Gordon, T.R. An initial assessment of genetic relationships among populations of *Fusarium circinatum* in different parts of the world. *Canadian Journal of Botany* **2000**, *78*, 709-717.
15. Fru, F.F.; Steenkamp, E.T.; Wingfield, M.J.; Roux, J. High genetic diversity of *Fusarium circinatum* associated with the first outbreak of pitch canker on *Pinus patula* in South Africa. *Southern Forests: a Journal of Forest Science* **2019**, *81*, 69-78.
16. Wikler, K.; Gordon, T.R.; Clark, S.L.; Wingfield, M.J.; Britz, H. Potential for outcrossing in an apparently asexual population of *Fusarium circinatum*, the causal agent of pitch canker disease. *Mycologia* **2000**, *92*, 1085-1090.
17. Santana, Q.C.; Coetzee, M.P.A.; Wingfield, B.D.; Wingfield, M.J.; Steenkamp, E.T. Nursery-linked plantation outbreaks and evidence for multiple introductions of the pitch canker pathogen *Fusarium circinatum* into South Africa. *Plant Pathology* **2016**, *65*, 357-368.
18. Pérez-Sierra, A.; Landeras, E.; León, M.; Berbegal, M.; García-Jiménez, J.; Armengol, J. Characterization of *Fusarium circinatum* from *Pinus spp.* in northern Spain. *Mycological Research* **2007**, *111*, 832-839.
19. Wöstemeyer, J.; Kreibich, A. Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. *Current Genetics* **2002**, *41*, 189-198.
20. Kempken, F.; Kück, U. Transposons in filamentous fungi—facts and perspectives. *BioEssays* **1998**, *20*, 652-659.
21. Soucy, S.M.; Huang, J.; Gogarten, J.P. Horizontal gene transfer: building the web of life. *Nature Reviews Genetics* **2015**, *16*, 472-482.
22. Gladyshev, E. Repeat-induced point mutation and other genome defense mechanisms in fungi. *Microbiology Spectrum* **2017**, *5*, 5-4.
23. Amselem, J.; Lebrun, M.-H.; Quesneville, H. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. *BMC Genomics* **2015**, *16*, 1-14.
24. Van Wyk, S.; Wingfield, B.D.; De Vos, L.; van der Merwe, N.A.; Santana, Q.C.; Steenkamp, E.T. Repeat-induced point mutations drive divergence between *Fusarium circinatum* and its close relatives. *Pathogens* **2019**, *8*, 298.
25. Steenkamp, E.T.; Wingfield, M.J.; McTaggart, A.R.; Wingfield, B.D. Fungal species and their boundaries matter–Definitions, mechanisms and practical implications. *Fungal Biology Reviews* **2018**, *32*, 104-116.
26. Van der Nest, M.A.; Beirn, L.A.; Crouch, J.A.; Demers, J.E.; De Beer, Z.W.; De Vos, L.; Gordon, T.R.; Moncalvo, J.-M.; Naidoo, K.; Sanchez-Ramirez, S. Draft genomes of *Amanita jacksonii*, *Ceratocystis albifundus, Fusarium circinatum, Huntiella omanensis, Leptographium procerum, Rutstroemia sydowiana*, and *Sclerotinia echinophila*. *IMA Fungus* **2014**, *5*, 472-485.
27. Slinski, S.; Kirkpatrick, S.; Gordon, T. Inheritance of virulence in *Fusarium circinatum*, the cause of pitch canker in pines. *Plant Pathology* **2016**, *65*, 1292-1296.
28. Jeffares, D.C.; Jolly, C.; Hoti, M.; Speed, D.; Shaw, L.; Rallis, C.; Balloux, F.; Dessimoz, C.; Bähler, J.; Sedlazeck, F.J. Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nature Communications* **2017**, *8*, 14061.

29. Steenwyk, J.; Rokas, A. Extensive copy number variation in fermentation-related genes among *Saccharomyces cerevisiae* wine strains. *G3: Genes, Genomes, Genetics* **2017**, *7*, 1475-1485.
30. Syvänen, A.-C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Reviews Genetics* **2001**, *2*, 930-942.
31. Suárez-Vega, A.; Gutiérrez-Gil, B.; Klopp, C.; Tosser-Klopp, G.; Arranz, J.J. Variant discovery in the sheep milk transcriptome using RNA sequencing. *BMC Genomics* **2017**, *18*, 170.
32. Das, A.; Panitz, F.; Gregersen, V.R.; Bendixen, C.; Holm, L.-E. Deep sequencing of Danish Holstein dairy cattle for variant detection and insight into potential loss-of-function variants in protein coding genes. *BMC Genomics* **2015**, *16*, 1043.
33. Daboussi, M.-J.; Capy, P. Transposable elements in filamentous fungi. *Annual Reviews in Microbiology* **2003**, *57*, 275-299.
34. Katju, V.; Bergthorsson, U. Copy-number changes in evolution: rates, fitness effects and adaptive significance. *Frontiers in Genetics* **2013**, *4*, 273.
35. Chuma, I.; Isobe, C.; Hotta, Y.; Ibaragi, K.; Futamata, N.; Kusaba, M.; Yoshida, K.; Terauchi, R.; Fujita, Y.; Nakayashiki, H. Multiple translocation of the *AVR-Pita* effector gene among chromosomes of the rice blast fungus *Magnaporthe oryzae* and related species. *PLoS Pathogens* **2011**, *7*, e1002147.
36. Kirkpatrick, M. How and why chromosome inversions evolve. *PLoS Biology* **2010**, *8*, e1000501.
37. Kirkpatrick, M.; Barton, N. Chromosome inversions, local adaptation and speciation. *Genetics* **2006**, *173*, 419-434.
38. Hu, G.; Wang, J.; Choi, J.; Jung, W.H.; Liu, I.; Litvintseva, A.P.; Bicanic, T.; Aurora, R.; Mitchell, T.G.; Perfect, J.R. Variation in chromosome copy number influences the virulence of *Cryptococcus neoformans* and occurs in isolates from AIDS patients. *BMC Genomics* **2011**, *12*, 526.
39. Steenwyk, J.L.; Soghigian, J.S.; Perfect, J.R.; Gibbons, J.G. Copy number variation contributes to cryptic genetic variation in outbreak lineages of *Cryptococcus gattii* from the North American Pacific Northwest. *BMC Genomics* **2016**, *17*, 700.
40. De Vos, L.; Myburg, A.A.; Wingfield, M.J.; Desjardins, A.E.; Gordon, T.; Wingfield, B.D. Complete genetic linkage maps from an interspecific cross between *Fusarium circinatum* and *Fusarium subglutinans*. *Fungal Genetics and Biology* **2007**, *44*, 701-714.
41. Wikler, K.; Gordon, T.R. An initial assessment of genetic relationships among populations of *Fusarium circinatum* in different parts of the world. *Canadian Journal of Botany* **2000**, *78*, 709-717.
42. Steenkamp, E.; Makhari, O.; Coutinho, T.; Wingfield, B.; Wingfield, M. Evidence for a new introduction of the pitch canker fungus *F usarium circinatum* in S outh Africa. *Plant Pathology* **2014**, *63*, 530-538.
43. Murray, M.; Thompson, W.F. Rapid isolation of high molecular weight plant DNA. *Nucleic Acids Research* **1980**, *8*, 4321-4326.
44. Koren, S.; Walenz, B.P.; Berlin, K.; Miller, J.R.; Bergman, N.H.; Phillippy, A.M. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* **2017**, *27*, 722-736.
45. Bolger, A.M.; Lohse, M.; Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **2014**, *30*, 2114-2120.
46. Chin, C.-S.; Alexander, D.H.; Marks, P.; Klammer, A.A.; Drake, J.; Heiner, C.; Clum, A.; Copeland, A.; Huddleston, J.; Eichler, E.E. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nature Methods* **2013**, *10*, 563.
47. Loman, N.J.; Quick, J.; Simpson, J.T. A complete bacterial genome assembled *de novo* using only nanopore sequencing data. *Nature Methods* **2015**, *12*, 733.
48. Walker, B.J.; Abeel, T.; Shea, T.; Priest, M.; Abouelliel, A.; Sakthikumar, S.; Cuomo, C.A.; Zeng, Q.; Wortman, J.; Young, S.K. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PloS One* **2014**, *9*, e112963.

49. De Vos, L.; Steenkamp, E.T.; Martin, S.H.; Santana, Q.C.; Fourie, G.; van der Merwe, N.A.; Wingfield, M.J.; Wingfield, B.D. Genome-wide macrosynteny among *Fusarium* species in the *Gibberella fujikuroi* complex revealed by amplified fragment length polymorphisms. *PLoS One* **2014**, *9*, e114682.
50. Harris, R. Improved pairwise alignment of genomic DNA Ph. D. *The Pennsylvania State University. United States—Pennsylvania* **2007**.
51. Kearse, M.; Moir, R.; Wilson, A.; Stones-Havas, S.; Cheung, M.; Sturrock, S.; Buxton, S.; Cooper, A.; Markowitz, S.; Duran, C. Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics* **2012**, *28*, 1647-1649.
52. Wiemann, P.; Sieber, C.M.; Von Bargen, K.W.; Studt, L.; Niehaus, E.-M.; Espino, J.J.; Huß, K.; Michielse, C.B.; Albermann, S.; Wagner, D. Deciphering the cryptic genome: genome-wide analyses of the rice pathogen *Fusarium fujikuroi* reveal complex regulation of secondary metabolism and novel metabolites. *PLoS Pathogens* **2013**, *9*, e1003475.
53. Wingfield, B.D.; Barnes, I.; de Beer, Z.W.; De Vos, L.; Duong, T.A.; Kanzi, A.M.; Naidoo, K.; Nguyen, H.D.; Santana, Q.C.; Sayari, M. Draft genome sequences of *Ceratocystis eucalypticola*, *Chrysoporthe cubensis*, *C. deuterocubensis*, *Davidsoniella virescens*, *Fusarium temperatum*, *Graphilbum fragrans*, *Penicillium nordicum*, and *Thielaviopsis musarum*. *IMA Fungus* **2015**, *6*, 493-506.
54. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv preprint arXiv* **2013,** *1303*, 3997.
55. Sedlazeck, F.J.; Rescheneder, P.; Smolka, M.; Fang, H.; Natestad, M.; von Haeseler, A. Accurate detection of complex structural variations using single molecule sequencing. *Nature Methods* **2018**, *15*, 461-468.
56. Li, H.; Handsaker, B.; Wysoker, A.; Fennell, T.; Ruan, J.; Homer, N.; Marth, G.; Abecasis, G.; Durbin, R.; Subgroup, G.P.D.P. The sequence alignment/map format and SAMtools. *Bioinformatics* **2009**, *25*, 2078-2079.
57. Barnett, D.W.; Garrison, E.K.; Quinlan, A.R.; Strömberg, M.P.; Marth, G.T. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **2011**, *27*, 1691-1692.
58. Simão, F.A.; Waterhouse, R.M.; Ioannidis, P.; Kriventseva, E.V.; Zdobnov, E.M. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **2015**, *31*, 3210-3212.
59. Hoff, K.J.; Stanke, M. WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes. *Nucleic Acids Research* **2013**, *41*, W123-W128.
60. Wingfield, B.D.; Steenkamp, E.T.; Santana, Q.C.; Coetzee, M.; Bam, S.; Barnes, I.; Beukes, C.W.; Yin Chan, W.; De Vos, L.; Fourie, G. First fungal genome sequence from Africa: a preliminary analysis. *South African Journal of Science* **2012**, *108*, 01-09.
61. Wingfield, B.D.; Liu, M.; Nguyen, H.D.; Lane, F.A.; Morgan, S.W.; De Vos, L.; Wilken, P.M.; Duong, T.A.; Aylward, J.; Coetzee, M.P. Nine draft genome sequences of *Claviceps purpurea s. lat.*, including *C. arundinis*, *C. humidiphila*, and *C. cf. spartinae*, pseudomolecules for the pitch canker pathogen *Fusarium circinatum*, draft genome of *Davidsoniella eucalypti*, *Grosmannia galeiformis*, *Quambalaria eucalypti*, and *Teratosphaeria destructans*. *IMA Fungus* **2018**, *9*, 401.
62. Robinson, J.T.; Thorvaldsdóttir, H.; Winckler, W.; Guttman, M.; Lander, E.S.; Getz, G.; Mesirov, J.P. Integrative genomics viewer. *Nature Biotechnology* **2011**, *29*, 24.
63. Cingolani, P.; Platts, A.; Wang, L.L.; Coon, M.; Nguyen, T.; Wang, L.; Land, S.J.; Lu, X.; Ruden, D.M. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w1118; iso-2; iso-3. *Fly* **2012**, *6*, 80-92.
64. Ruden, D.M.; Cingolani, P.; Patel, V.M.; Coon, M.; Nguyen, T.; Land, S.J.; Lu, X. Using *Drosophila melanogaster* as a model for genotoxic chemical mutational studies with a new program, SnpSift. *Frontiers in Genetics* **2012**, *3*, 35.

65. Conesa, A.; Götz, S.; García-Gómez, J.M.; Terol, J.; Talón, M.; Robles, M. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* **2005**, *21*, 3674-3676.
66. Bankevich, A.; Nurk, S.; Antipov, D.; Gurevich, A.A.; Dvorkin, M.; Kulikov, A.S.; Lesin, V.M.; Nikolenko, S.I.; Pham, S.; Prjibelski, A.D. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology* **2012**, *19*, 455-477.
67. Garrison, E.; Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv* **2012,** *1207*, 3907.
68. Drillon, G.; Carbone, A.; Fischer, G. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. *PLoS One* **2014**, *9*, e92621.
69. Ozer, E.A.; Allen, J.P.; Hauser, A.R. Characterization of the core and accessory genomes of *Pseudomonas aeruginosa* using bioinformatic tools Spine and AGEnt. *BMC Genomics* **2014**, *15*, 737.
70. Delcher, A.L.; Phillippy, A.; Carlton, J.; Salzberg, S.L. Fast algorithms for large-scale genome alignment and comparison. *Nucleic Acids Research* **2002**, *30*, 2478-2483.
71. Kurtz, S.; Phillippy, A.; Delcher, A.L.; Smoot, M.; Shumway, M.; Antonescu, C.; Salzberg, S.L. Versatile and open software for comparing large genomes. *Genome Biology* **2004**, *5*, R12.
72. Ozer, E.A. ClustAGE: a tool for clustering and distribution analysis of bacterial accessory genomic elements. *BMC Bioinformatics* **2018**, *19*, 150.
73. Jankowsky, E. RNA helicases at work: binding and rearranging. *Trends in Biochemical Sciences* **2011**, *36*, 19-29.
74. Strandberg, A.K.; Salter, L.A. A comparison of methods for estimating the transition: transversion ratio from DNA sequences. *Molecular Phylogenetics and Evolution* **2004**, *32*, 495-503.
75. Yin, L.; Zhang, H.; Tang, Z.; Xu, J.; Yin, D.; Zhang, Z.; Yuan, X.; Zhu, M.; Zhao, S.; Li, X. rMVP: A Memory-efficient, Visualization-enhanced, and Parallel-accelerated tool for Genome-Wide Association Study. *Genomics, Proteomics & Bioinformatics* **2020,** *19*, 619-628.
76. Emms, D.M.; Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **2019**, *20*, 1-14.
77. Walkowiak, S.; Rowland, O.; Rodrigue, N.; Subramaniam, R. Whole genome sequencing and comparative genomics of closely related Fusarium Head Blight fungi: *Fusarium graminearum*, *F. meridionale* and *F. asiaticum*. *BMC Genomics* **2016**, *17*, 1014.
78. Vesth, T.C.; Nybo, J.L.; Theobald, S.; Frisvad, J.C.; Larsen, T.O.; Nielsen, K.F.; Hoof, J.B.; Brandl, J.; Salamov, A.; Riley, R. Investigation of inter-and intraspecies variation through genome sequencing of *Aspergillus* section *Nigri*. *Nature Genetics* **2018**, *50*, 1688-1695.
79. Abd-Elsalam, K.; Schnieder, F.; Asran-Amal, A.; Khalil, M.; Verreet, J. Intra-species genomic groups in *Fusarium semitectum* and their correlation with origin and cultural characteristics. *Journal of Plant Diseases and Protection* **2003**, *110*, 409-418.
80. Ellison, C.E.; Hall, C.; Kowbel, D.; Welch, J.; Brem, R.B.; Glass, N.; Taylor, J.W. Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences USA* **2011**, *108*, 2831-2836.
81. Nielsen, R.; Paul, J.S.; Albrechtsen, A.; Song, Y.S. Genotype and SNP calling from next-generation sequencing data. *Nature Reviews Genetics* **2011**, *12*, 443-451.
82. Olson, N.D.; Lund, S.P.; Colman, R.E.; Foster, J.T.; Sahl, J.W.; Schupp, J.M.; Keim, P.; Morrow, J.B.; Salit, M.L.; Zook, J.M. Best practices for evaluating single nucleotide variant calling methods for microbial genomics. *Frontiers in Genetics* **2015**, *6*, 235.
83. Santana, Q.C.; Coetzee, M.P.; Steenkamp, E.T.; Mlonyeni, O.X.; Hammond, G.N.; Wingfield, M.J.; Wingfield, B.D. Microsatellite discovery by deep sequencing of enriched genomic libraries. *Biotechniques* **2009**, *46*, 217-223.

84. Maphosa, M.N.; Steenkamp, E.T.; Wingfield, B.D. Genome-based selection and characterization of *Fusarium circinatum*-specific sequences. *G3: Genes, Genomes, Genetics* **2016**, *6*, 631-639.
85. Phasha, M.M.; Wingfield, B.D.; Coetzee, M.P.; Santana, Q.C.; Fourie, G.; Steenkamp, E.T. Architecture and distribution of introns in core genes of four *Fusarium* species. *G3: Genes, Genomes, Genetics* **2017**, *7*, 3809-3820.
86. Wallace, M.M.; Covert, S.F. Molecular Mating Type Assay for *Fusarium circinatum*. *Applied Environmental Microbiology* **2000**, *66*, 5506-5508.
87. Xu, J.-R.; Yan, K.; Dickman, M.B.; Leslie, J.F. Electrophoretic karyotypes distinguish the biological species of *Gibberella fujikuroi* (*Fusarium* section *Liseola*). *Molecular Plant Microbe Interactions* **1995**, *8*, 74-84.
88. Waalwijk, C.; Taga, M.; Zheng, S.-L.; Proctor, R.H.; Vaughan, M.M.; O'Donnell, K. Karyotype evolution in *Fusarium*. *IMA Fungus* **2018**, *9*, 13-33.
89. Zakian, V.A. Telomeres: beginning to understand the end. *Science* **1995**, *270*, 1601-1607.
90. Cohn, M.; Liti, G.; Barton, D.B. Telomeres in fungi. In *Comparative Genomics*; Springer: 2006; pp. 101-130.
91. Barredo, J.L.; Díez, B.; Alvarez, E.; Martín, J.F. Large amplification of a 35-kb DNA fragment carrying two penicillin biosynthetic genes in high penicillin producing strains of *Penicillium chrysogenum*. *Current Genetics* **1989**, *16*, 453-459.
92. Fierro, F.; García-Estrada, C.; Castillo, N.I.; Rodríguez, R.; Velasco-Conde, T.; Martín, J.-F. Transcriptional and bioinformatic analysis of the 56.8 kb DNA region amplified in tandem repeats containing the penicillin gene cluster in *Penicillium chrysogenum*. *Fungal Genetics and Biology* **2006**, *43*, 618-629.
93. Smith, D.J.; Bull, J.H.; Edwards, J.; Turner, G. Amplification of the isopenicillin N synthetase gene in a strain of *Penicillium chrysogenum* producing high levels of penicillin. *Molecular and General Genetics* **1989**, *216*, 492-497.
94. Fogel, S.; Welch, J.W. Tandem gene amplification mediates copper resistance in yeast. *Proceedings of the National Academy of Sciences USA* **1982**, *79*, 5342-5346.
95. Marcotte, E.M.; Pellegrini, M.; Ng, H.-L.; Rice, D.W.; Yeates, T.O.; Eisenberg, D. Detecting protein function and protein-protein interactions from genome sequences. *Science* **1999**, *285*, 751-753.
96. Seoighe, C.; Federspiel, N.; Jones, T.; Hansen, N.; Bivolarovic, V.; Surzycki, R.; Tamse, R.; Komp, C.; Huizar, L.; Davis, R.W. Prevalence of small inversions in yeast gene order evolution. *Proceedings of the National Academy of Sciences USA* **2000**, *97*, 14433-14437.
97. Strauss, B.S. Frameshift mutation, microsatellites and mismatch repair. *Mutation Research/Reviews in Mutation Research* **1999**, *437*, 195-203.
98. Simmonds, J.; Scott, P.; Brinton, J.; Mestre, T.C.; Bush, M.; Del Blanco, A.; Dubcovsky, J.; Uauy, C. A splice acceptor site mutation in TaGW2-A1 increases thousand grain weight in tetraploid and hexaploid wheat through wider and longer grains. *Theoretical and Applied Genetics* **2016**, *129*, 1099-1112.
99. Fenn, J.; Boursnell, M.; Hitti, R.J.; Jenkins, C.A.; Terry, R.L.; Priestnall, S.L.; Kenny, P.J.; Mellersh, C.S.; Forman, O.P. Genome sequencing reveals a splice donor site mutation in the SNX14 gene associated with a novel cerebellar cortical degeneration in the Hungarian Vizsla dog breed. *BMC Genetics* **2016**, *17*, 1-8.
100. Golicz, A.A.; Bayer, P.E.; Bhalla, P.L.; Batley, J.; Edwards, D. Pangenomics comes of age: from bacteria to plant and animal applications. *Trends in Genetics* **2020**, *36*, 132-145.
101. Tautz, D.; Domazet-Lošo, T. The evolutionary origin of orphan genes. *Nature Reviews Genetics* **2011**, *12*, 692-702.