RESEARCH ARTICLE

Structural variants contribute to pangenome evolution of a plant pathogenic fungus

Li Guo^{1,2,#}, Quanbin Dong^{2,#}, Bo Wang¹, Mengyao Guo², Kai Ye^{1,2,†}

¹ MOE Key Laboratory for Intelligent Networks & Network Security, Faculty of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

² School of Life Science and Technology, Xi'an Jiaotong University, Xi'an 710049, China

†: Corresponding author:

kaiye@xjtu.edu.cn (K.Y.) ORCID: 0000-0002-2851-6741

#: Equal contribution

Running Title: *Fusarium graminearum* Structural Variation Landscape

Emails and ORCIDs:

Li Guo: guo_li@xjtu.edu.cn ORCID: 0000-0001-6100-3481

Quanbin Dong: yy3118113002@stu.xjtu.edu.cn ORCID: 0000-0002-0849-8136

Bo Wang: wangboxjtu@xjtu.edu.cn ORCID: 0000-0002-9041-878X

Mengyao Guo: myguo123@stu.xjtu.edu.cn ORCID: 0000-0002-7895-1290

Word count: 5566

Number of Figures: 4

Number of Tables: 2

Number of Supplementary Figures: 5

Number of Supplementary Tables: 4

ABSTRACT

Genetic variation is the driving force of plant-pathogen co-evolution. Large-scale genetic variations such as structural variations (SVs) often alter genome stability and organismal fitness. However, the pangenomic landscape and functional implications of SVs remain largely unexplored in plant pathogens. Here, we characterized the pangenomic and SV landscape in wheat head blight fungus *Fusarium graminearum* by producing and comparing chromosome-level (average contig N50 of 8.9 Mb) genome assemblies of 98 accessions using a reference-guided approach. Accounting for 29.05% and 19.01% of *F. graminearum* pangenome, respectively, accessory and private genomes are enriched with functions related to membrane trafficking, metabolism of fatty acids and tryptophans, with the private also enriched with putative effectors. Furthermore, using chromosome-level assemblies, we detected 52,420 SVs, 69.51% of which are inaccessible using read-mapping based approach. Over a half (55.65%) of 52,645 merged SVs affected 1,660 protein-coding genes, the most variable of which are involved in fungal virulence, cellular contact and communications. Interestingly, highly variable effectors and secondary metabolic enzymes are co-localized with SVs at subtelomeric and centromeric regions. Collectively, this landmark study shows the prevalence and functional relevance of SVs in *F. graminearum*, providing a valuable resource for future pangenomic studies in this cosmopolitan pathogen of cereal crops.

Keywords: *Fusarium graminearum*, head blight, genome evolution, population genomics, genome assembly, next-generation sequencing

61 INTRODUCTION

62 Fungal pathogens contribute to a substantial fraction of crop diseases and challenge
 63 global food safety, economic and social stability (Savary & Willocquet, 2020). For
 64 example, rice blast disease caused by *Magnaporthe oryzae* threatens rice productions
 65 worldwide (Dean et al., 2012). Fusarium head blight caused by *Fusarium*
 66 *graminearum* is a devastating disease of wheat and barley causing huge yield and
 67 economic losses (Goswami & Kistler, 2004). FHB also threatens human and animal
 68 health through mycotoxins such as trichothecenes and the estrogenic zearalenone
 69 (Chanda et al., 2016). A major obstacle of battling against many devastating crop
 70 diseases including FHB is the constant and rapid evolution of pathogen virulence and
 71 drug resistance through gene mutation and natural selection, an inevitable problem
 72 further deteriorated by fungicide abuses and resistant cultivar monoculture widely
 73 adopted in modern agriculture. Drug resistance in agricultural pathogens also poses
 74 dangers to human health through opportunistic fungal infections in
 75 immunocompromised individuals (Benitez & Carver, 2019). It is thus necessary to
 76 investigate the landscape and function of genetic mutations leading to evolution of
 77 fungal traits such as virulence and antifungal resistance, so that effective and
 78 environment-friendly strategies can be developed for plant disease prevention and
 79 management.

80 Genetic variants arisen from DNA mutations are the driving force behind evolution
 81 (Kronenberg *et al.*, 2018) including host-pathogen co-evolution with a boom-and-bust
 82 cycle (De la Concepcion *et al.*, 2018). Genetic variations come in various forms
 83 including single nucleotide polymorphisms (SNPs), small (<50bp) deletions or
 84 insertions (Indels) and structural variations (SVs) (>50bp) (Mahmoud *et al.*, 2019).
 85 Generally, genetic variants modify gene coding or non-coding sequences leading to
 86 altered gene functions and ultimately organismal fitness (Kronenberg *et al.*, 2018).
 87 Because these variants contribute to the formation of genetically diverse populations,
 88 any reference genome assembly of a single individual hardly represents the complete
 89 genetic information of any species known as pangenome (Parfrey *et al.*, 2008).

Pangenome represents a non-redundant complement of genome sequences for all individuals within a species (Tettelin *et al.*, 2005). First defined in bacteria, pangenome has been conceptually recognized and explored across all major kingdoms ranging from human (Li *et al.*, 2010), animals (Li *et al.*, 2017; Tian *et al.*, 2020), plants (Bayer *et al.*, 2020) to bacteria (Ding *et al.*, 2018) and fungi such as *Saccharomyces cerevisiae*, *Candida albicans*, *Cryptococcus neoformans*, *Aspergillus fumigatus* (Golicz *et al.*, 2016; Peter *et al.*, 2018), and also several plant pathogens including *Parastagonospora* spp, *Zymoseptoria tritici* (Plissonneau *et al.*, 2018; Syme *et al.*, 2018). Therefore, characterization of genetic variants is vital to mapping the pangenomes and understanding the mechanisms of species evolution.

Despite the importance of both small and large variants, our current understanding of fungal genetic variations generally focuses on SNPs that are widely used in population genetics and genome-wide association studies to link genotypes with phenotypes. So far, *F. graminearum* population genetic studies have emphasized on analysis of SNPs. For example, a link between local polymorphisms and pathogen specificity has been identified in *F. graminearum* genome (Cuomo *et al.*, 2007). Firasl *et al.* associated SNP diversity with genes crucial for *F. graminearum* phenotype including trichothecene chemotypes and virulence (Talas *et al.*, 2012). By contrast, there is to date a lack of both interest and effort in studying large variants such as indels and SVs in population genetic studies of fungal pathogens. However, compared to SNPs, SVs are more likely to disrupt the genome stability and function such as altering gene structure, copy numbers, and gene regulation given their large size (Alonge *et al.*, 2020). For example, SVs have already been implicated in development of various genetic disorders in certain human pedigrees or populations (Friedman *et al.*, 1994; Nattestad *et al.*, 2018). Therefore, a lack of population-wide mapping of structural variants in plant pathogens has led to an underestimation of their genetic diversities as well as impact on fungal pangenome evolution.

The overall lack of SV knowledge in plant pathogenic fungi is largely down to the technical challenges to detect SVs based on widely-used next-generation sequencing

(NGS) data due to its small read-length (Mahmoud *et al.*, 2019). Although variant detection tools such as *Pindel* (Ye *et al.*, 2009), *Delly* (Rausch *et al.*, 2012), *Lumpy* (Layer *et al.*, 2014) are available, application of these tools to NGS data are mostly ideal for detecting small variants with limited power in large variant discovery. Third-generation sequencing technology (*i.e.*, Pacific Bioscience or Oxford Nanopore), able to span most repetitive and complex regions in genome assembly and variant detection given the long reads (Mahmoud *et al.*, 2019), presents an ideal alternative to identify SVs. However, long-read sequencing remains expensive for variant detection in large-scale population genomic studies of plants and fungi. Recently, an alternative strategy has been proposed for variant detection based on a chromosome-scale reference genome and population-scale resequencing datasets. It involves reference-guided scaffolding of draft genome assemblies from NGS data, followed by assembly-based detection of variants. Several computational tools have been developed for this task including *Ragout2* (Kolmogorov *et al.*, 2014) and *RaGOO* (Alonge *et al.*, 2019) etc., providing a fast and affordable option to characterize pan-SV landscape at population level. The chromosome-scale genome assemblies also facilitate the analysis of pangenomes for the species being studied.

In this study, we sought to identify SVs in a large collection of *F. graminearum* accessions using chromosome-level genome assemblies, generated by referenced-guided genome scaffolding of NGS-based assembly, followed by SV identification. We also constructed the pangenome of *F. graminearum* based on these assemblies, revealing the contribution of accessory and private genomes to species adaptation. Intersecting the SVs with pangenome components highlighted the important role of SV in the genome evolution and pathogenesis of *F. graminearum*. This study not only presents a valuable resource for future population genomic and pangenomic investigation in this cosmopolitan fungal pathogen, but also demonstrates how SVs could be analyzed in fungal population genomic datasets solely based on NGS.

147

148 MATERIALS AND METHODS

149 Sequencing data and quality control

150 NGS (Illumina paired end) raw data of 104 *F. graminearum* isolates from five
 151 countries (China, USA, United Kingdom, France and Australia) around the globe
 152 were downloaded from National Center of Biological Information (NCBI) Sequence
 153 Read Archives (SRA) (Table S1). The SRA data were then converted to FASTQ
 154 format using SRA Toolkit (<https://github.com/ncbi/sra-tools>). The quality of the
 155 FASTQ data were assessed from two perspectives. Firstly, *FASTP* (Chen *et al.*, 2018)
 156 was used to check the read quality such as base quality, guanine-cytosine (GC)
 157 content, adapters etc. of the fastq files, followed by filtering reads with the poor
 158 quality and adapters with default parameters settings. Secondly, the software
 159 *Sourmash* (Ondov *et al.*, 2016) was used to check k-mer distributions of each dataset,
 160 finding and filtering out samples with abnormal k-mer frequencies. In total, 98 of 104
 161 samples passed the quality control and these cleaned data were used for the
 162 downstream analysis.

163 Chromosome-level genome assembly

164 *SPAdes* (Prjibelski *et al.*, 2020) was used to *de novo* assemble the cleaned reads, with
 165 the parameters: -k 33,55,77 --careful -t 28, and then the contigs.fasta and
 166 scaffolds.fasta were generated. *RaGOO* (Alonge *et al.*, 2019) was used to assemble
 167 contigs on the chromosome level based on the results of *SPAdes* (Prjibelski *et al.*,
 168 2020). The running parameter was -b -t 4-g 100-s-i 0.2, and the fasta file at the
 169 chromosome level was obtained. To evaluate the genome assemblies, we run *QUAST*
 170 (Gurevich *et al.*, 2013) with default parameters.

171 Genome annotations and effector prediction

172 For *F. graminearum* genome annotation, *de novo* gene structure was predicted by
 173 *GeneMark-ES* with parameters '--ES --fungus' (Lomsadze *et al.*, 2005;

174 Ter-Hovhannisyan *et al.*, 2008). A *Fusarium* gene model was then used to train
 175 *AUGUSTUS* v. 3.1 (Stanke *et al.*, 2008). *MAKER2* pipeline (Min *et al.*, 2017) with
 176 *RepeatMasker* v. 4.0.7 (Saha *et al.*, 2008) option on to find and mask repetitive
 177 elements, was used to find protein-coding genes integrating gene models predicted
 178 from *GeneMark-ES* and *AUGUSTUS*, and protein sequences of the *F. graminearum*.
 179 The *F. graminearum* putative effectors were predicted as follows: candidate secreted
 180 proteins have a secretion signal as determined by *EffectorP* (Sperschneider *et al.*,
 181 2018) and have no transmembrane domain as determined by *TMHMM* 2.0 (Krogh *et*
 182 *al.*, 2001). Eventually, *WoLF-PSort* v. 0.2 (Horton *et al.*, 2007) software was used to
 183 estimate the located sites and only those proteins that were credibly positioned in the
 184 extracellular space (i.e., extracellular score >15) were included into in the final
 185 secretome (Kaundal *et al.*, 2010). Small secreted proteins (SSPs) are defined here as
 186 proteins that are smaller than 200 amino acids and labeled as ‘cysteine rich’ when the
 187 percentage of cysteine residues in the protein was at least twice as high as the average
 188 percentage of cysteine residues in all predicted proteins of that organism.

189 **Variant detection**

190 Structural variant detection was conducted using two different approaches: mapping
 191 based approach (MBA) and assembly-based approach (ABA). For MBA, we first
 192 mapped NGS short reads to *F. graminearum* PH1 genome using *BWA-mem* (Li &
 193 Durbin, 2009), and performed structural variant detection using three mainstream SV
 194 callers *Lumpy* (Layer *et al.*, 2014), *Delly* (Rausch *et al.*, 2012) and *Manta* (Chen *et al.*,
 195 2016), followed by merging the detected SV of each caller (only considering SVs that
 196 are detected by at least two of four SV callers) using *SURVIVOR* (Jeffares *et al.*,
 197 2017). Alternatively, with ABA we aligned each of the 98 chromosome scale genome
 198 assemblies against *F. graminearum* PH1 genome, followed by structural variant
 199 detection using *Assemblytics* (Nattestad & Schatz, 2016). The chromosome-level
 200 genome assembly for each of 98 *F. graminearum* isolates was aligned to the reference
 201 genome PH1 using *minimap2* (Li, 2018) with the parameter settings: *minimap2 -k19*
 202 *-w19 reference.fasta contigs.fasta*, where "reference.fasta" and "contigs.fasta"

203 represents the PH1 reference genome and genome assembly results given by *RaGOO*,
 204 respectively. The alignments (.pav files) were then converted to delta format, and then
 205 used as input to *Assemblytics* for structural variant discovery with the parameter
 206 settings: *assemblytics contigs.delta contig_SV 1000000 1 1000000*. Structural variants
 207 detected recorded in .bed files as the output of *Assemblytics* were converted to VCF
 208 (variant call format) files using *SURVIVOR* v2.0.1. Structural variants of multiple
 209 isolates were filtered, compared and merged using *SURVIVOR* to identify common
 210 and distinct variants. SNP and indels were identified using Genome analysis tool kit
 211 (*GATK*) (DePristo *et al.*, 2011). *dN/dS* (the ratio of non-synonymous to synonymous
 212 substitutions) data were obtained from a previous report by Sperschneider *et al*
 213 (Sperschneider *et al.*, 2015).

214 **Structural variant effect analysis**

215 The effects of structural variants on genome functions were analyzed using
 216 *ANNOVAR* (Wang *et al.*, 2010). Genome annotation files (.gtf) and VCF files storing
 217 structural variant calls and genome coordinates were used as input to *ANNOVAR* for
 218 calculating the effects of each structural variant including overlaps with gene coding
 219 regions (introns and exons), UTRs, intergenic regions etc. The fungal genes affected
 220 by structural variants were obtained by overlapping the gene annotation information
 221 with the variant information stored in .bed files given by *Assemblytics* using *Bedtools*
 222 (Quinlan & Hall, 2010). A threshold of 10% or more in gene coding regions
 223 overlapping with any structural variant was used to identify genes affected by the
 224 variant.

225 **Pangenomic analysis**

226 The pangenome analysis was conducted using two different approaches:
 227 genome-based and gene-based approach. For genome-based approach, ppsPCP
 228 pipeline (Tahir Ul Qamar *et al.*, 2019) was used for pan-genome analysis to find full
 229 complement of genome sequences from all 98 genomes with default parameters. For

the gene-based approach, we used protein sequences of all 98 isolates and PH1 to identify ortholog groups (orthogroups) shared by all proteomes, among different proteomes and unique to each proteome using *OrthoFinder* (Emms & Kelly, 2019). Core genome is defined as orthogroups present in all isolates, whereas accessory genome is defined as orthogroups shared by some but not all isolates. Private genome is defined as orthogroups unique to each isolate. The three parts of pangenomes were compared with genes encoding for effectors, carbohydrate-degrading enzymes, virulence factors (PHI-base records (Urban *et al.*, 2020)) and trichothecene biosynthetic enzymes to evaluate the evolution of these gene functions in *F. graminearum*. The pangenome components were also compared with genes affected by structural variants to assess the contribution of the variants to these gene functions and fungal evolution. Pangenome openness was determined by fitting the pangenome profile curve model: $y = Ax^B + C$ (Tettelin *et al.*, 2005), where y and x represent pangenome size and genome number respectively, and A, B and C are fitting parameters.

Functional enrichment analysis

For gene function annotation, KEGG pathway analysis was performed using KOBAS3.0 (Xie *et al.*, 2011), protein domain was annotated by InterProScan (Jones *et al.*, 2014), and Gene Ontology was annotated by BLAST2GO (<https://www.blast2go.com/>), and then enrichment analysis was completed by TBtools (Chen *et al.*, 2020).

Data availability

The genome assemblies and variants reported in this paper have been deposited in the Genome Sequence Archive in National Genomics Data Center, China National Center for Bioinformation / Beijing Institute of Genomics, Chinese Academy of Sciences, under the BioProject ID PRJCA004286 and accession numbers WGS018715-WGS018812 that are publicly accessible at <https://bigd.big.ac.cn/gsa>.

257 RESULTS

258 Chromosome-level genome assembly of 98 *F. graminearum* accessions

259 To reconstruct the pangenome and identify structural variants in *F. graminearum*, we
 260 produced chromosome-level genome assemblies for a collection of *F. graminearum*
 261 accessions via a reference-guided approach. We first downloaded from public
 262 domains NGS data for 104 *F. graminearum* isolates originally sampled from five
 263 countries: China (CN), the United States (US), France (FR), United Kingdom (UK)
 264 and Australia (AUS) (Table S1). Quality of these NGS data were assessed, followed
 265 by removing six problematic datasets (showing abnormal Kmer frequencies) and
 266 poor-quality reads and sequence adapters, yielding a total of 98 high quality datasets
 267 including 60, 24, 6, 4 and 4 from US, AUS, FR, UK and CN, respectively (Figure 1A).
 268 Cleaned reads were then *de novo* assembled by *SPAdes* to generate 98 draft genome
 269 assemblies (Figure 1B; Table S1) with genome sizes ranged from 34.3Mb to 37.4Mb
 270 and an average GC content of 48.20%. Unsurprisingly, these assemblies were overall
 271 fragmented with the number of contigs ranging from 72 to 805 (Figure 1B and 1C),
 272 and contig N50 ranging from 93.8kb to 2.3Mb (Figure 1C).

273 High-quality genome assemblies are needed for optimal pangenome construction and
 274 efficient SV identification based on whole genome alignments. Recently, several
 275 algorithms such as *RaGOO* (Alonge *et al.*, 2019) for reference-guided genome
 276 assembly have been developed to scaffold contig-level assemblies into
 277 chromosome-level assemblies using a reference genome. From the NGS-based draft
 278 genomes of the 98 isolates, we further generated chromosome-scale genome assembly
 279 for each isolate using *F. graminearum* PH1 reference genome as a guide (Figure 1B).
 280 We obtained 98 final genome assemblies of high contiguity with contig N50 ranging
 281 from 8.3Mb to 10Mb (Figure 1C), a significant leap of quality over the draft
 282 assemblies (Table S1). We also observed that the draft contigs of each isolate could
 283 not be fully aligned into four chromosomes of PH1 genome, suggesting that each
 284 isolate has underwent substantial evolution carrying unique genome sequences. It

demonstrated the need to characterize the fungal pangenome because any individual genome is insufficient to represent the genetic information in the whole species.

Pangenomic analysis of *F. graminearum*

We recovered the *F. graminearum* pangenome sequence by a genome-based approach from the 98 chromosome-scale genome assemblies using ppsPCP pipeline (Tahir Ul Qamar *et al.*, 2019). First, each genome assembly was iteratively compared with the PH1 reference genome, followed by the presence-absence variation identification via scanning unique sequences (>100bp) of each accession relative to the reference genome. For each iteration, the unique sequences and the reference genome were merged into a non-redundant sequence file. The process was repeated for all 98 accessions to complete the pangenome construction for *F. graminearum*. The final pangenome size of the 98 accessions is 42.6Mb, about 5.6Mb larger than the PH1 reference genome. These extra sequences encoded a total of 1,203 protein-coding genes, and functional enrichment showed that they were mostly significantly enriched in pathways such as carbohydrate, fatty acid and tyrosine metabolism, transporters (Figure S1). Fatty acid, carbohydrate and amino acid metabolism produces primary metabolites that are not only essential for fungal cellular functions, but also precursors for fungal secondary metabolism (Chroumpi *et al.*, 2020).

For any species, pangenome typically consists of gene sets conserved in all, some or none of the isolates, which are defined as core, accessory and private genomes, respectively. To systematically identify the core, accessory and private genomes in *F. graminearum* pangenome, we first predicted the protein-coding genes from the chromosome-scale genomes of the 98 *F. graminearum* accessions using *AUGUSTUS* (Stanke *et al.*, 2008) based on *Fusarium*-specific gene model (Table S1). *Orthofinder* was then used to identify orthologs between PH1 and 98 samples, classifying genes into 15,408 orthogroups, among which 8,003 (51.94%) were present in all samples defined as the core genomes. Additionally, 2,928 (19.01%) orthogroups associated with a single accession, defined as private genomes. Finally, the remaining 4,476

(29.05%) orthogroups associated with at least two but not all accessions were defined as accessory genomes (Figure 2A and 2C). We found that the pangenome size increased before reaching a plateau as the number of accessions increased, but the size of core genomes decreased (Figure 2B), suggesting that *F. graminearum* has a closed pangenome. Interestingly, we found significant smaller dN/dS ratios were associated with the *F. graminearum* core genes than with the accessory genes and private genes, suggesting a different selection pressure likely being exerted on the three types of genomes (Figure 2D). Furthermore, functional enrichment showed that the accessory genes were enriched in membrane trafficking (SNARE mediated vesicle trafficking, exocytosis and autophagy), ribosome and protein translation. Private genes were enriched in transcription factors, metabolism of amino acids (valine, leucine and tryptophan) and fatty acids (Figure 2E), consistent with the finding using genome-based approach (Figure S1). By contrast, core genomes were enriched in pathways related to the basic metabolism and house-keeping cellular processes (Figure 2E). Collectively, the pangenomic analysis indicated that *F. graminearum* field populations have evolved accessory and private genomes with stronger diversifying selection compared to core genomes, reflecting the pangenome evolution behind the fungal adaptation.

331 Mapping structural variants in *F. graminearum*

332 Genetic variants play a central role in genome evolution. With the identified *F.*
333 *graminearum* pangenome, we are curious about what genomic variations each
334 accession went through to shape the current fungal genome. We characterized the
335 structural variations (SVs) in all 98 *F. graminearum* accessions, as SNPs and indels
336 have already been reported in these isolates previously by others (Cuomo *et al.*, 2007;
337 Talas *et al.*, 2012). More importantly, SVs are genetic variations typically larger than
338 50bp such as deletions, insertions, inversions, and translocations, and tend to have
339 more severe consequences to genome stability and organismal fitness (Medvedev *et al.*, 2009; Escaramís *et al.*, 2015). Here, we focused on detecting large deletions and
340 insertions, two most common SV types, in 98 *F. graminearum* isolates using two
341

different approaches: mapping based approach (MBA) and assembly-based approach (ABA). For MBA, we first mapped NGS short reads to *F. graminearum* PH1 genome using *BWA-mem* (Li & Durbin, 2009), and performed structural variant detection using three mainstream SV callers *Lumpy* (Layer *et al.*, 2014), *Delly* (Rausch *et al.*, 2012) and *Manta* (Chen *et al.*, 2016), followed by merging variants (only considering SVs that are detected by at least two of three SV callers) using *SURVIVOR* (Jeffares *et al.*, 2017). Alternatively, for ABA we aligned each of the 98 chromosome-scale genome assemblies against *F. graminearum* PH1 genome, followed by structural variant detection using *Assemblytics* (Nattestad & Schatz, 2016).

In total, the MBA method detected 10,253 SVs (> 50bp) including 10,118 deletions and 135 insertions from 98 *F. graminearum* isolates (Figure 3A). Conversely, the ABA method discovered a total of 52,420 SVs including 30,191 insertions (57.59%) and 22,229 deletions (42.41%) (Figure 3A). The fact that more SVs were detected by ABA than by MBA showed the power of chromosome-scale genome assemblies used for SV discovery. A comparison of SVs found that 8,855 SVs were captured by both MBA and ABA, occupying 86.36% and 16.89% of total SVs discovered by MBA and ABA, respectively (Figure 3A). Interestingly, 69.51% SVs (57.15% deletions, 99.55% insertions) detected by ABA were not detected by MBA. The size distribution showed that smaller and larger SVs are more detectable by MBA and ABA, respectively (Figure 3B). Harnessing the strength of both methods, we obtained a merged SV callset by incorporating variants identified by MBA and ABA, yielding a total of 52,645 SVs (Figure S2) for *F. graminearum*, including 23,614 deletions and 29,031 insertions which were used for downstream characterization of their population landscape and functional effects. Interestingly, SVs tend to be clustered at subteleomeric and centromeric regions of *F. graminearum* genome, although SVs were distributed throughout the genome (Figure 3C), consistent with previous reports that SVs occur more frequently in highly complex genomic regions (Sudmant *et al.*, 2015).

Biosynthesis of trichothecene mycotoxins is controlled by *Tri* gene cluster in *F.*

371 *graminearum* and other trichothecene-producing species (Gauthier *et al.*, 2015). Three
 372 trichothecene chemotypes have been found in natural isolates of *F. graminearum*:
 373 15-acetyl-deoxynivalenol (15ADON), 3-acetyl-deoxynivalenol (3ADON), and
 374 nivalenol (NIV). Studies have shown that gene presence and absence variation within
 375 the cluster leads to the fungal chemotypic diversity. In current study, we detected a
 376 large deletion event (2,379bp) contributing to the loss of *Tri7* gene in all 3ADON and
 377 NX2 chemotype, but not 15ADON chemotype of *F. graminearum* accession from
 378 USA (Figure S3). This is consistent with current knowledge that *Tri7*, a trichothecene
 379 biosynthesis gene encoding an acetyltransferase catalyzing a C-4 oxygenation essential
 380 for T2-toxin production in *F. sporotrichioides* (Brown *et al.*, 2001), is a pseudogene in
 381 *F. graminearum* 15ADON chemotypes and absent in 3ADON chemotypes (Rep &
 382 Kistler, 2010). However, this deletion event was not observed in *F. graminearum*
 383 accessions from China, France, Australia and England. In addition, we also detected a
 384 large segment of deletion (7,640bp) contributing to the loss of *Tri4*, *Tri5* and *Tri6*
 385 genes in 16 of 24 accessions from Australia, which are deletion mutants of the three
 386 genes generated using CRISPR-cas9 genome editing in *F. graminearum* (Table
 387 S2)(Gardiner & Kazan, 2018). The fact that these deletion events are consistent with
 388 the previous reports or prior knowledge indicates the reliability of the structural
 389 variant detection procedure in this study.

390 With the merged SVs, we further examined their population distributions and effects
 391 on coding genes. First, a principle component analysis using a SV presence/absence
 392 matrix revealed that the 98 isolates belonged to distinct clusters that overall
 393 correspond to their geographical regions (Figure 3D). Second, we found the UK
 394 isolates and US isolates had the lowest and highest number of SVs per sample,
 395 respectively (Figure S4), although this discrepancy of SV frequency could well be a
 396 result of insufficient sampling of the *F. graminearum* population of UK compared to
 397 US regions. Third, the genome-wide distribution of SVs showed that majority (84.4%)
 398 of SVs intersected with gene exonic regions and their upstream and downstream
 399 regulatory regions (Figure S5). These SVs affected a total of 1,660 protein-coding

genes *F. graminearum* enriched with pathways such as signal transduction and energy metabolism (Figure 3E), suggesting potential disruptive effects of SVs on the gene function and potential fitness. Lastly, the number of common SVs between isolates gradually decreased as the number of compared isolates increased. For instance, 1,660, 597, 145 and 8 protein-coding genes (1kb flanking each side) intersected by SVs were shared by at least 2, 10, 50 and 90 isolates (Table S3). We further identified highly variable genes among 98 accessions intersecting with the greatest number of SVs. The top 20 highly variable genes encode proteins involved in cell contact during mating (agglutinin like proteins), cell surface associated proteins (Mucins), myosins and kinesin proteins, virulence-associated proteins and 2OG-Fe oxygenase etc. (Table 1), suggesting these highly variable genes in *F. graminearum* pangenomes are likely associated with virulence, fungal cell communications and interactions with either other cells, or the environment.

Impact of SVs on *F. graminearum* pangenome and pathobiology gene functions

We next investigated how much SVs may have shaped *F. graminearum* pangenomes, by examining the fractions of genes affected by SVs associated with core, accessory and private genomes for each accession. Compared to the proportion of core (52%), accessory (29%) and private (19%) genes in pangenome (Figure 2A), 45%, 29% and 26% genes affected by SVs belong to core, accessory and private genomes, significantly overrepresented on private genomes but underrepresented on core genomes (Figure 4A; Table 2). This suggests a clear skewed contribution of SVs (large deletions and insertions) towards the evolution of private and accessory genomes, compared to core genomes in *F. graminearum*. As such, SVs would have caused extensive gene loss and gain in the fungal populations, leading to a diverse range of dispensable gene content in different accessions. Conversely, the under-representation of SV-affected genes in core genomes might be a consequence of purifying selection against disrupting conserved genes, many of which perform essential house-keeping functions.

Next, we examined how structural variants have affected specific groups of genes that are associated with pathogenesis or secondary metabolism of *F. graminearum*, including carbohydrate-active enzymes (CAZYme), effectors, secondary metabolic gene clusters and transcription factors. For each of these gene groups, we performed statistical test (Fisher's exact tests) (Table 2) to determine whether their distribution on each compartment (core, accessory and private) of pangenome significantly deviated from a random distribution of three pangenomic compartments, followed by testing whether such distribution also significantly deviated from the distribution of these gene groups intersecting with SVs on each compartment of pangenome (Table S4). For example, we predicted 584 effector proteins in *F. graminearum* pangenome, small secreted fungal proteins that typically promote pathogenesis, of which 32%, 23% and 45% located in core, accessory and private genome, respectively (Figure 4B), with private genome significantly enriched with effectors (Table 2). We found 65 effectors intersected with SVs, of which 26%, 32% and 42% belong to core, accessory and private genome, respectively (Figure 4B), without enrichment on any compartment (Table 2). Similarly, we analyzed SV impact on 29 transcription factors (TFs), a list of 764 *F. graminearum* CAZYme-encoding genes (Figure 4B) downloaded from dbCAN meta server (Zhang *et al.*, 2018), and 696 secondary metabolic genes (SMG) (Figure 4C) we predicted using antismash. The results show that no deviation of distribution was observed for SMGs, global or SV-affected, on any compartment (Table 2). Although TFs and CAZYmes are overall enriched on core genome, no significant enrichment of SV-affected TFs or CAZYmes was found on any compartment (Table 2). Despite such a lack of significant enrichment, an increased proportion on accessory compartment was found for SV-affected SMG (33.62%) and CAZYmes (30.10%) compared to pangenomic ratio (29.05%), suggesting that SVs have contributed to increased variability of these proteins among *F. graminearum* isolates.

Finally, we showed that SMG clusters and effectors harbor substantial structural variations among isolates across different countries (Figure 4E and 4F). We found 22 (33.85%) effectors and 40 (29.41%) SMGs are affected by a deletion or insertion in at

least ten isolates, respectively. These highly variable SMGs and effector genes are mostly located at subteleomeric and centromeric regions of chromosomes, consistent with the genomic distribution of SVs (Figure 3C, track g-h). Given likely associations of CAZymes, effectors and SMG clusters to fungal pathobiology, and the disruptive effects of structural variations on the coding and flanking sequence of these genes, our results indicate that the pathogen pangenome is likely experiencing rapid evolution in these genes allowing the fungus to adapt to host and environmental cues.

DISCUSSION

The landscape and functional roles of structural variants in fungal pathogens remain an overall uncharted area of research in plant pathogens. Focusing on *F. graminearum*, one of the most researched plant fungal pathogens, we for the first time performed systematic identification of large-scale genome structural variants in a collection of 98 fungal isolates with resequencing data. Knowledge-wise, our study have made new discoveries in three major aspects. Firstly, through reference-guided genome assembly and alignment followed by variant detection, we discover that structural variants are prevalent in *F. graminearum* field populations. Secondly, we show that many of these deletion and insertion variants co-localize with coding genes and thus may disrupt their normal functions. The most highly variable genes (found in over 80% of the *F. grmainearum* accessions analyzed in this study) caused by SVs are involved in agglutin proteins, mucins and kinesins that mediate cell to cell contact and communications during mating or interaction with environment. A high proportion of isolates carrying these mutations indicates pathogen adaptation to surrounding cells or environment is likely under strong selection. Thirdly, although these variants can be found throughout the genome, a high density of SVs is associated with genomic regions near centromere and telomeres. SVs in these highly polymorphic regions intersected with genes encoding putative effectors and secondary metabolic enzymes. Whether SVs play similar roles in evolution of other fungal pathogens of plants and humans would be intriguing to investigate.

Our study also showcased a computational strategy to characterize SVs of plant pathogenic fungi from large populations. The technical challenge of structural variation detection using short reads has been a major reason why these variants are left unnoticed in *F. graminearum*. In this study, we showed that the assembly-based method detected 44,569 structural variants that are inaccessible to traditional read-mapping method, highlighting the limitation of large variant detection based on short reads. Recently, variant callers are being developed to identify SVs in human samples based on single-molecule sequencing data (PacBio and Oxford Nanopore). Therefore, plant and fungal structural variant detections are bound to be improved using these long-read sequencing data given their advantage in detecting large and complex variants, although the cost of producing and analyzing these data from a massive plant or fungal populations remains a tremendous challenge for most large-scale population genomics studies so far. The approach (reference-guided assembly followed by SV detection) we adopted in this study enabled the SV analysis solely based on short reads, proving its efficacy working with population scale resequencing data in pathogenic fungi. With the cost of sequencing continuously plummeting in the near future, it will be possible to obtain long-read-based fungal resequencing data from hundreds or thousands of field isolates or experimental strains to reveal a more complete pangenomic and pan-SV landscape.

F. graminearum SVs detected in this work represent a valuable resource for future population genomic and pangenomic studies in this cereal pathogen, which is important for two reasons. First, the prevalence of large scale genome variants in *F. graminearum* genome clearly shows the inadequacy of a single reference genome in population genetic studies, since it tends to introduce geographic bias in interpreting the genomic data. A pangenomic database integrating all types of variants is essential to a more robust interpretation of genetic variations genotyped in various *F. graminearum* populations. Second, failure to characterize the full spectrum of genome variants by missing the structural variants represents a blind spot for discovering the genotype and phenotype associations in *F. graminearum*. Despite the effects of SNPs

514 in gene expression and regulation, they are less disruptive to gene functions and
515 phenotypes than large-scale variations such as SVs and chromosomal aberrations.
516 Therefore, it's critical to take into consideration the impacts of a broader spectrum of
517 variants for identifying the causal mutations behind trait evolution such as tolerance
518 of antifungal drugs or evasion of host resistance.

519 In conclusion, we have produced genome assemblies for a large collection of *F.*
520 *graminearum* isolates, based on which the fungal pangenome and structural variants
521 were comprehensively analyzed. Our study demonstrates that SVs are ubiquitous in *F.*
522 *graminearum* genomes disrupting functions of genes possibly associated with
523 pathogenesis and secondary metabolism, providing insights into the fungal genome
524 evolution. The computational strategies and structural variant resources developed by
525 this study will be valuable to future population genetic researches of *F. graminearum*
526 and other plant pathogenic fungi.

527 **AUTHOR CONTRIBUTIONS**

528 LG and KY conceived and designed the project. LG, QBD and MG performed the
529 quality control, variant detection and pangenome analysis. WB conducted the genome
530 assembly and annotation. LG, QBD and KY wrote the manuscript. All authors revised
531 and approved the manuscript.

532 **ACKNOWLEDGEMENT**

533 This project was supported by the National Natural Science Foundation of China
534 (31701739, 31970317 and 3167372) and National Key R&D Program of China
535 (2017YFC0907500, 2018YFC0910400), China Postdoctoral Foundation Grant
536 (2017□M623188) and the Fundamental Research Fund of Xi'an Jiaotong University
537 (1191329155). The authors also would like to thank anonymous reviewers for their
538 comments and suggestions to improve this manuscript.

539 **CONFLICT OF INTEREST**

540 The authors declare no conflict of interest.

541

542

543 REFERENCES

544

545 **Alonge M, Soyk S, Ramakrishnan S, Wang X, Goodwin S, Sedlazeck FJ,**
546 **Lippman ZB, Schatz MC. 2019.** RaGOO: fast and accurate reference-guided
547 scaffolding of draft genomes. *Genome biology* **20**(1): 224.

548 **Alonge M, Wang X, Benoit M, Soyk S, Pereira L, Zhang L, Suresh H,**
549 **Ramakrishnan S, Maumus F, Ciren D, et al. 2020.** Major Impacts of
550 Widespread Structural Variation on Gene Expression and Crop Improvement
551 in Tomato. *Cell* **182**(1).

552 **Bayer PE, Golicz AA, Scheben A, Batley J, Edwards D. 2020.** Author Correction:
553 Plant pan-genomes are the new reference. *Nature plants* **6**(11): 1389.

554 **Benitez LL, Carver PL. 2019.** Adverse Effects Associated with Long-Term
555 Administration of Azole Antifungal Agents. *Drugs* **79**(8): 833-853.

556 **Brown DW, McCormick SP, Alexander NJ, Proctor RH, Desjardins AE. 2001.** A
557 genetic and biochemical approach to study trichothecene diversity in
558 *Fusarium sporotrichioides* and *Fusarium graminearum*. *Fungal Genet Biol*
559 **32**(2): 121-133.

560 **Chanda A, Gummadidala PM, Goma OM. 2016.** Mycoremediation with
561 mycotoxin producers: a critical perspective. *Applied microbiology and*
562 *biotechnology* **100**(1): 17-29.

563 **Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. 2020.** TBtools:
564 An Integrative Toolkit Developed for Interactive Analyses of Big Biological
565 Data. *Molecular plant* **13**(8): 1194-1202.

566 **Chen S, Zhou Y, Chen Y, Gu J. 2018.** fastp: an ultra-fast all-in-one FASTQ
567 preprocessor. *Bioinformatics (Oxford, England)* **34**(17): i884-i890.

568 **Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox**
569 **AJ, Kruglyak S, Saunders CT. 2016.** Manta: rapid detection of structural
570 variants and indels for germline and cancer sequencing applications.
571 *Bioinformatics (Oxford, England)* **32**(8): 1220-1222.

572 **Chroumpi T, Mäkelä MR, de Vries RP. 2020.** Engineering of primary carbon
573 metabolism in filamentous fungi. *Biotechnology advances* **43**: 107551.

574 **Cuomo CA, Güldener U, Xu J-R, Trail F, Turgeon BG, Di Pietro A, Walton JD,**
575 **Ma L-J, Baker SE, Rep M, et al. 2007.** The *Fusarium graminearum* genome
576 reveals a link between localized polymorphism and pathogen specialization.
577 *Science (New York, N.Y.)* **317**(5843): 1400-1402.

578 **De la Concepcion JC, Franceschetti M, Maqbool A, Saitoh H, Terauchi R,**
579 **Kamoun S, Banfield MJ. 2018.** Polymorphic residues in rice NLRs expand
580 binding and response to effectors of the blast pathogen. *Nature plants* **4**(8):
581 576-585.

582 **Dean R, Van Kan JAL, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu**

583 **PD, Rudd JJ, Dickman M, Kahmann R, Ellis J, et al. 2012.** The Top 10
584 fungal pathogens in molecular plant pathology. *Molecular plant pathology*
585 **13**(4): 414-430.

586 **DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C,**
587 **Philippakis AA, del Angel G, Rivas MA, Hanna M, et al. 2011.** A
588 framework for variation discovery and genotyping using next-generation DNA
589 sequencing data. *Nature genetics* **43**(5): 491-498.

590 **Ding W, Baumdicker F, Neher RA. 2018.** panX: pan-genome analysis and
591 exploration. *Nucleic acids research* **46**(1): e5.

592 **Emms DM, Kelly S. 2019.** OrthoFinder: phylogenetic orthology inference for
593 comparative genomics. *Genome biology* **20**(1): 238.

594 **Escaramís G, Docampo E, Rabionet R. 2015.** A decade of structural variants:
595 description, history and methods to detect structural variation. *Briefings in*
596 *functional genomics* **14**(5): 305-314.

597 **Friedman LS, Ostermeyer EA, Szabo CI, Dowd P, Lynch ED, Rowell SE, King**
598 **MC. 1994.** Confirmation of BRCA1 by analysis of germline mutations linked
599 to breast and ovarian cancer in ten families. *Nature genetics* **8**(4): 399-404.

600 **Gardiner DM, Kazan K. 2018.** Selection is required for efficient Cas9-mediated
601 genome editing in *Fusarium graminearum*. *Fungal biology* **122**(2-3): 131-137.

602 **Gauthier L, Atanasova-Penichon V, Chéreau S, Richard-Forget F. 2015.**
603 Metabolomics to Decipher the Chemical Defense of Cereals against *Fusarium*
604 *graminearum* and Deoxynivalenol Accumulation. *International journal of*
605 *molecular sciences* **16**(10): 24839-24872.

606 **Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK,**
607 **Severn-Ellis A, McCombie WR, Parkin IAP, et al. 2016.** The pangenome of
608 an agronomically important crop plant *Brassica oleracea*. *Nature*
609 *communications* **7**: 13390.

610 **Goswami RS, Kistler HC. 2004.** Heading for disaster: *Fusarium graminearum* on
611 cereal crops. *Molecular plant pathology* **5**(6): 515-525.

612 **Gurevich A, Saveliev V, Vyahhi N, Tesler G. 2013.** QUAST: quality assessment tool
613 for genome assemblies. *Bioinformatics (Oxford, England)* **29**(8): 1072-1075.

614 **Horton P, Park K-J, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai**
615 **K. 2007.** WoLF PSORT: protein localization predictor. *Nucleic acids research*
616 **35**(Web Server issue): W585-W587.

617 **Jeffares DC, Jolly C, Hoti M, Speed D, Shaw L, Rallis C, Balloux F, Dessimoz C,**
618 **Bähler J, Sedlazeck FJ. 2017.** Transient structural variations have strong
619 effects on quantitative traits and reproductive isolation in fission yeast. *Nature*
620 *communications* **8**: 14061.

621 **Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, McWilliam H,**
622 **Maslen J, Mitchell A, Nuka G, et al. 2014.** InterProScan 5: genome-scale
623 protein function classification. *Bioinformatics (Oxford, England)* **30**(9):
624 1236-1240.

625 **Kaundal R, Saini R, Zhao PX. 2010.** Combining machine learning and
626 homology-based approaches to accurately predict subcellular localization in

627 *Arabidopsis*. *Plant physiology* **154**(1): 36-54.

628 **Kolmogorov M, Raney B, Paten B, Pham S. 2014.** Ragout-a reference-assisted
629 assembly tool for bacterial genomes. *Bioinformatics (Oxford, England)* **30**(12):
630 i302-i309.

631 **Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001.** Predicting
632 transmembrane protein topology with a hidden Markov model: application to
633 complete genomes. *Journal of molecular biology* **305**(3): 567-580.

634 **Kronenberg ZN, Fiddes IT, Gordon D, Murali S, Cantsilieris S, Meyerson OS,
635 Underwood JG, Nelson BJ, Chaisson MJP, Dougherty ML, et al. 2018.**
636 High-resolution comparative analysis of great ape genomes. *Science (New
637 York, N.Y.)* **360**(6393).

638 **Layer RM, Chiang C, Quinlan AR, Hall IM. 2014.** LUMPY: a probabilistic
639 framework for structural variant discovery. *Genome biology* **15**(6): R84.

640 **Li H. 2018.** Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics
641 (Oxford, England)* **34**(18): 3094-3100.

642 **Li H, Durbin R. 2009.** Fast and accurate short read alignment with Burrows-Wheeler
643 transform. *Bioinformatics (Oxford, England)* **25**(14): 1754-1760.

644 **Li M, Chen L, Tian S, Lin Y, Tang Q, Zhou X, Li D, Yeung CKL, Che T, Jin L, et
645 al. 2017.** Comprehensive variation discovery and recovery of missing
646 sequence in the pig genome using multiple de novo assemblies. *Genome
647 research* **27**(5): 865-874.

648 **Li R, Li Y, Zheng H, Luo R, Zhu H, Li Q, Qian W, Ren Y, Tian G, Li J, et al.
649 2010.** Building the sequence map of the human pan-genome. *Nature
650 biotechnology* **28**(1): 57-63.

651 **Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. 2005.** Gene
652 identification in novel eukaryotic genomes by self-training algorithm. *Nucleic
653 acids research* **33**(20): 6494-6506.

654 **Mahmoud M, Gobet N, Cruz-Davalos DI, Mounier N, Dessimoz C, Sedlazeck FJ.
655 2019.** Structural variant calling: the long and the short of it. *Genome biology*
656 **20**(1): 246.

657 **Medvedev P, Stanciu M, Brudno M. 2009.** Computational methods for discovering
658 structural variation with next-generation sequencing. *Nature methods* **6**(11
659 Suppl): S13-S20.

660 **Min B, Grigoriev IV, Choi I-G. 2017.** FunGAP: Fungal Genome Annotation
661 Pipeline using evidence-based gene model evaluation. *Bioinformatics (Oxford,
662 England)* **33**(18): 2936-2937.

663 **Nattestad M, Goodwin S, Ng K, Baslan T, Sedlazeck FJ, Rescheneder P, Garvin
664 T, Fang H, Gurtowski J, Hutton E, et al. 2018.** Complex rearrangements
665 and oncogene amplifications revealed by long-read DNA and RNA sequencing
666 of a breast cancer cell line. *Genome research* **28**(8): 1126-1135.

667 **Nattestad M, Schatz MC. 2016.** Assemblytics: a web analytics tool for the detection
668 of variants from an assembly. *Bioinformatics (Oxford, England)* **32**(19):
669 3021-3023.

670 **Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S,**

671 **Phillippy AM. 2016.** Mash: fast genome and metagenome distance estimation
672 using MinHash. *Genome biology* **17**(1): 132.

673 **Parfrey LW, Lahr DJG, Katz LA. 2008.** The dynamic nature of eukaryotic genomes.
674 *Molecular biology and evolution* **25**(4): 787-794.

675 **Peter J, De Chiara M, Friedrich A, Yue J-X, Pflieger D, Bergström A, Sigwalt A,**
676 **Barre B, Freel K, Llored A, et al. 2018.** Genome evolution across 1,011
677 *Saccharomyces cerevisiae* isolates. *Nature* **556**(7701): 339-344.

678 **Plissonneau C, Hartmann FE, Croll D. 2018.** Pangenome analyses of the wheat
679 pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic
680 eukaryotic genome. *BMC biology* **16**(1): 5.

681 **Prjibelski A, Antipov D, Meleshko D, Lapidus A, Korobeynikov A. 2020.** Using
682 SPAdes De Novo Assembler. *Current protocols in bioinformatics* **70**(1): e102.

683 **Quinlan AR, Hall IM. 2010.** BEDTools: a flexible suite of utilities for comparing
684 genomic features. *Bioinformatics (Oxford, England)* **26**(6): 841-842.

685 **Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. 2012.** DELLY:
686 structural variant discovery by integrated paired-end and split-read analysis.
687 *Bioinformatics (Oxford, England)* **28**(18): i333-i339.

688 **Rep M, Kistler HC. 2010.** The genomic organization of plant pathogenicity in
689 *Fusarium* species. *Current opinion in plant biology* **13**(4): 420-426.

690 **Saha S, Bridges S, Magbanua ZV, Peterson DG. 2008.** Empirical comparison of ab
691 initio repeat finding programs. *Nucleic acids research* **36**(7): 2284-2294.

692 **Savary S, Willocquet L. 2020.** Modeling the Impact of Crop Diseases on Global
693 Food Security. *Annual review of phytopathology* **58**: 313-341.

694 **Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. 2018.** Improved
695 prediction of fungal effector proteins from secretomes with EffectorP 2.0.
696 *Molecular plant pathology* **19**(9): 2094-2110.

697 **Sperschneider J, Gardiner DM, Thatcher LF, Lyons R, Singh KB, Manners JM,**
698 **Taylor JM. 2015.** Genome-Wide Analysis in Three *Fusarium* Pathogens
699 Identifies Rapidly Evolving Chromosomes and Genes Associated with
700 Pathogenicity. *Genome biology and evolution* **7**(6): 1613-1627.

701 **Stanke M, Diekhans M, Baertsch R, Haussler D. 2008.** Using native and
702 syntenically mapped cDNA alignments to improve de novo gene finding.
703 *Bioinformatics (Oxford, England)* **24**(5): 637-644.

704 **Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J,**
705 **Zhang Y, Ye K, Jun G, Fritz MH-Y, et al. 2015.** An integrated map of
706 structural variation in 2,504 human genomes. *Nature* **526**(7571): 75-81.

707 **Syme RA, Tan K-C, Rybak K, Friesen TL, McDonald BA, Oliver RP, Hane JK.**
708 **2018.** Pan-Parastagonospora Comparative Genome Analysis-Effector
709 Prediction and Genome Evolution. *Genome biology and evolution* **10**(9):
710 2443-2457.

711 **Tahir Ul Qamar M, Zhu X, Xing F, Chen L-L. 2019.** ppsPCP: a plant
712 presence/absence variants scanner and pan-genome construction pipeline.
713 *Bioinformatics (Oxford, England)* **35**(20): 4156-4158.

714 **Talas F, Würschum T, Reif JC, Parzies HK, Miedaner T. 2012.** Association of

single nucleotide polymorphic sites in candidate genes with aggressiveness and deoxynivalenol production in *Fusarium graminearum* causing wheat head blight. *BMC genetics* **13**: 14.

Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome research* **18**(12): 1979-1990.

Tettelin H, Masignani V, Cieslewicz MJ, Donati C, Medini D, Ward NL, Angiuoli SV, Crabtree J, Jones AL, Durkin AS, et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences of the United States of America* **102**(39): 13950-13955.

Tian X, Li R, Fu W, Li Y, Wang X, Li M, Du D, Tang Q, Cai Y, Long Y, et al. 2020. Building a sequence map of the pig pan-genome from multiple de novo assemblies and Hi-C data. *Science China. Life sciences* **63**(5): 750-763.

Urban M, Cuzick A, Seager J, Wood V, Rutherford K, Venkatesh SY, De Silva N, Martinez MC, Pedro H, Yates AD, et al. 2020. PHI-base: the pathogen-host interactions database. *Nucleic acids research* **48**(D1): D613-D620.

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**(16): e164.

Xie C, Mao X, Huang J, Ding Y, Wu J, Dong S, Kong L, Gao G, Li C-Y, Wei L. 2011. KOBAS 2.0: a web server for annotation and identification of enriched pathways and diseases. *Nucleic acids research* **39**(Web Server issue): W316-W322.

Ye K, Schulz MH, Long Q, Apweiler R, Ning Z. 2009. Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics (Oxford, England)* **25**(21): 2865-2871.

Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic acids research* **46**(W1).

755

756

757

758 TABLES

759 **Table 1.** Summary of the 20 most variable genes intersecting with structural variants

760 and their annotated functions in *Fusarium graminearum* pangenome. The number of

761 variants is the total number of structural variants intersecting with the protein-coding

762 sequene and its 1kb flanking region.

763

Gene ID	Annotations	Number of variants	Number of accessions with variants
FGRAMPH1_01G05643	agglutinin-Like Protein	568	93
FGRAMPH1_01G15613	agglutinin-Like Protein	489	94
FGRAMPH1_01G27087	agglutinin-Like Protein	232	85
FGRAMPH1_01G21813	myosin light chain kinase	187	89
FGRAMPH1_01G25011	mucin 1, cell surface associated (MUC1)	185	85
FGRAMPH1_01G15427	ankyrin-3 protein	184	93
FGRAMPH1_01G12267	mucin 22 protein	157	85
FGRAMPH1_01G25295	virulence-associated lipoprotein MIA	122	84
FGRAMPH1_01G13139	vacuolar carboxypeptidase	118	83
FGRAMPH1_01G22029	nucleoside phosphorylase	117	81
FGRAMPH1_01G08911	Extracellular serine/threonine-rich Protein	116	84
FGRAMPH1_01G08231	cell surface proteins containing the conserved peptide motif (LPXTG)	113	82
FGRAMPH1_01G12003	kinesin light chain	109	81
FGRAMPH1_01G28273	peptidase c14	108	69
FGRAMPH1_01G27923	2OG-Fe oxygenase	104	87
FGRAMPH1_01G11565	Unknown protein	103	73
FGRAMPH1_01G28289	kinesin light chain	102	96
FGRAMPH1_01G04545	zinc finger transcription factor	98	93
FGRAMPH1_01G10821	SNF5-component of SWI SNF transcription activator complex	97	49

764

765

766

767 **Table 2.** A summary of the core, accessory and private gene fractions in *Fusarium*
 768 *graminearum* pangenome (Global), SV-affected genes (Pan-SV), and genes belonging
 769 to four different functional groups (effectors, CAZyme, SMG and TF). Underneath
 770 the fractions are p-values given by two-tail Fisher's exact tests conducted to determine
 771 the statistical significance of gene enrichment. SV: structural variants. O:
 772 overrepresented. U: underrepresented. N: nonsignificant. NA: nonapplicable. SMG:
 773 secondary metabolic genes. TF: transcription factors.

774

Number of genes		Global	Effector	CAZYme	SMG	TF
Pangenome (15,407)	Core (8,003)	51.94%	32.36% $p = 2.3\text{e-}08$ (U)	68.98% $p = 1.2\text{e-}06$ (O)	56.03% $p = 0.2549$ (N)	71.01% $p = 0.0108$ (O)
	Accessory (4,476)	29.05%	22.60% $p = 0.01155$ (U)	30.10% $p = 0.6724$ (N)	33.62% $p = 0.0647$ (N)	28.99% $p = 0.9817$ (N)
	Private (2,928)	19.01%	45.03% $p < 2.2\text{e-}16$ (O)	0.92% $p < 2.2\text{e-}16$ (U)	10.34% $p = 1.1\text{e-}06$ (U)	0% NA
Pan-SV (1,660)	Core (842)	50.72% $p = 0.6084$ (N)	26.15% $p = 0.54$ (N)	61.90% $p = 0.6839$ (N)	53.68% $p = 0.8471$ (N)	65.52% $p = 0.9248$ (N)
	Accessory (659)	39.70% $p = 2.1\text{e-}10$ (O)	32.31% $p = 0.2344$ (N)	38.10% $p = 0.42$ (N)	37.50% $p = 0.6084$ (N)	34.48% $p = 0.8212$ (N)
	Private (159)	9.59% $p = 4.3\text{e-}16$ (U)	41.54% $p = 0.8282$ (N)	0% NA	8.82% $p = 0.7388$ (N)	0% NA

775

776

777

778

779

780

781 **Figure 1. Geographic distribution and genome assembly of 98 *Fusarium***
782 ***graminearum* accessions. A.** World map displaying the countries of origin for the *F.*
783 *graminearum* accessions included in this study. The color scale is proportional to the
784 number of accessions marked on the map. **B.** Whole genome alignments of *F.*
785 *graminearum* reference genome PH1 against the genome assembly using Illumina
786 short reads alone (left) and using *RaGOO* to perform a scaffolding based on the NGS
787 assembly (right), using UK2999 isolate as an example. **C.** Density distribution of
788 contig counts (left) and contig N50 (right) for the 98 genome assemblies using short

reads alone (blue) or using reference-guided assembly of the short reads (red).

790

Figure 2. Pan-genome analysis of *Fusarium graminearum*. **A.** Core, accessory and private genomes represent 51.94%, 29.05% and 19.01% of *F. graminearum* pan-genome, respectively. **B.** Variation of gene families in the pan-genome and core-genome along with an additional *F. graminearum* genome. **C.** The number of genes counted for each pan-genome composition (core, accessory and private) in 98 individual genomes. **D.** Boxplot of dN/dS ratio (nonsynonymous substitution rate divided by synonymous substitution rate) distribution for *F. graminearum* genes located on each pan-genome composition (core, accessory and private). The lower-case letter a, b and c represents the significant difference ($p < 0.05$) using Student's t-test. **E.** A bubble plot summarizing the functional enrichment analysis of each composition of *F. graminearum* pangenome. Y-axis and X-axis denotes the enriched KEGG terms and p value ($p < 0.05$).

803

804

805 **Figure 3. An overview of structural variant landscape in 98 *Fusarium***
806 ***graminearum* accessions. A.** Comparison of *F. graminearum* structural variants
807 detected using two different approaches: mapping-based approach (MBA) and
808 assembly-based approach (ABA). **B.** The size distribution of structural variants
809 showed that smaller and larger structural variants are more easily detectable by MBA
810 and ABA, respectively. **C.** Genome circos plot displaying the distributions of key
811 genomic features for *F. graminearum*. (a-h) GC content, Gene density, SNP density,
812 indel density, structural variant (SV-deletion, SV-insertion) density, effector and
813 secondary metabolic (SM) gene density calculated in 100-kb windows. Black bars (g
814 and h) represent the highly variable effectors and SM genes intersected with structural
815 variants among at least 80% of *F. graminearum* accessions. **D.** Principal components
816 analysis of the structural variants and geographical locations based on a
817 presence/absence matrix of the 98 accessions. **E.** Kyoto Encyclopedia of Genes and

818 Genomes (KEGG) pathway analyses of SV genes.

819

820 **Figure 4. Structural variations contribute to accessory genome evolution in *F.***
821 ***graminearum*.** A. Proportions of genes affected by structural variants (SV) across the
822 pangenome. B-D. Pan-SV categories of carbohydrate-active enzymes (B), effectors
823 (C) and secondary metabolite (SM) gene clusters (D). E-F. Heatmaps showing SV
824 frequency of effector (E) and secondary metabolic (F) genes.

825