Determination and Structural Analysis of the Whole-Genome Sequence of Fusarium equiseti D25-1

Xueping LI (lixueping@gsagr.ac.cn)
Gansu Academy of Agricultural Sciences https://orcid.org/0000-0001-5728-7768
Jianhong Li
Gansu Agricultural University
Yonghong Qi
Gansu Academy of Agricultural Sciences
Yonggang Liu
Gansu Academy of Agricultural Sciences
Minquan Li
Gansu Academy of Agricultural Sciences

Research article

Keywords: Fusarium equiseti, plant pathogen, Illumina HiSeq 4000, PacBio, whole genome

Posted Date: February 27th, 2020

DOI: https://doi.org/10.21203/rs.2.24663/v1

Abstract

Background

Fusarium equiseti is a plant pathogen with a wide range of hosts and diverse effects, including probiotic activity. However, the underlying molecular mechanisms remain unclear, hindering its effective control and utilization. In this study, the Illumina HiSeq 4000 and PacBio platforms were used to sequence and assemble the whole genome of *Fusarium equiseti* D25-1.

Results

The assembly included 16 fragments with a GC content of 48.01%, gap number of zero, and size of 40,776,005 bp. There were 40,110 exons and 26,281 introns having a total size of 19,787,286 bp and 2,290,434 bp, respectively. The genome had an average copy number of 333, 71, 69, 31, and 108 for tRNAs, rRNAs, sRNAs, snRNAs, and miRNAs, respectively. The total repetitive sequence length was 1,713,918 bp, accounting for 4.2033% of the genome. In total, 13,134 functional genes were annotated, accounting for 94.97% of the total gene number. Toxin-related genes, including two related to zearalenone and 23 related to trichothecene, were identified. A comparative genomic analysis supported the high quality of the *F. equiseti* assembly, exhibiting good collinearity with the reference strains, 3,483 species-specific genes, and 1,805 core genes. A gene family analysis revealed more than 2,500 single-copy orthologs. *F. equiseti* was most closely related to *Fusarium pseudograminearum* based on a phylogenetic analysis at the whole-genome level.

Conclusions

Our comprehensive analysis of the whole genome of *F. equiseti* provides basic data for studies of gene expression, regulatory and functional mechanisms, evolutionary processes, as well as disease prevention and control.

Background

Fusarium equiseti, also known as *Gibberella intricans* in its sexual stage, has complex and diverse beneficial or pathogenic properties. As a probiotic, it can produce many metabolites beneficial to plants or humans. For example, it produces linamarase [1], equisetin, cellulase, polyketides [2], glucose, alcohol [3], and various metabolites that inhibit hepatitis virus infections [4]. Additionally, Marín et al. (2012) have found that *F. equiseti* can produce certain type A and B trichothecenes, butenolides, beauvericin, zearalenone, and fusarochromanone [5]. *F. equiseti* is also employed in plant insect control [6–7], benzopyrene degradation, and heavy metal pollution abatement [8]. Conversely, *F. equiseti* itself also acts as a plant pathogen with a broad host range and produces various toxins. In fact, it is a pathogen of multiple cereals, including wheat [9], corn [10], triticale, oats, and barley [11], and other crops, such as cowpea [12], beans [13], melon [14], lettuce [15], sunflower [16], jujube, and radish [17]. Some flowers, trees, and Chinese medicinal plants, including hickory [18], white pine seedlings [19], cumin [20], fishtail palm [21], hybrid cymbidium, clover, rice flatsedge [22], white mangrove [23], mulberry [24], largehead atractylodes rhizome, and Chinese magnolia vine fruit [25] can also be infected with *F. equiseti*.

However, the mechanisms underlying the beneficial or pathogenic properties of *F. equiseti* are not fully understood. A direct and effective approach for clarifying the action mechanisms of *F. equiseti* is the identification of the genetic factors directly related to its functions. For instance, Stępień et al. have analyzed the *tef-1α* sequences of *F. equiseti*, as well as the toxin-related genes *PKS13*, *PKS4,* and *TRI5* [26], and Kari has successfully cloned a protease gene [27]. However, there are few in-depth studies on the functional genes. Moreover, the traditional methods for functional gene discovery are insufficient to meet the modern-day research demands due to the high error rates and low efficiencies.

Advances in whole-genome sequencing technologies have enabled numerous key discoveries. In this study, the complete genomic sequence of *Fusarium equiseti* D25-1 was determined using multiple sequencing platforms. The assembly was characterized, including analyses of the genomic structure, with the aim of providing data to support the discovery of beneficial or harmful genes and to improve our understanding of the molecular pathogenesis.

Results

Illumina HiSeq4000 data

Sequencing was performed using the Illumina HiSeq 4000 platform. A total of 56,458,678 reads were produced from D25-1, with an insert size of 500 bp and a read length of 125 bp. The proportion of adapter sequences was 0.25%, the proportion of repetitive sequences was

2.38%, and the low-quality sequences accounted for only 7.04%. After filtering out 9.69% of the sequences from the raw data (7,057 Mb), 6,372 Mb of clean data were obtained (Table 1).

Table 1
Summary statistics of D25-1 sequencing data

Insert size (bp)	Read length (bp)	Raw data (Mb)	Adapter (%)	Duplication (%)	Total reads	Filtered reads (%)	Low-quality filtered reads (%)	Clean data (Mb)
500	125	7,057	0.25	2.38	56,458,678	9.69	7.04	6,372

Table 2
Summary statistics of the data obtained by the PacBio platform

Polymerase Read Number	Mean Read Length (bp)	Polymerase Read Bases (bp)	Polymerase Read Quality	Sub-read Number	Sub-read Mean Length (bp)	Sub-read Bases (bp)	Subread Quality	Utilization Ratio
294,876	15,189	4,957,271,514	0.84	447,618	9,970	4,463,016,422	0.84	0.92

In the base distribution analysis, the frequencies of A and T, as well as the frequencies of G and C, were correlated, indicating an equilibrium base composition. Further, base quality values were relatively high, suggesting a low error rate and high-quality reads (Fig. 1).

PacBio data

Given the large quantities of adapter sequences and low-quality or erroneous sequences in the raw sequencing data obtained using the PacBio platform, extensive filtering was performed to generate clean data, which were deposited at DDBJ/ENA/GenBank under accession number QOHM00000000. The lengths and qualities of the reads were well distributed (Fig. 2.). The version described in this paper is version QOHM01000000. Summary statistics are provided in Table. 2.

Genome Evaluation

Prior to assembly, the genome size, heterozygosity, and repetitive sequence information were determined by a K-mer analysis. The D25-1 genome was approximately 44.69 Mb in size, with a coverage depth of 144.34×. The details are shown in Fig. 3.

Assembly statistics

Data assembly was performed by combining the data from the Illumina HiSeq4000 and PacBio sequencing platforms. As shown in Table 3, the D25-1 genome was well spliced. A total of 16 Fusarium equiseti chromosomes with a total length of 40,776,005 bases were assembled. The longest chromosome was 8,344,890 bp, and the shortest chromosome was 2,316 bp. The gap number was 0, and the GC content was 48.01%.

Table 3
Assembly statistics

Seq Type	Total Number	Total Length (bp)	N50 Length (bp)	N90 Length (bp)	Max Length (bp)	Min Length (bp)	Gap Number (bp)	GC Content (%)
Scaffold	16	40,776,005	6,178,397	2,783,306	8,344,890	2,316	0	48.01
Contig	16	40,776,005	6,178,397	2,783,306	8,344,890	2,316	-	48.01

GC content

Through calculating GC content and average depth, one can analyze whether a GC bias exists. If not seriously biased, the corresponding scatter diagram will assume a shape similar to that of Poisson distribution and have a peak near the GC content estimate of the genome. The more the graph deviates from this peak, the lower the depth is. The GC bias in the genome of D25-1 was analyzed after assembly, and a Poisson distribution was observed, suggesting no substantial bias (Fig. 4).

Gene components

Gene prediction was performed using the assembly to obtain the open reading frames and gene length distribution, as summarized in Table 4. The total number, total length, average length, and percentage of total chromosome lengths were analyzed for the genes, exons, CDS, and introns. The introns were abundant but short, accounting for a relatively low proportion of the genome.

Table 4
Gene statistics

Type	Total Number	Total Length (bp)	Average Length (bp)	Length/Genome Length (%)
Gene	13,829	22,077,720	1,596.48	54.14
Exon	40,110	19,787,286	493.33	48.53
CDS	13,829	19,787,286	1,430.85	48.53
Intron	26,281	2,290,434	87.15	5.62

Table 5
Summary statistics for non-coding RNAs

Type	Copy number	Average Length (bp)	Total Length (bp)	Percentage of Genome (%)
tRNA	333	86.91	28,943	0.0710
rRNA (de novo prediction)	71	116.01	8,237	0.0202
sRNA	69	66.69	4,602	0.0113
snRNA	31	140.12	4,344	0.0107
miRNA	108	55.77	6,024	0.0148

Genes with lengths between 500–999 bp were most abundant (n = 3,663), followed by those of 1000–1499 bp. Genes shorter than 200 bp and longer than 10,000 bp were infrequent (Fig. 5).

Non-coding RNAs

The non-coding RNAs in the Fusarium equiseti D25-1 genome were analyzed. With a copy number of 333, an average length of 86.91 nt, and a total length of 28,943 nt, tRNAs accounted for 0.0171% of the entire genome. The rRNA copy number was 71; the average length was 116.01 nt, and the total length was 8237 nt, accounting for 0.0202% of the genome. sRNAs represented 0.0113% of the genome, with a copy number of 69, an average length of 66.69 nt, and a total length of 4,602 nt. Small nuclear rRNAs accounted for 0.0107% of the genome, with a copy number of 31, an average length of 140.12 nt, and a total length of 4,334 nt. The microRNA copy number was 108; the average length was 55.77 nt, and the total length was 6,024 nt, accounting for 0.0148% of the genome.

Repetitive sequences

Repetitive sequences, including DNA transposons, tandem repeats (TRs), and transposable elements, have important roles in chromosomal spatial structure, regulation of gene expression, and genetic recombination. Transposable elements are further classified into long terminal repeats (LTRs) and non-LTRs, and the latter category includes long interspersed nuclear elements (LINEs) and short interspersed nuclear elements (SINEs). Tables 6 and 7 list the repetitive sequences in the Fusarium equiseti D25-1 genome, as determined by various prediction algorithms and databases. For instance, the TRs predicted by the TRF software spanned 224,134 bp, which accounted for only 0.5497% of the genome. The total predicted repetitive sequences spanned 1,713,918 bp, accounting for 4.2033% of the genome.

Table 6
Repetitive sequence statistics

Method	Repeat Size (bp)	Percentage of Genome (%)
Repbase	561,215	1.3763
ProMask	420,889	1.0322
de novo	1,169,760	2.8687
TRF	224,134	0.5497
Total	1,713,918	4.2033

Table 7
Transposon classification statistics

	Repbase TEs		ProteinMask TEs		De novo TEs		Combined TEs	
Type	Length (bp)	% of Genome	Length (bp)	% of Genome	Length (bp)	% of Genome	Length (bp)	% of Genome
DNA	220,036	0.5396	195,998	0.4807	164,159	0.4026	311,230	0.7633
LINE	63,265	0.1552	59,402	0.1457	27,174	0.0666	98,240	0.2409
LTR	279,482	0.6854	165,768	0.4065	352,607	0.8647	527,778	1.2943
SINE	2,627	0.0064	0	0.0000	9,329	0.0229	11,283	0.0277
Other	0	0.0000	0	0.0000	0	0.0000	0	0.0000
Unknown	1,940	0.0048	0	0.0000	619,726	1.5198	621,666	1.5246
Total	561,215	1.3763	420,889	1.0322	1,169,760	2.8687	1,530,931	3.7545

Gene annotation

As shown in Table 8, 13134 genes, accounting for 94.97% of all the genes in D25-1 genome, were annotated after BLAST searches against all databases. Over 70% of annotations were based on the NR, NOG, and IPR databases. Furthermore, 1422 genes (10.28%) were annotated using the PHI database.

Table 8
Summary of overall annotation results

Sample	Total	ARDB	CAZy	COG	GO	IPR	KEGG	KOG
D25-1	13829	2	397	1516	7261	9995	4646	2231
Proportion		0.01%	2.87%	10.96%	52.5%	72.27%	33.59%	16.13%
Sample	NOG	NR	P450	PHI	SwissProt	T3SS	VFDB	Overall
D25-1	11332	12640	1209	1422	3284	3804	55	13134
Proportion	81.94%	91.4%	8.74%	10.28%	23.74%	27.5%	0.39%	94.97%

Genes encoding toxins were mined based on the annotation results. Two genes related to zearalenone were identified (D25-1_GLEAN_10000531 and D25-1_GLEAN_10000533). D25-1_GLEAN_10000531 was related to the non-reductive iterative type I polyketide synthase involved in the synthesis of zearalenone, whereas D25-1_GLEAN_10000533 was related to the highly reductive iterative type I polyketide synthase (Table 9).

Table 9

Genes related to zearalenone

Gene ID	Identity	Database	Database gene ID	Function
D25-1_GLEAN_10000531	54.84	KEGG	pcs:Pc21g12450	Zearalenone synthase, nonreducing iterative type I polyketide synthase
D25-1_GLEAN_10000533	57.27	KEGG	pcs:Pc21g12440	Zearalenone synthase, highly reducing iterative type I polyketide synthase

Additionally, 23 genes related to trichothecene mycotoxin were identified, including 19 genes related to the active efflux pump of trichothecene mycotoxin, two related to trioxyacetyltransferase of trichothecene mycotoxin, and two related to the biosynthesis of trichothecene mycotoxin. Many of these genes were obtained by searches against the NOG database (Table 10).

Table 10

Genes related to trichothecene

Gene ID	Identity (%)	Database	Database gene ID	Function
D25-1_GLEAN_10001590	42.83	sordNOG	JGI67026	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10001811	76.04	ascNOG; euNOG; fuNOG; opiNOG	EFQ35752	Trichothecene 3-O-acetyltransferase
D25-1_GLEAN_10002084	87.89	hypNOG; necNOG	FGSG_10823P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10002452	76.6	hypNOG; necNOG; sordNOG	FGSG_12768P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10003513	78.59	hypNOG	NechaP36294	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10003680	94.59	hypNOG; necNOG	FGSG_08749P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10005110	91.85	hypNOG; necNOG	FGSG_05352P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10005621	80.1	hypNOG; necNOG; sordNOG	XP_387212.1	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10006448	69.65	necNOG	NechaP53897	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10006579	88.71	ascNOG; hypNOG; necNOG; opiNOG; sordNOG	XP_383902.1	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10006973	71.71	NOG; ascNOG; euNOG; hypNOG; fuNOG; necNOG; opiNOG; sordNOG	FVEG_00056T0	Trichothecene 3-O-acetyltransferase
D25-1_GLEAN_10006980	85.6	NOG; ascNOG; euNOG; hypNOG; fuNOG; necNOG; opiNOG; sordNOG	FGSG_03537P0	Trichothecene biosynthesis
D25-1_GLEAN_10009924	89.23	ascNOG; fuNOG; opiNOG	FGSG_02343P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10010021	85.26	hypNOG; necNOG	FGSG_12141P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10010553	91.59	necNOG	FVEG_04795T0	Fungal trichothecene efflux pump (TRI12)

Gene ID	Identity (%)	Database	Database gene ID	Function
D25-1_GLEAN_10011281	86	hypNOG	FOXG_01267P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10011497	83.28	hypNOG; sordNOG	FGSG_11815P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10011948	82.65	hypNOG; necNOG; sordNOG	XP_381015.1	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10013026	92.83	necNOG; sordNOG	XP_389873.1	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10013028	94.49	ascNOG; fuNOG; opiNOG; hypNOG; sordNOG	FGSG_09701P0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10013718	54.74	necNOG	FVEG_04795T0	Fungal trichothecene efflux pump (TRI12)
D25-1_GLEAN_10011175	53.63	NR	gi|751354543|gb|KIL92265.1|	tri7-trichothecene biosynthesis gene cluster
D25-1_GLEAN_10009924	/	IPR	PF06609	Fungal trichothecene efflux pump (TRI12)

Comparative genomic analyses

As summarized in Table 11, the newly assembled F. equiseti genome had the most intact sequence, with 16 contigs and an N50 of 6,178,397 bp, indicating a high-quality assembly. In a comparative analysis, the best-assembled F. oxysporum genome only had 33 scaffolds, with an N50 of 4,490,135 bp, which was superior to the assembly results for other strains. The differences in quality might be explained by the differences in assembly technology and platform. In this study, data were obtained using the third-generation PacBio platform and second-generation Illumina platform for joint assembly; whereas, most previously reported data had been obtained using the second-generation sequencing technology alone.

Table 11
Genome information

Sample_Name	Seq Type	Total Number	Total Length (bp)	N50 Length (bp)	N90 Length (bp)	Max Length (bp)	Min Length (bp)	Gap Number (bp)	GC Content (%)
Bipolaris sorokiniana	Scaffold	154	34,409,167	1,789,485	1,003,746	3,642,493	2,011	1,196,549	49.84
F. avenaceum	Scaffold	83	41,590,745	1,436,644	424,894	4,337,333	602	29,282	48.47
F. oxysporum	Scaffold	33	52,908,293	4,490,135	2,466,030	6,470,671	4,587	0	47.67
F. pseudograminearum	Scaffold	281	36,973,259	8,840,934	7,724,594	11,688,822	502	40,400	47.75
Nectria haematococca	Scaffold	209	51,286,497	1,255,602	96,667	4,937,060	865	56,137	50.79
F. equiseti D25-1	Contig	16	40,776,005	6,178,397	2,783,306	8,344,890	2,316	-	48.01

Structural variation

Figure 6 summarizes the genomic collinearity results. The highest collinearity (i.e., greatest conservation) was found for F. pseudograminearum, which had high similarity with F. equiseti D25-1, followed by F. avenaceum and F. oxysporum, whereas B. sorokiniana had the lowest similarity. In addition, the gene number in the reference B. sorokiniana genome was 5,133; however, in a collinearity analysis, only 37.12% of the target genes were covered. The N. haematococca reference genome had 7,123 genes, covering 51.51% of the target genes. The F. pseudograminearum genome had 10,052 genes, covering 72.69% of the target genes. The F. avenaceum genome had 9884 genes, covering 71.47% of the target genes. The F. oxysporum genome had 10,236 genes, covering 74.02% of the target genes. Such findings might be related to the large number of reference genes from F. oxysporum.

Core-pan genome

Based on core-pan genome analysis, the six strains shared 1,805 core genes. As an outgroup, Bipolaris sorokiniana had the most species-specific genes (n = 8,912), followed by Nectria haematococca (n = 5,759), Fusarium oxysporum (n = 4,946), Fusarium equiseti (n = 3,483), Fusarium avenaceum (n = 2,614), and Fusarium pseudograminearum (n = 2,299), respectively (Fig. 7).

COG was used to annotate the core genes (Fig. 8) in four modules (cellular processes, genetic information storage and transmission, metabolism, and unknown function). Regarding cellular processes, 3 genes were associated with cell cycle control, cell division, and chromosome partitioning; 14 were involved in cell wall/membrane/envelop biogenesis; 1 gene was related to cell mobility; 5 were related to defense mechanisms; 1 was associated with the cytoskeleton; 1 was associated with intracellular trafficking, secretion, and vesicular transport; 51 were related to posttranslational modification, protein turnover, and chaperones; and 6 were related to signal transduction mechanisms. Five terms in the genetic information storage and transmission module were overrepresented; 1 gene was responsible for chromatin structure and dynamics; 1 was related to RNA processing and modification; 17 were related to replication, recombination, and modification; 14 were related to transcription; and 80 were related to translation, ribosomal structure, and biogenesis. In the metabolism module, 84 genes were associated with amino acid transport and metabolism; 63 were related to carbohydrate transport and metabolism; 38 were associated with coenzyme transport and metabolism; 69 were related to energy production and conversion; 20 were related to inorganic ion transport and metabolism; 58 were linked with lipid transport and metabolism; 35 were related to nucleotide transport and metabolism; and 36 were related to secondary metabolite biosynthesis, transport, and catabolism.

Genes specific to the Fusarium equiseti D25-1 genome were annotated using COG (Fig. 9) based on four modules (cellular processes, genetic information storage and transmission, metabolism, and unknown function). In the cellular processes module, 2 genes were associated with cell cycle control, cell division, and chromosome partitioning; 3 were associated with cell wall/membrane/envelop biogenesis; 1 gene was related to defense mechanisms; 5 were related to posttranslational modification, protein turnover, and chaperones; and 5 were related to signal transduction mechanisms. In the genetic information storage and transmission module, 17 genes were associated with replication, recombination, and modification, and 7 were related to translation, ribosomal structure, and biogenesis. In the metabolism module, 12 genes were associated with amino acid transport and metabolism; 13 were related to carbohydrate transport and metabolism; 7 were associated with coenzyme transport and metabolism; and 13 were related to secondary metabolites biosynthesis, transport, and catabolism.

Phylogenetic analysis based on the whole genomes (Fig. 10) indicated that F. equiseti was most closely related to F. pseudograminearum, followed by F. avenaceum and F. oxysporum.

Gene family analysis

As shown in Table 12, the clustered genes in the outgroup taxon Bipolaris sorokiniana accounted for 72% of the total genes, whereas the genes assigned to clusters in Fusarium solani accounted for over 90% of the genes. Typically, Fusarium equiseti D25-1 gene families exceed 60%, with 54 species-specific gene families.

Table 12
Gene family clustering analysis

Sample	Gene Number	Clustered Genes	Unclustered Genes	Families	Unique Families
Bipolaris sorokiniana	12,154	8,692	3,462	5,708	213
Fusarium oxysporum	16,792	15,288	1,504	8,726	139
Nectria haematococca	15,647	14,091	1,556	7,901	109
Fusarium equiseti D25-1	13,829	12,654	1,175	8,493	54
Fusarium avenaceum	13,092	12,476	616	8,181	9
Fusarium pseudograminearum	12,348	11,446	902	8,117	7

Single-copy orthologs were detected in the genome of each strain (n > 2,500). Multiple-copy orthologs were also detected in each genome, but the counts differed among the species. For instance, Fusarium oxysporum and Nectria haematococca had high numbers of multi-copy orthologs (Fig. 11).
Gene family-based clustering analysis (Fig. 12) indicated that F. equiseti D25-1 was most similar to F. pseudograminearum, followed by F. avenaceum and F. oxysporum. These results were consistent with the results of the genome-wide sequence analysis, as well as the evolutionary position of F. equiseti in a genome-based phylogenetic analysis.

Discussion

The whole-genome sequencing assembly of Fusarium equiseti D25-1 was based on a combination of results from Illumina Hiseq 4000 and PacBio platforms. The primary assembly was carried out on Illumina Hiseq 4000 platform, and then the rough analysis of the genomic data and pre-determination of the genome size were conducted, providing a basic reference for an advanced assembly on PacBio platform. This platform provides more accurate results, and the assembly effect is better.

Meanwhile, the gene composition and distribution, and the number of non-coding RNA genes, repeat sequences, and transposon types were identified. Non-coding RNAs play a crucial role in the regulation of mRNA translation, localization, and stability, in addition to other processes [28]. Numerous studies have shown that miRNAs play key roles in the development of cancer [29, 30]. Thus, the roles of F. equiseti miRNAs in pathogenesis should be evaluated. Transposons can affect the coding capacity of genes and can even lead to chromosomal rearrangements [31]. Identification, classification, and annotation of transposons can be accomplished more accurately over whole genomes. Therefore, the genome-wide annotation of F. equiseti D25-1 transposons can provide a basis for the detection of functionally and evolutionarily important genes, including those associated with pathogenicity.

On the basis of these basic data, gene function was annotated. Gene function annotation was used to comprehensively annotate the basic functions of F. equiseti D25-1 through various databases. Numerous studies have shown that F. equiseti is pathogenic [32, 33]. Additionally, fusarium can produce multiple toxins, including zearalenone, trichothecene mycotoxin, moniliformin, and fumonisins [34]. Hence, the genes that confer pathogenicity to fungi were also annotated using the carbohydrate-related enzyme (CAZy), PHI, KEGG, and cytochrome P450 databases. The derived data can be used to explore the factors and mechanism(s) underlying the pathogenicity of F. equiseti.

Currently, the genomic sequences of many plant pathogenic fungi can be retrieved from the NCBI database. This study selected Fusarium oxysporum, Nectria haematococca, Fusarium pseudograminearum, Fusarium avenaceum, and Bipolaris sorokiniana as the reference strains for Fusarium equiseti D25-1 because these strains are crop root rot pathogens. For example, F. pseudograminearum and N.haematococca can cause root rot in barley [35] and physic nut [36], respectively. Our study showed that these pathogens have many common genes, such as the core genes described here. This similarity may cause them to give rise to the same disease. In addition, genes

belonging to many families have been identified in the genomes of the six strains through gene family analysis, and the functional characteristics of each gene family need to be further studied.

Conclusion

In this study, the complete genome sequence of F. equiseti D25-1 was assembled based on the Illumina HiSeq 4000 and PacBio platforms. The primary assembled genome had a size of 40.55 Mb and a GC content of 47.92%. The advanced assembly included 16 chromosomes, with a GC content of 48.01%, a total size of 40,776,005 bp, and a gap number of 0. The introns, exons, gene lengths, non-coding RNA, and repetitive sequences were characterized. Based on 14 databases, 13,134 functional genes were annotated in the Fusarium equiseti D25-1 genome, accounting for a high proportion (94.97%) of the total genes. Moreover, multiple genes were related to cellular processes, metabolism, molecular functions, and pathogenicity, including two genes related to zearalenone and 23 genes associated with trichothecene mycotoxin. High collinearity with the genome sequence of the reference strain was observed, with 3,483 specific genes, 1,805 common genes, and over 2,500 single-copy orthologs. The genome had the highest sequence similarity with the genome of F. pseudograminearum, followed by F. avenaceum, providing insight into its evolutionary position. The genomic data provide a basis for studies of gene expression, regulatory mechanisms, functional mechanisms, evolutionary processes, and disease control in F. equiseti and other fungi.

Methods

Materials

F. equiseti D25-1 was isolated from the rotten root of highland barley in our preliminary study; it was identified as a pathogen of highland barley root rot disease according to Koch's postulates.

Strain cultivation and genomic DNA extraction

F. equiseti D25-1 was inoculated into a 500-mL flask containing 200-mL sterile potato dextrose broth and cultured at 25 °C in a 120 r/min shaker for 5 d. After filtration, mycelia were collected by centrifugation at 7,168 g for 1 min, followed by DNA extraction using the OMEGA Fungal DNA Kit (Biel, Switzerland) as per the manufacturer's instructions.

Illumina HiSeq 4000 sequencing

After quality control, DNA samples were used to construct a library. Large DNA fragments (genomic DNA, BACs, or long-range PCR products) were randomly cleaved by Covaris or Bioruptor ultrasonication to generate a series of DNA fragments with the main band of 800 bp or less. Next, T4 DNA Polymerase, Klenow DNA Polymerase, and T4 PNK were used to repair the ensuing sticky ends into blunt ends, and the base "A" was added to the 3' ends to allow the DNA fragments to become ligated to special adapters containing "T" at their 3' ends. Next, the desired ligation product was identified by electrophoresis, and the DNA fragments with adapters at both ends were amplified by PCR. Finally, cluster preparation and sequencing were performed using a qualified library (sequencing was completed with the BGI Tech APAC microbial line).

PacBio sequencing

The genomic DNA was processed by g-TUBE into fragments of appropriate size, after which damage- and end-repair were performed. Both ends of the DNA fragments were ligated to the hairpin adapters to form a dumbbell structure designated SMRTbell. The annealed SMRTbell was mixed with polymerase at the bottom of ZWM for final sequencing (sequencing was completed with the BGI Tech APAC microbial line).

Illumina HiSeq 4000 data filtering

To filter low-quality data and thereby ensure the accuracy and reliability of the subsequent analyses, 1–125 bp of read 1 and read 2 were first obtained. Reads with quality values ≤ 2 were removed, as were those with a total of 40% "N" bases. Adapters and duplications were eliminated.

PacBio data filtering

The raw sequencing data from the PacBio platform contained the adapter sequences, low-quality sequences, erroneous sequences, and other reads requiring processing. To improve the assembly results, the following steps were taken: 1) polymerase reads < 1,000 bp were

removed; 2) polymerase reads with qualities < 0.80 were removed; 3) sub-reads were extracted from polymerase reads, and adapter sequences were removed; 4) sub-reads < 1,000 bp were removed.

Genome assembly

The sequencing data were assembled using a variety of tools. 1) During the sub-read correction, multiple algorithms were used for self-correction (Pbdagcon and FalconConsensus) or hybrid correction (Proovread) of the sub-reads to obtain highly reliable corrected reads. 2) For corrected read assembly, Celera and Falcon were used to finalize the optimal assembly. 3) For assembly correction, single-base correction (GATK) was performed using the second-generation Illumina short reads to obtain a highly reliable assembly sequence. 4) For scaffolding, SSPACE Basic v2.0 was used based on the second-generation Illumina long reads, and the scaffold gaps were filled using PBjelly2 [37–39].

Gene prediction

Gene component analyses were primarily performed using Homology [40], SNAP [41], Augustus [42], and GeneMark-ES [43]. Homology prediction was implemented in GeneWise. SNAP and Augustus were used for predictions using training sets for the reference species.

Non-coding RNA analysis

For the non-coding RNA analysis, rRNAs were detected by comparing the rRNA library with the non-coding RNAs predicted by RNAmmer 1.2 [44]. The domains and secondary structures of the tRNAs were predicted using tRNAscan 1.3.1 [45]. The sRNAs were identified by comparing with the Rfam 9.1 [46] database using Infernal.

Repetitive sequence analysis

Transposons were predicted via three approaches–Repeat Masker 4-0-6 (using Repbase database), RepeatProteinMasker (using its own transposon protein library), or de novo (a database of transposon sequences was generated using buildXDFDatabase, and then a transposon model was constructed according to this database using Repeat Modeler, followed by transposon identification from the constructed model using Repeat Masker). The tandem repeats were predicted using TRF 4.04 (Tandem Repeat Finder) [47].

Gene annotation and virulence genes

The protein sequences were functionally annotated. Gene sequences were aligned with the available sequences in databases to obtain corresponding annotations. The highest quality alignment was chosen for gene annotation. Functional annotation was performed by BLAST searches against the following databases: Gene Ontology (GO) [48], Kyoto Encyclopedia of Genes and Genomes (KEGG) [49], Cluster of Orthologous Groups of proteins (COG) [50], Swiss-Prot [51], Trembl 2016, NR 2015, EggNOG 4.5, Antibiotic Resistance Genes Database (ARDB) 1.1 [52], Pathogen Host Interactions (PHI) 4.0 [53], Fungal Cytochrome P450 Database1.1 [54], Carbohydrate-Active enzymes Database (CAZy), virulence factor database (VFDB) [55], Type III secretion system Effector protein (T3SS) 1.0 [56], and TransportDB 2.0. Finally, the genes related to toxin production were identified.

Comparative Genomics

A comparative genomic analysis was performed with the whole-genome sequences of Bipolaris sorokiniana (INSDC: AEIN00000000.1), Fusarium avenaceum (INSDC: JPYM00000000.1), Fusarium oxysporum (INSDC: MABQ00000000.2), Fusarium pseudograminearum (INSDC: AFNW00000000.1), and Nectria haematococca (INSDC: ACJF00000000.1) as the reference.

Structural variation

The sequence of the target fungus was ordered according to that of the reference fungus using MUMmer [57]. Protein set P1 of the target fungus was aligned with protein set P2 of the reference fungus. P1 was first aligned with P2 using BLASTP by setting P2 as the database, and the best hit for each protein was selected. The reciprocal alignment was then performed (by setting P1 as the database). Finally, the results with the best hit values for both alignments were retained, and the average of two consistent values was obtained.

Core-pan genome

Genes from the reference genome were used as the gene pool. Predicted genes from Query samples were BLAST-searched against the gene pool, and the BLAST results were filtered by length and identity. The BLAST coverage ratios of the genes from the gene pool and Query samples were calculated separately. Finally, a tree was constructed based on the multiple sequence alignment obtained using Muscle by the neighbor-joining method implemented in Treebest [58]. Completed core and specific genes were annotated using BLAST and the COG database [59].

Gene family analysis

Gene families were constructed for the reference genes and the genes of the target fungus. Protein sequences were aligned using BLAST, and redundancy was eliminated using SOLAR. The gene family database TreeFam was used for the clustering analysis using Hclustersg. The protein alignment results for gene families were converted to the amino acid sequences of the CDS regions using Muscle [60]. The gene family tree was constructed based on the multiple sequence alignment results by the neighbor-joining method using Treebest.

Abbreviations

BLAST
Basic local alignment search tool
GO
Gene Ontology
KEGG
Kyoto Encyclopedia of Genes and Genomes
COG
Cluster of Orthologous Groups of proteins
ARDB
Antibiotic Resistance Genes Database
PHI
Pathogen Host Interactions
CAZy
Carbohydrate-Active enzymes Database
VFDB
Virulence factor database
T3SS
Type III secretion system Effector protein
NOG
Non-supervised Orthologous Groups

Declarations

Availability of data and materials

The datasets generated and/or analyzed in the current study can be obtained from the corresponding/first author or DDBJ/ENA/GenBank repository under the accession number QOHM00000000, https://www.ncbi.nlm.nih.gov/bioproject/?term=QOHM00000000&utm_source=gquery&utm_medium=search.

Acknowledgements

Not applicable

Funding

This study was supported by the Youth Fund for Gansu Academy of Agricultural Sciences

(Grant/Award Number: 2019GAAS34) and the Special Fund for Agro-Scientific Research in the

Public Interest (201503112).

Author information

Affiliations

Institute of Plant Protection, Gansu Academy of Agricultural Sciences, Lanzhou, 730070, China

Xueping Li, Yonghong Qi, Yonggang Liu & Minquan Li

College of Prataculture, Gansu Agricultural University, Lanzhou, 730070, China

Jianhong Li

Contributions

Xueping Li was the lead investigator of this study. Jianhong Li assisted in DNA extraction.

Yonghong Qi helped to collect the samples. Yonggang Liu and Minquan Li served as consultants on

the experimental designs and suggested some modifications.

Corresponding author

Correspondence to Xueping Li

Ethics declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no competing interests.

References

1. Ikediobi CO, Ibrahim S, Ikoku OA: Linamarase from *Fusarium equiseti*. *Appl Microbiol Biot* 1987, 25:327-333. https://doi.org/10.1007/BF00252542
2. Yoshihito S, Fumiaki S, Tetsuya M: A polyketide metabolite from an Endophytic *Fusarium equiseti* in a Medicinal Plant. *Z. Naturforsch* 2013, 68:289-292. https://doi.org/10.5560/znb.2013-3014
3. Wang H, Liu T, Xin Z: A new glucitol from an endophytic fungus *Fusarium equiseti* Salicorn 8. *Euro Food Res Technol* 2014, 239:365-376. https://doi.org/10.1007/s00217-014-2230-z
4. Usama WH, Radwan AF, Lamia TA, Adnan JT: Different culture metabolites of the Red Sea fungus *Fusarium equiseti* optimize the inhibition of hepatitis C virus NS3/4A protease (HCV PR). *Mar Drugs* 2016, 14:2-12. https://doi.org/10.3390/md14100190
5. Marín P, Moretti A, Ritieni A: Phylogenetic analyses and toxigenic profiles of *Fusarium equiseti* and *Fusarium acuminatum* isolated from cereals from Sotuhern Europe. *Food Microbiol* 2012, 31:229-237. https://doi.org/10.1016/j.fm.2012.03.014
6. Horinouchi H, Muslim A, Hyakumachi M: Biocontrol of *Fusarium* wilt of spinach by the plant growth promoting fungus *Fusarium equiseti* gf183. *J Plant Pathol* 2010, 92:249-254. https://doi.org/10.2307/41998793
7. Horinouchi H, Watanabe H, Taguchi Y: Biological control of Fusarium wilt of tomato with *Fusarium equiseti* GF191 in both rock wool and soil systems. *Bio Control* 2011, 56:915-923. https://doi.org/10.1007/s10526-011-9369-3
8. Wahab AA, Awang ASAH, Azham Z. Factors affecting toxic lead (II) ion bioremediation by *Fusarium equiseti* isolated from the mangrove soil environment of sotuheast Borneo. *Malays J Microbiol* 2015, 11:215-222.
9. Shikur GE, Sharma PD, Paulitz TC: Identity and pathogenicity of *Fusarium* species associated with crown rot on wheat (*Triticum* spp.) in Turkey. *Euro J Plant Pathol* 2018, 150:387-399. https://doi.org/10.1007/s10658-017-1285-7.
10. Li PP, Cao ZY, Wang K: First report of *Fusarium equiseti* causing a sheath rot of corn in China. *Plant Dis* 2014, 98:998. https://doi.org/10.1094/PDIS-10-13-1088-PDN
11. Kosiak B, Thorp M, Skjere E, Thrane U: The prevalence and distribution of *Fusarium* species in Norwegian cereals: a survey. *Acta Agr Scand B-S P* 2003, 53:168-176. https://doi.org/10.1080/09064710310017272
12. Li YG, Zhang SQ, Sun L P: First report of root rot of cowpea caused by *Fusarium equiseti* in Georgia in the United States. *Plant Dis* 2017, 101:1674. https://doi.org/10.1094/PDIS-03-17-0358-PDN

13. Zhou QX, Li NN, Chang KF: Genetic diversity and aggressiveness of *Fusarium* species isolated from soybean in Alberta, Canada. *Crop Prot* 2018, 105:49-58. https://doi.org/10.1016/j.cropro.2017.11.005
14. Seo Y, Kim YH: Potential reasons for prevalence of *Fusarium* wilt in oriental melon in Korea. *Plant Pathol J* 2017, 33:249-263. https://doi.org/10.5423/PPJ.OA.02.2017.0026
15. Garibaldi A, Gilardi G, Ortu G, Gullino ML: First report of leaf spot of lettuce (*Lactuca sativa*) caused by *Fusarium equiseti* in Italy. 2016, *Plant Dis* 100:531. https://doi.org/10.1094/PDIS-06-15-0686-PDN
16. Gao J, Zhang YY, Wang K: Identification of sunflower wilt pathogen and its biological characteristics. *Chin J Oil Crop Sci* 2016, 38:214-222. https://doi.org/10.7505/j. issn.1007-9084
17. Garibaldi A, Gilardi G, Matic S, Gullino ML: Occurrence of *Fusarium equiseti* on *Raphanus sativus* seedlings in Italy. *Plant Dis* 2017, 101: 1548. https://doi.org/10.1094/PDIS-03-17-0417-PDN
18. Lazarotto M, Muniz MFB, Santos RF: First report of *Fusarium equiseti* associated on pecan (*Carya illinoinensis*) seeds in Brazil. 2014, *Plant Dis* 98:847. https://doi.org/10.1094/PDIS-09-13-0976-PDN
19. Lazreg F, Belabid L: First report of *Fusarium equiseti* causing damping-off disease on aleppo pine in Algeria. *Plant Dis* 2014, 98:1268. https://doi.org/10.1094/PDIS-02-13-0194-PDN
20. Ramchandra S, Bhatt PN: First report of *Fusarium equiseti* causing vascular wilt of cumin in India. *Plant Dis* 2012, 96:1821. https://doi.org/10.1094/PDIS-03-12-0236-PDN
21. Ashfaq M, Anjum MA, Hafeez R: First report of *Fusarium equiseti* causing brown leaf spot of fishtail palm (*Caryota mitis*) in Pakistan. *Plant Dis* 2017, 101:840. https://doi.org/10.1094/PDIS-11-16-1585-PDN
22. Gupta V, Razdan VK, John D, Sharma BC: First report of leaf blight of *Cyperus iria* caused by *Fusarium equiseti* in India. *Plant Dis* 2013, 97:838. https://doi.org/10.1094/PDIS-07-12-0690-PDN
23. Lu NH, Huang QZ, He H: First report of black stem of *Avicennia marina* caused by *Fusarium equiseti* in China. *Plant Dis* 2014, 98:843. https://doi.org/10.1094/PDIS-08-13-0873-PDN
24. Zhou JH, Yan SJ, Wang CF, Mao WL: Identification and pathogenicity analysis of a fungal strain isolated from mulberry plant with blight disease. *Sci Sericult* 2013, 39:1066-1070. https://doi.org/10.13441/j.cnki.cykx.2013.06.014
25. Xue CY, Yan XR, Lin TX: A preliminary report on occurrence of *Schisandra* berry stalk rot. *Plant Prot* 2007, 33:96-100.
26. Stępień Ł[1], Gromadzka K, Chełkowski J: Polymorphism of mycotoxin biosynthetic genes among *Fusarium equiseti* isolates from Italy and Poland. *J Appl Genet* 2012, 53:227-236. https://doi.org/10.1007/s13353-012-0085-1
27. Kari Juntunen Susanna Mäkinen, Sari Isoniemi: A New Subtilase-Like Protease Deriving from *Fusarium equiseti* with High Potential for Industrial Applications. *Appl Biochem Biotechnol* 2015, 177: 407-430.
28. Belew AT, Meskauskas A, Musalgaonkar S: Ribosomal frameshifting in the CCR5 mRNA is regulated by miRNAs and the NMD pathway. *Nature* 2014, 512:265-269. https://doi.org/10.1038/nature13429
29. Chen LL, Gu H, Peng J: Non-coding RNA and pancreatic cancer. *Journal of Central South University* 2014, 39:1672-7347. https://doi.org/10.11817/j.issn.1672-7347.2014.05.015
30. Chen L, Shan G: A brief introduction of noncoding RNA research. *Chinese Sci Bull* 2017, 62:3236-3244. https://doi.org/10.1360/N972017-00384
31. Xu HE, Zhang HH, Han MJ: Computational approaches for identification and classification of transposable elements in eukaryotic genomes. *Hereditas* 2012, 34: 1009-1019. https://doi.org/10.3724/SP.J.1005.2012.01009
32. Marin P, Moretti A, Ritieni A, Jurado M, Vazquez C, Gonzalez-Jaen MT: Phylogenetic analyses and toxigenic profiles of *Fusarium equiseti* and *Fusarium acuminatum* isolated from cereals from Southern Europe. *Food Microbiol* 2012, 31, 229-237.
33. Guo MP, Chen K, Wang GZ: First Report of Stipe Rot Disease on Morchella imporituna Caused by Fusarium incarnatum-equiseti Species Complex in China. *Plant Dis* 2016, 100(12): 2530.
34. Peng J, Wu XP, Huang HQ, Bao SX: Research development of *Fusarium* toxins. *Chinese Agr Sci Bull* 2009, 25: 25-27.
35. Fernandez MR, Zentner RP, DePauw RM: Impacts of crop production factors on fusarium head blight in barley in eastern Saskatchewan. *Crop Sci* 2007, (47): 1574-1584.
36. Wu Y, Ou G, Yu J: First report of Nectria haematococca causing root rot disease of physic nut (Jatropha curcas) in China. *Australasian Plant Dis* 2011, 6: 39-42.
37. Tsuji M, Kudoh S, Hoshino T: Draft genome sequence of cryophilic basidiomycetous yeast *Mrakia blollopis* SK-4, isolated from an algal mat of Naga-ike Lake in the Skarvsnes ice-free area, East Antarctica. *Genome Announc* 3 2015, e01454-14.

https://doi.org/10.1128/genomeA.01454-14

38. Badouin H, Hood ME, Gouzy J: Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus *Microbotryum lychnidis-dioicae. Genetics* 2015, 200:1275-1284. https://doi.org/10.1534/genetics.115.177709
39. Faino L, Seidl MF, Datema E: Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *MBio* 2015, 6:e00936-15. https://doi.org/10.1128/mBio.00936-15
40. Birney E, Clamp M, Durbin R: GeneWise and Genomewise. *Genome Res* 2004, 14:988-995. https://doi.org/10.1101/gr.1865504
41. Johnson AD, Handsaker RE: SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap. *Bioinformatics* 2008, 24:2938-2939. https://doi.org/10.1093/bioinformatics/btn564
42. Stanke M, Diekhans M, Baertsch R: Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* 2008, 24:637-644. https://doi.org/10.1093/bioinformatics/btn013
43. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO: Gene prediction in novel fungal genomes using an abinitio algorithm with unsupervised training. *Genome Res* 2008, 18:1979-1990. https://doi.org/10.1101/gr.081612.108
44. Lagesen K, Hallin PF, Rødland E: RNAmmer: consistent and rapid annotation of ribosomal RNA Genes. *Nucleic Acids Res* 2007, 35:3100-3108. https://doi.org/10.1093/nar/gkm160
45. Lowe TM, Eddy SR: Trnascan-SE: A program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 1997, 25:0955-0964. https://doi.org/10.1093/nar/25.5.0955
46. Gardner PP, Daub J, Tate JG: Rfam: updates to the RNA families database. *Nucleic Acids Res* 2009, 37:D136-D140. https://doi.org/10.1093/nar/gkn766
47. Benson G: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res* 1999, 27:573-580. https://doi.org/10.1093/nar/27.2.573
48. Philip Jones, David Binns, Hsin-Yu Chang: InterProScan 5: genome-scale protein function classification. *Bioinformatics* 2014, btu031.
49. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M: KEGG as a reference resource for gene and protein annotation. *Nucleic Acids Res* 2016, 44:D457-D462. https://doi.org/10.1093/nar/gkv1070
50. Makarova, Kira S, Yuri I. Wolf, Eugene V: Koonin. Archaeal clusters of orthologous genes (arCOGs): an update and application for analysis of shared features between thermococcales, methanococcales, and methanobacteriales. *Life* 2015, 5:818-840. https://doi.org/10.3390/life5010818
51. UniProt Consortium: UniProt: a hub for protein information. *Nucleic Acids Res* 2014, 43: D204-D212. https://doi.org/10.1093/nar/gku989
52. Liu B, Pop M: ARDB-Antibiotic Resistance Genes Database. *Nucleic Acids Res* 2009, 37: D443-447. https://doi.org/10.1093/nar/gkn656
53. Trudy TA, Candace WC, Michelle GG: The Plant-Associated Microbe Gene Ontology (PAMGO) Consortium: community development of new Gene Ontology terms describing biological processes involved in microbe-host interactions. *BMC Microbiol* 2009, 9: S1-S1. https://doi.org/10.1186/1471-2180-9-S1-S1
54. Fischer M, Fischer M, Knoll M, Sirim D, Wagner F, Funke S, Pleiss J: The Cytochrome P450 Engineering Database: a navigation and prediction tool for the cytochrome P450 protein family. *Bioinformatics* 2007, 23:2015-2017. https://doi.org/10.1093/bioinformatics/btm268
55. Chen LH, Zheng DD, Liu B, Yang J, Jin Q: VFDB 2016: hierarchical and refined dataset for big data analysis-10 years on. *Nucleic Acids Res* 2016, 44:D694-D697. https://doi.org/10.1093/nar/gkv1239
56. Vargas WA, Martín JM, Rech GE , Rivera LP, Benito EP, Díaz MJM, Thon MR, Sukno SA: Plant defense mechanisms are activated during biotrophic and necrotrophic development of *Colletotricum graminicola* in maize. *Plant Physiol* 2012, 158:1342-1358. https://doi.org/10.1104/pp.111.190397
57. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzburg SL () Versatile and open software for comparing large genomes. *Genome Biol* 2004, 5R12. https://doi.org/10.1186/gb-2004-5-2-r12
58. Tannistha Nandi, Catherine Ong, Arvind Pratap Singh: A Genomic Survey of Positive Selection in Burkholderia pseudomallei Provides Insights into the Evolution of Accidental Virulence. *PLoS Pathogens* 2010, 6: 1-15.
59. Galperin MY, Makarova KS, Wolf YI: Expanded microbial genome coverage and improved protein family annotation in the COG database. *Nucleic Acids Res* 2015, 43:D261-D269. https://doi.org/10.1093/nar/gku1223
60. Edgar RC: MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 2004, 5:113. https://doi.org/10.1186/1471-2105-5-113

Figures

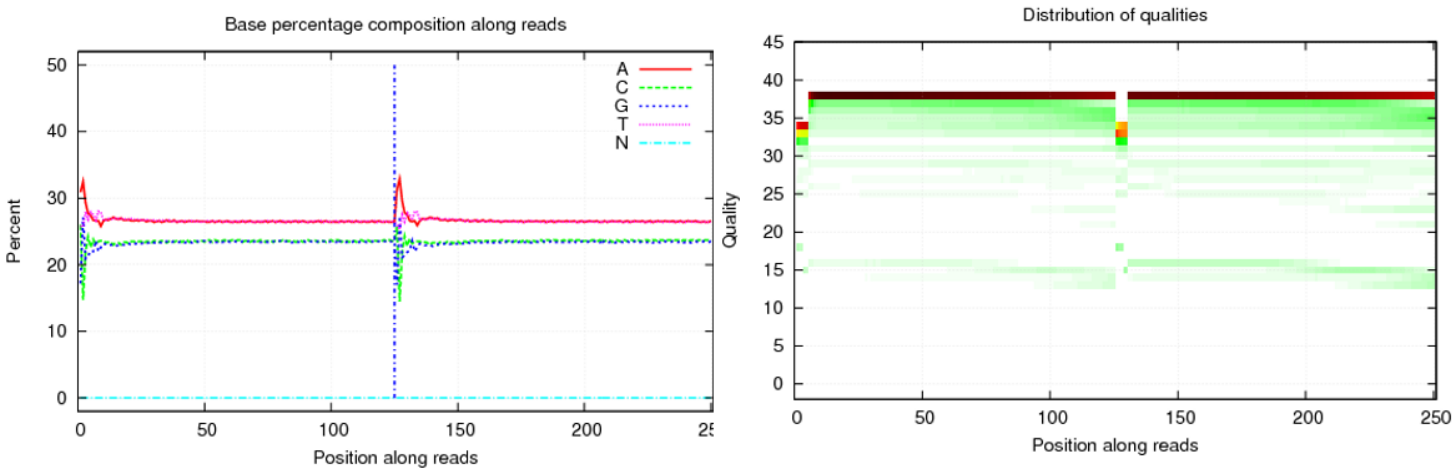

Figure 1

Base and quality distribution. The figure on the left shows the distribution of the bases after filtration. The X-axis represents the base positions of read 1 and read 2, and the vertical axis represents the distribution percentage of each base. The figure on the right shows the mass distribution of the bases, and each point in the figure represents the mass value of the base at the corresponding position in a read.

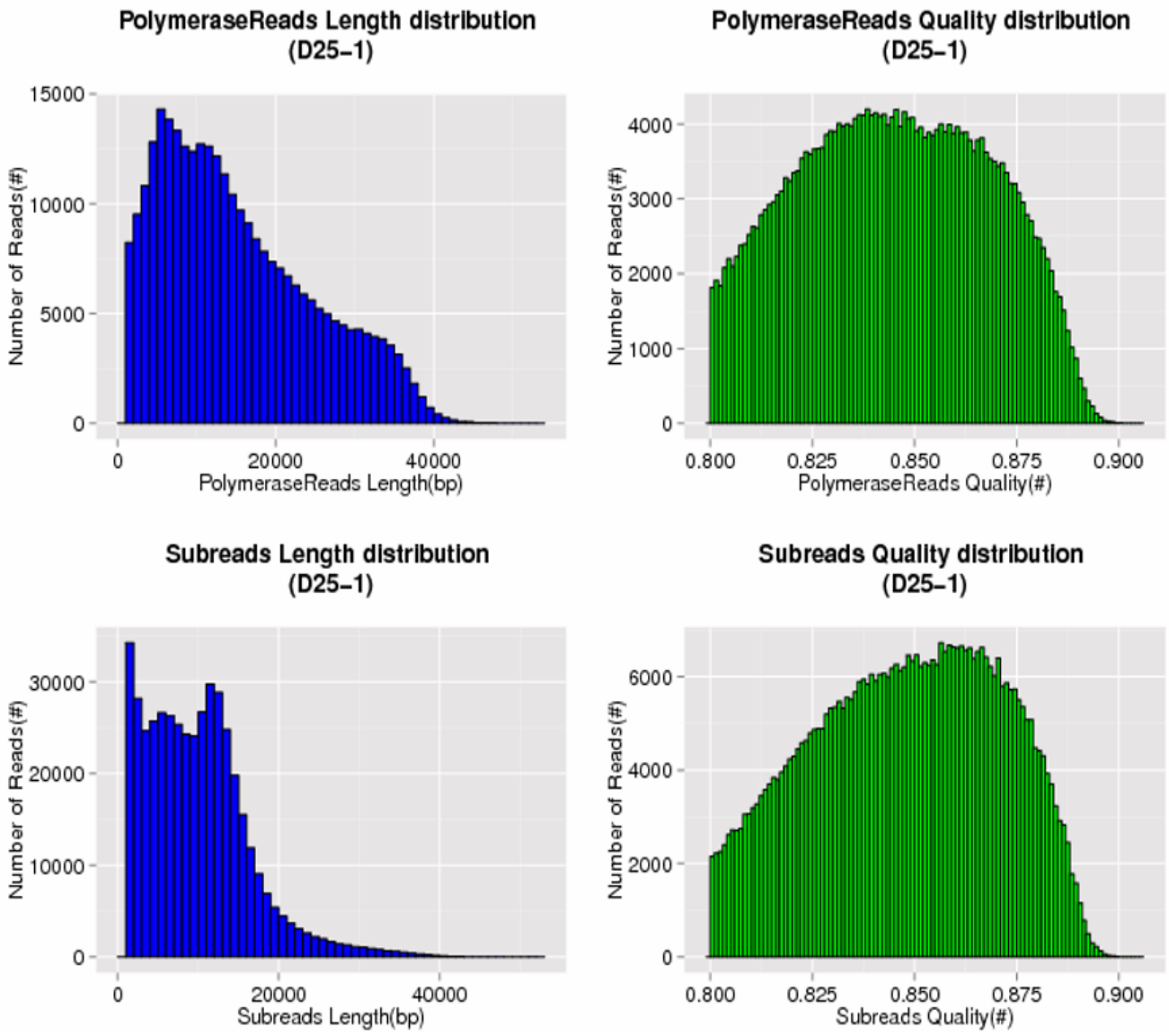

Figure 2

PacBio read length and quality distribution. The top and bottom left corners in the chart show the distributions of the polymerase read and sub-read lengths, respectively. Abscissa and ordinate represent the lengths and numbers of the polymerase reads/subreads, respectively. The top and bottom right corners in the chart show the quality distributions of the polymerase reads and sub-reads, respectively. Abscissa and ordinate represent the qualities and numbers of the polymerase reads/subreads, respectively.

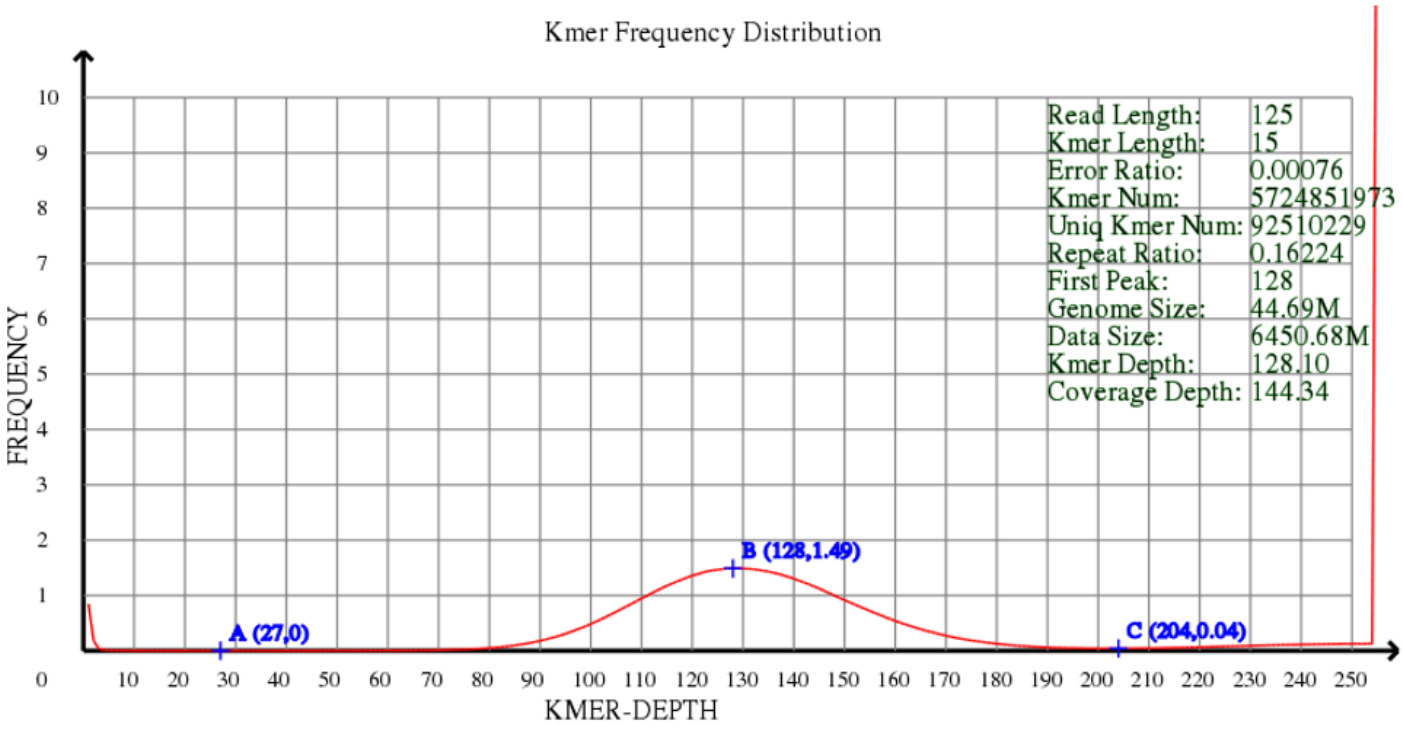

Figure 3

15-Kmer analysis. X and Y coordinates depict the depths and proportions, respectively. Regardless of any sequencing error and genomic heterozygosity or duplication, 15-mer distribution should follow the Poisson distribution. However, low-depth k-mer constitutes a high proportion due to sequencing errors. Heterozygosity may generate another peak at the 1/2 of the x-coordinate of the main peak, and duplication may cause repeating peaks near the integer times of the x-coordinate of the main peak.

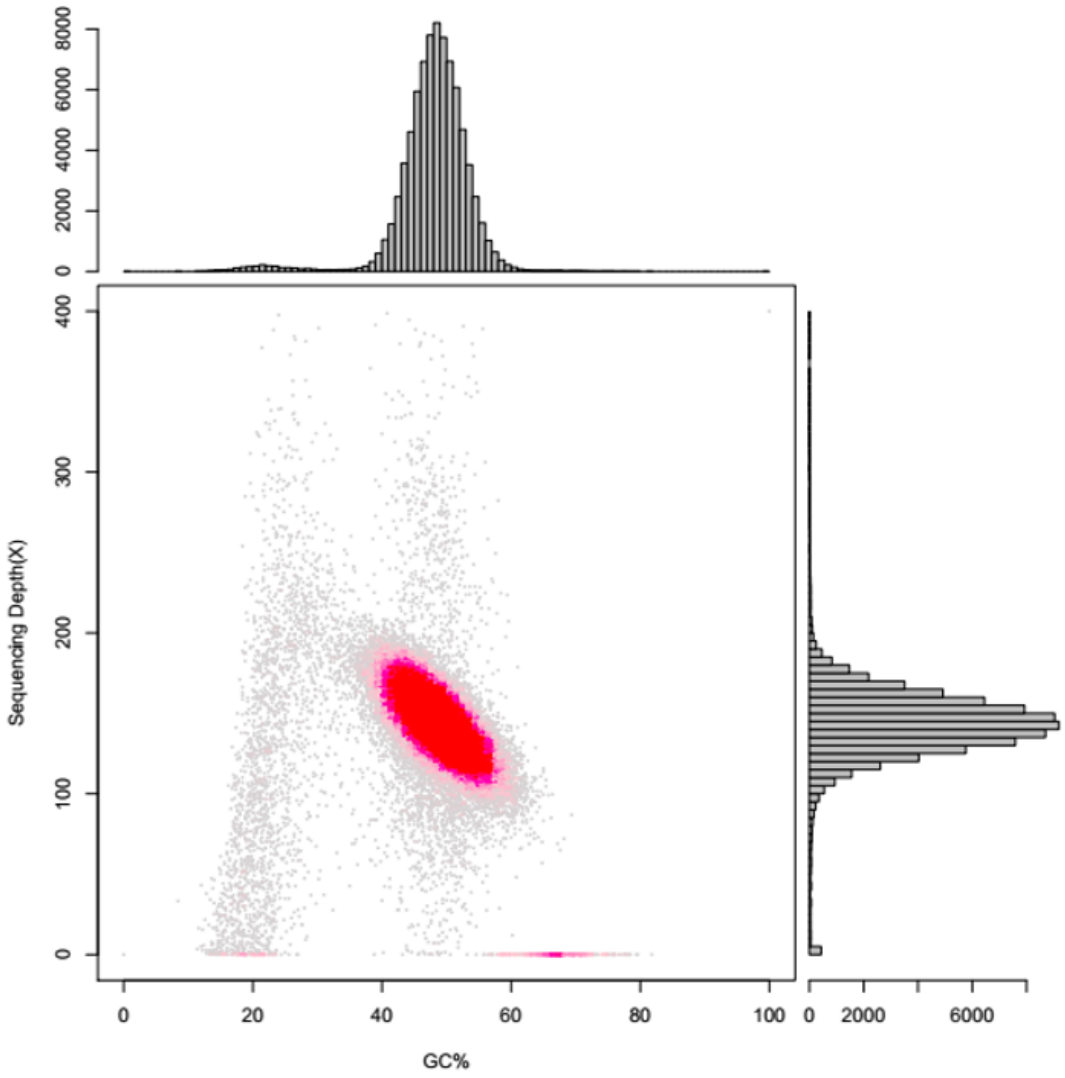

Figure 4

GC content and depth correlative analysis. X and Y coordinates show the GC content and average depth, respectively.

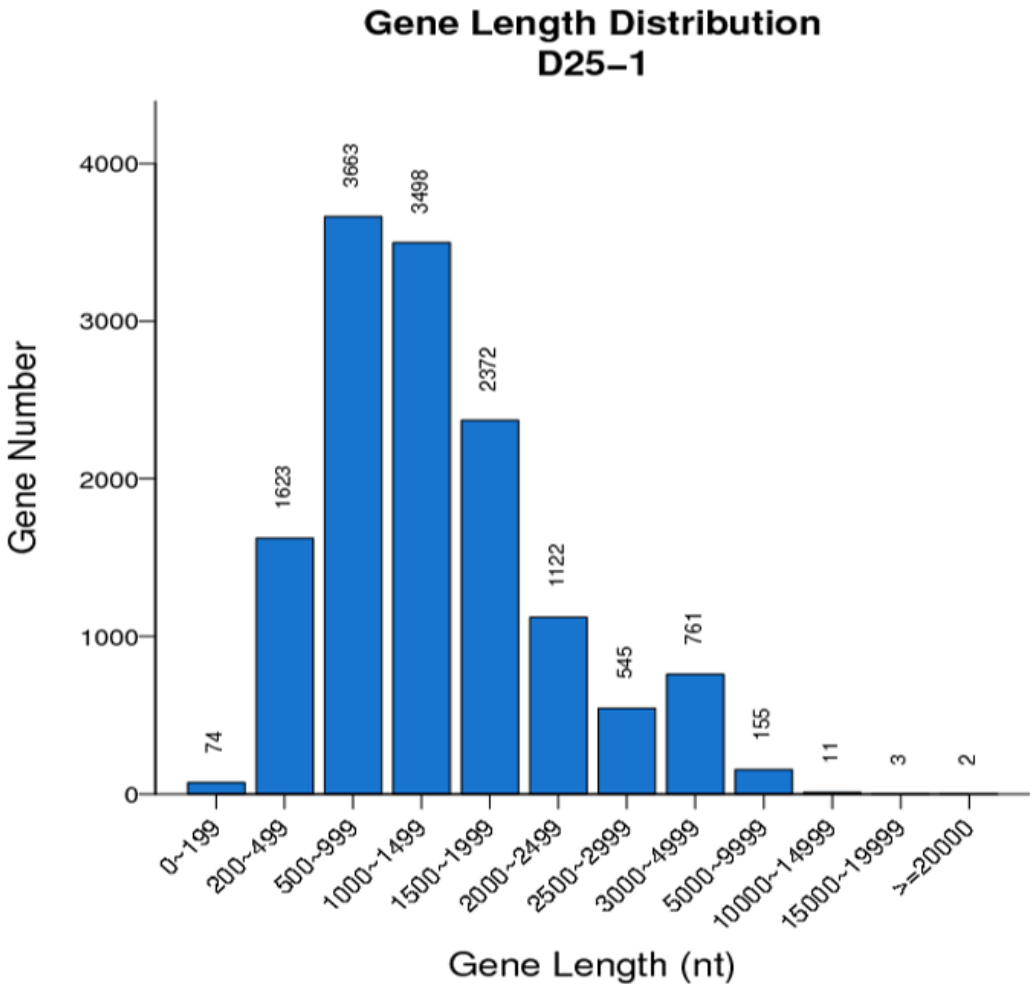

Figure 5

Gene length distribution. X and Y coordinates show the gene length and number, respectively.

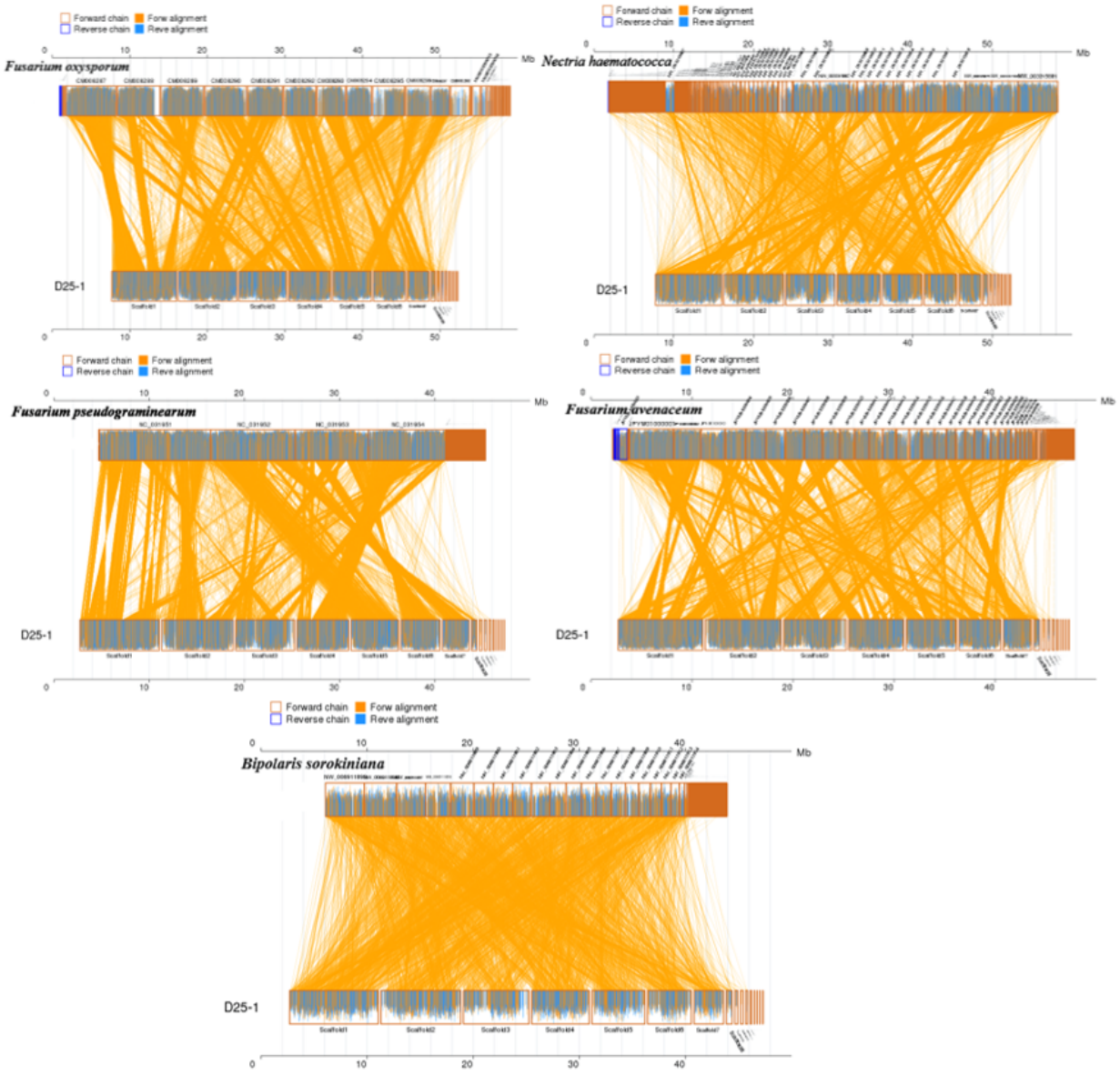

Figure 6

Synteny of Fusarium equiseti D25-1 with Fusarium oxysporum, Nectria haematococca, Fusarium pseudograminearum, Fusarium avenaceum, and Bipolaris sorokiniana. X and Y axes show the sequences of the target and reference genomes, respectively. The lighter parallel and vertical lines represent the split of each Scaffold, and the red line represents the corresponding positions of the genes with the best alignment result between two genomes.

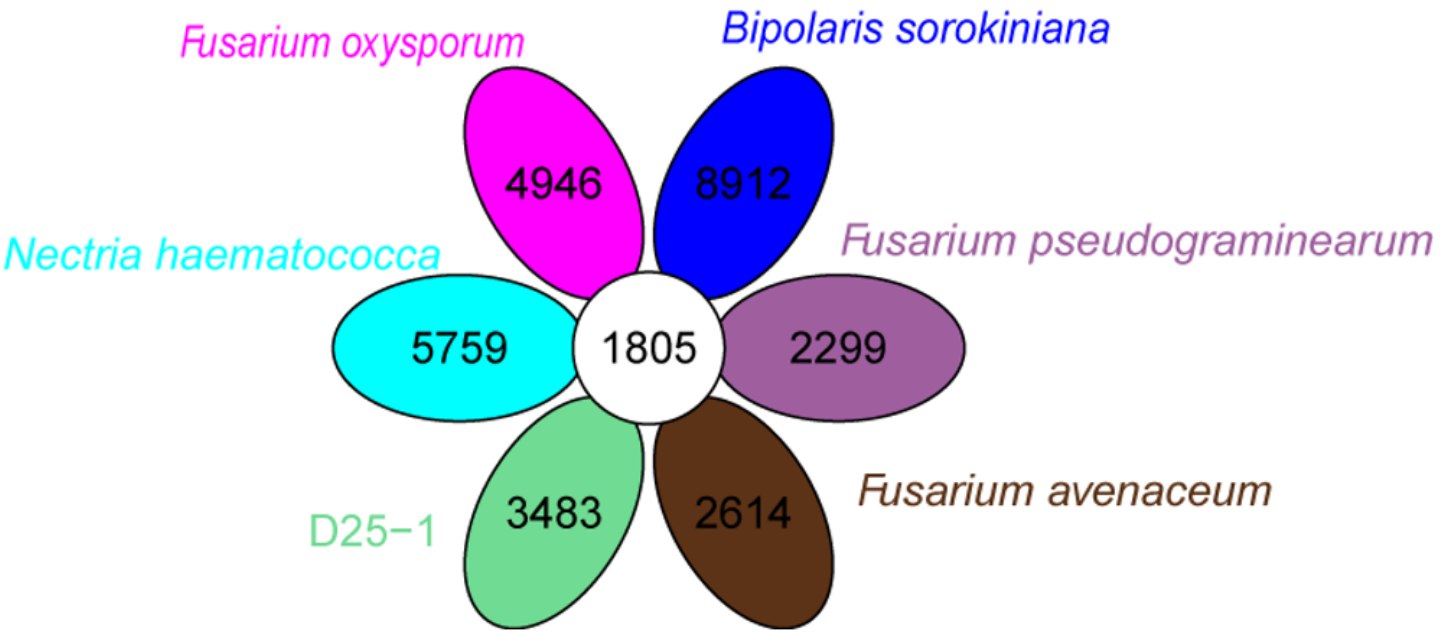

Figure 7

Core and specific gene counts. The number in the white center circle is the number of the genes common to all the six strains, and each number in the other circles represents the number of species-specific genes of the strain labeled with the same color as the corresponding circle.

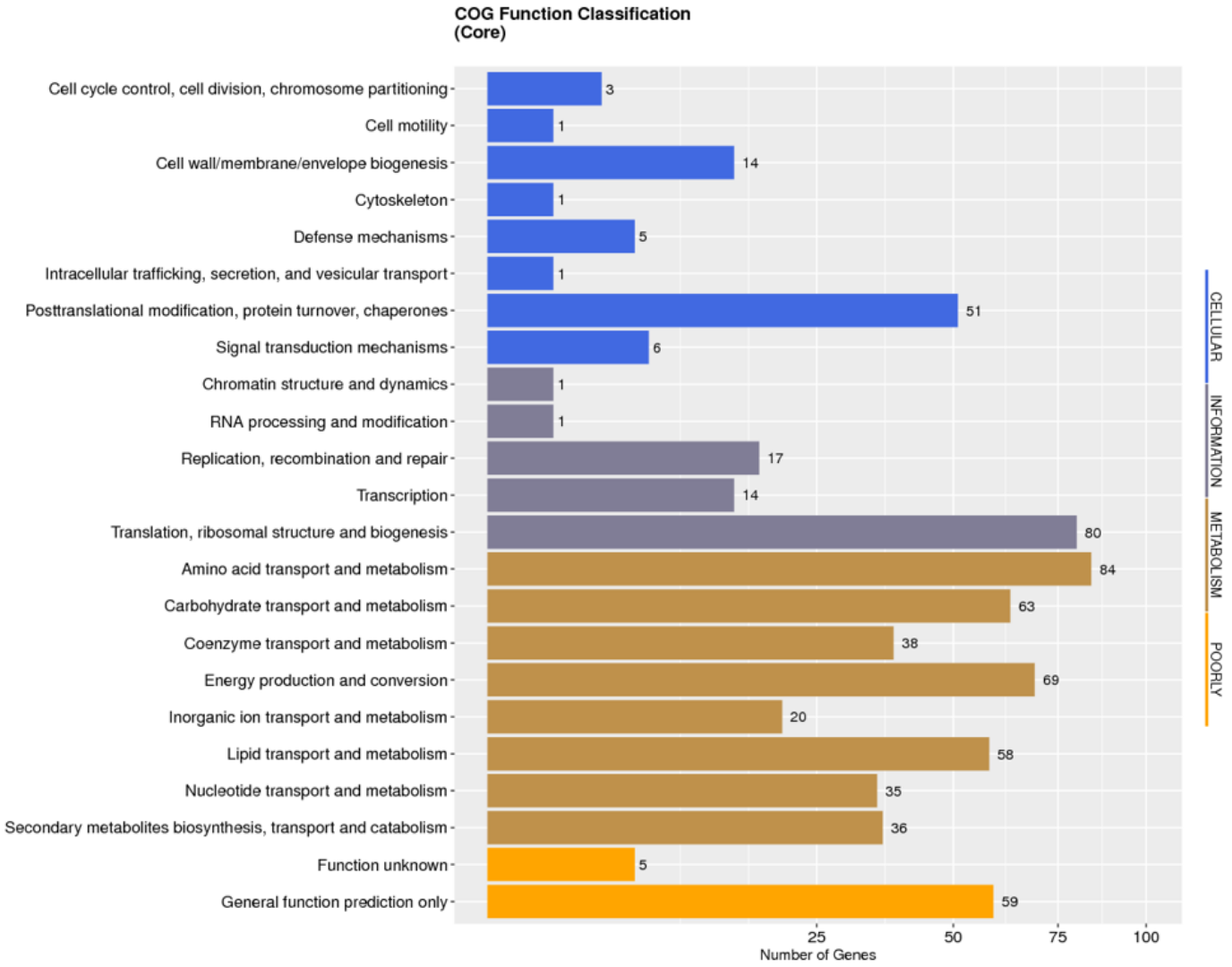

Figure 8

Annotation of core genes by COG. The ordinate and abscissa show the annotation type and corresponding gene number, respectively.

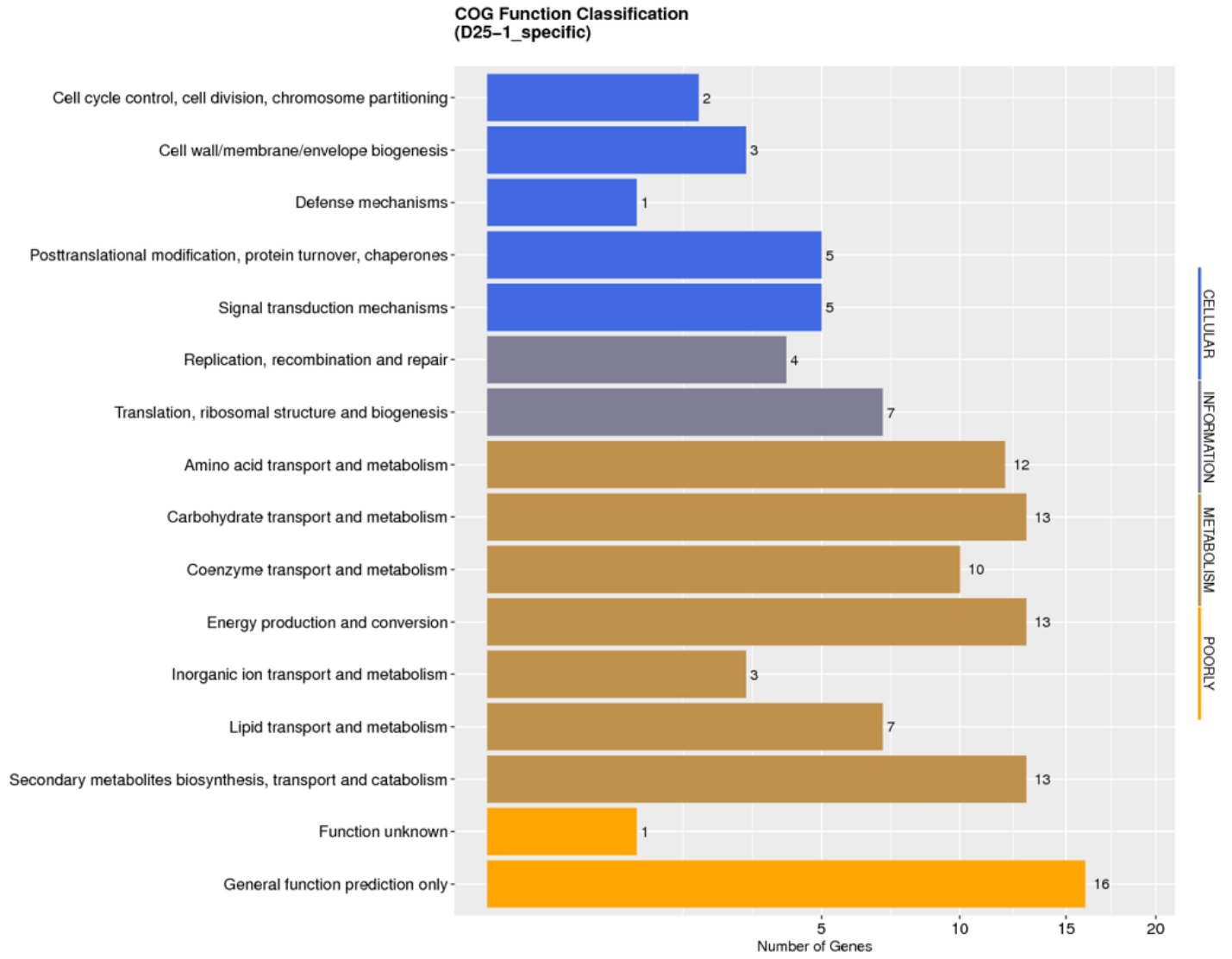

Figure 9

Annotation of species-specific genes by COG. The ordinate and abscissa show the annotation type and corresponding gene number, respectively.

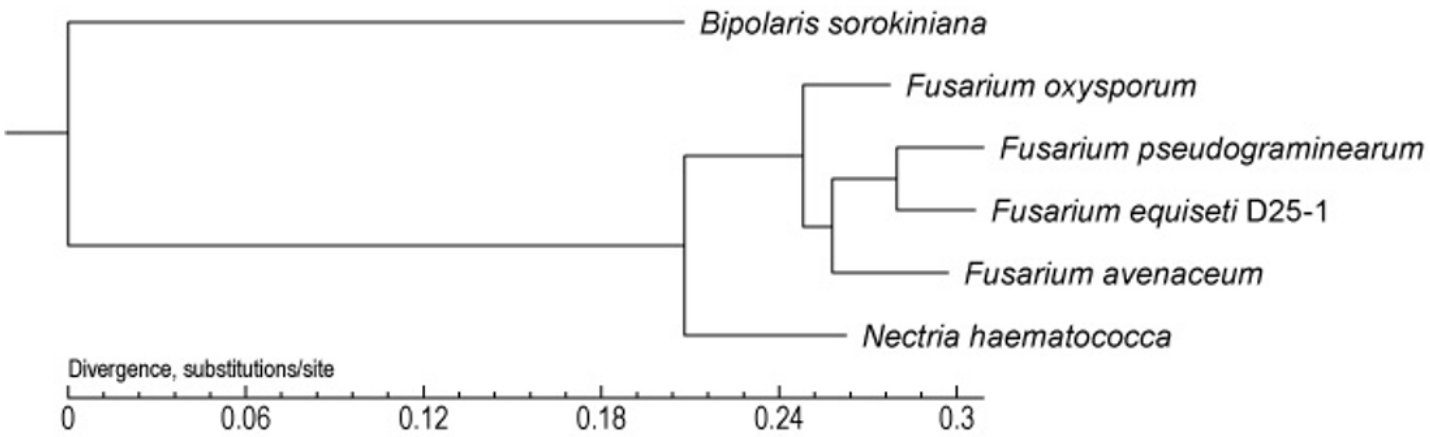

Figure 10

Phylogenetic tree of the strains and references. Core-Pan phylogenetic tree analysis of the strains Bipolaris sorokiniana, Fusarium oxysporum, Fusarium avenaceum, Fusarium pseudograminearum, Nectria haematococca, and Fusarium equiseti D25-1.

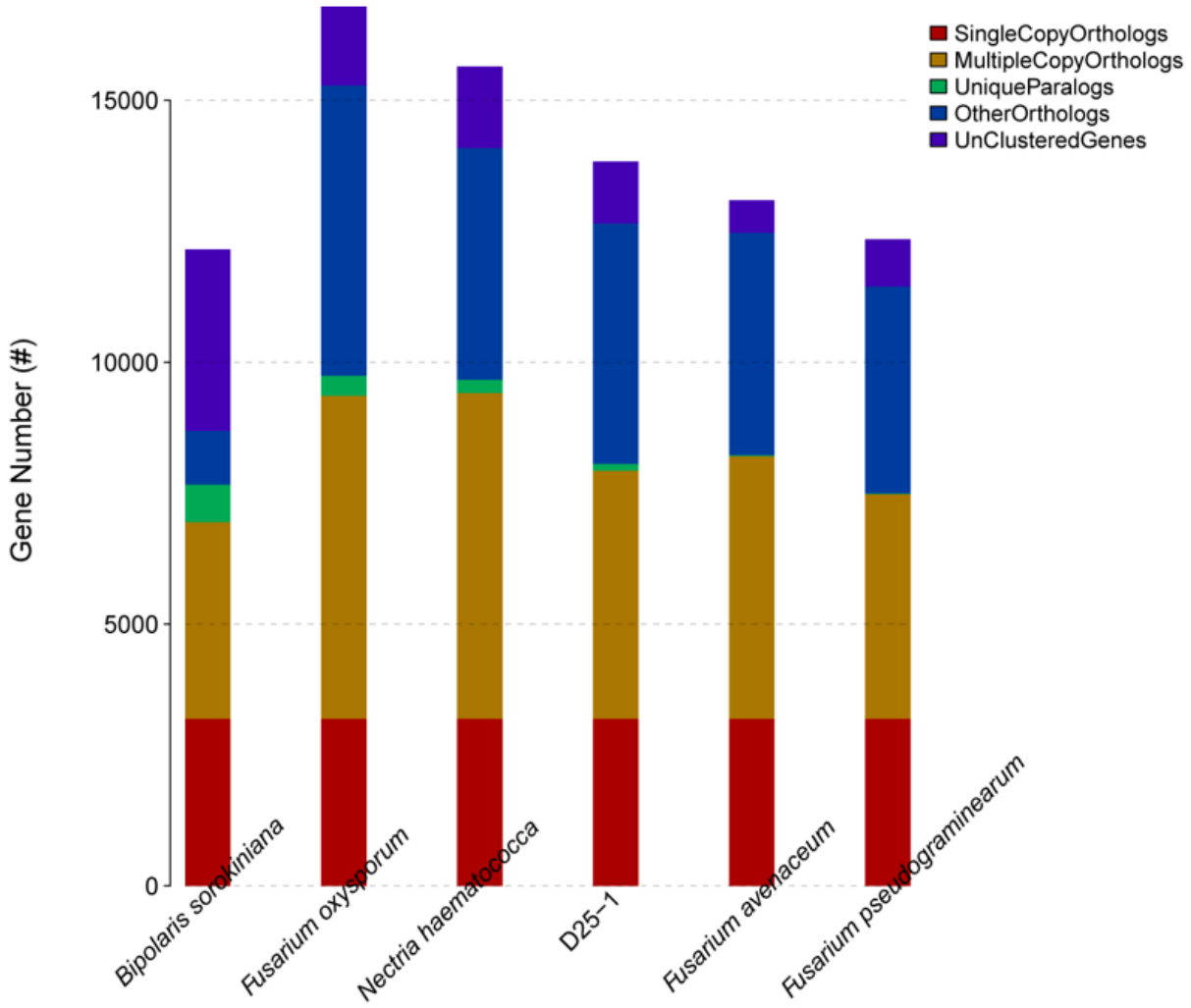

Figure 11

Gene family statistics. The abscissa and different colors represent the different strains and the number of different types of genes, respectively.

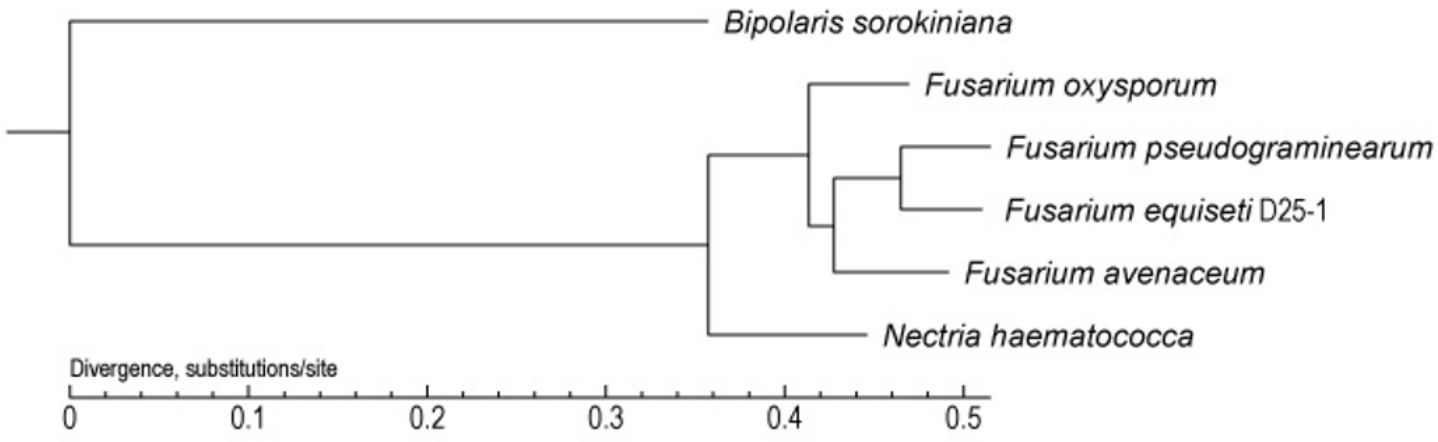

Figure 12

Phylogenetic tree of the test strains and references. Gene Family phylogenetic tree analysis of the strains Bipolaris sorokiniana, Fusarium oxysporum, Fusarium avenaceum, Fusarium pseudograminearum, Nectria haematococca, and Fusarium equiseti D25-1.