De novo genome assembly and functional annotation for *Fusarium langsethiae*

Ya Zuo¹, Carol Verheeecke-Vaessen², Corentin Molitor¹, Angel Medina², Naresh Magan² and Fady Mohareb^{1*}

¹ The Bioinformatics Group, Cranfield Soil and Agrifood Institute, School of Water, Energy and Environment, Cranfield University, College Road, Bedford, MK43 0AL, UK

² Applied Mycology Group, Environment and AgriFood Theme, Cranfield University, College Road, Bedford, MK43 0AL, UK

*To whom correspondence should be addressed.

Abstract

Motivation: *Fusarium langsethiae* is a T-2 and HT-2 mycotoxins producing *Fusarium* species firstly characterised in 2004. It is commonly isolated from oats in Northern Europe. T-2 and HT-2 mycotoxins exhibit immunological and haematological effects in animal health mainly through inhibition of protein, RNA and DNA synthesis. The development of a high-quality and comprehensively annotated assembly for this species is therefore essential in providing the molecular understanding and the mechanism of T-2 and HT-2 biosynthesis in *F. langsethiae* to help develop effective control strategies.

Results: The *F. langsethiae* assembly was produced using PacBio long reads, which were then assembled independently using Canu, SMARTdenovo and Flye; producing a genome assembly total length of 59Mb and N50 of 3.51Mb. A total of 19,336 coding genes were identified using RNA-Seq informed *ab-initio* gene prediction. Finally, predicting genes were annotated using the basic local alignment search tool (BLAST) against the NCBI non-redundant (NR) genome database and protein hits were annotated using InterProScan. Genes with blast hits were functionally annotated with Gene Ontology.

Contact: f.mohareb@cranfield.ac.uk

Data availability: Raw sequence reads and assembled genome can be downloaded from: GenBank under the accession JAFFKB000000000

1 Introduction

Fusarium langsethiae is a fungus belonging to the family *Nectriaceae*. It commonly infects ripening oats without showing any visible symptoms especially in Northern Europe, and contaminates the grains with the type A trichothecenes, T-2 and HT-2 (Torp and Nirenberg, 2004; Kalantari, 2010). These mycotoxins mainly inhibit proteins, RNA and DNA synthesis leading to immunological and haematological effects (CONTAM, 2011).

Originally, this species was considered to be a “powdery *F. poae*”. However, subsequently it was shown to be a separate species and classified as *F. langsethiae* in 2004 by Torp and Nirenberg (Torp and Nirenberg, 2004). More than 300 *Fusarium* species exist (Summerell, 2019) with different ecophysiological responses to parameters such as temperature and water availability. *F. langsethiae* is a relatively slow colonizer of temperate cereals although the ability to produce mycotoxins may provide some competitiveness in the phyllosphere microbiome (Medina and Magan, 2010, 2011; Kokkonen, 2012; Verheeecke, 2020).

Some other more economically important *Fusarium* species, e.g. *F. graminearum*, have received more attention as they produce visible symptoms in wheat (*Fusarium* head blight) and have thus been widely

sequenced and analyzed previously (King, 2015). However, many of these studies have focused on the toxin biosynthesis process (Rep, 2010; Mehrej, 2011). Some of these studies use sequencing methods to analyse protein structure or toxin-related biosynthetic gene clusters. However, most of the *Fusarium* sequences comes from short reads platforms, which may have thus excluded some important functional gene clusters and their proteins. The aim of this work was to produce a high quality, deep coverage genome assembly for *F. langsethiae* through a long-reads assembly strategy using the PacBio platform. Further analysis including identification of the clusters related to biosynthesis of T-2 and HT-2 mycotoxins was also examined. The relatively small length of fungal genomes makes them ideal candidates for long-reads assembly, since it is possible to achieve deep read coverage, allowing a higher degree of continuity with no noticeable implication on the sequencing costs.

2 Methods

2.1 Sequencing Data

To obtain long sequence fungal DNA, the protocol from Bacha (2015) was used. Briefly, 3-day-old colonies of *F. langsethiae* Fe2391 grown on

Potatoes Dextrose Agar (PDA) were harvested and frozen in liquid nitrogen. The mycelia were incubated for 10 min at 50°C in a modified lysis buffer (1% of hexadecyl-trimethyl-ammonium bromide, 100mM pH8 EDTA, 1.4 M NaCl, 20mM pH8 Tris-HCl). The DNA was then extracted 3 times in phenol:chloroform prior to precipitation in isopropanol. The pellet was resuspended with 25U of RNase prior to sending for sequencing. The DNA size (~20kb) was validated by gel electrophoresis. The samples were sent to Novogene, China, which generated raw sequencing data using the PacBio® Sequel platform.

2.2 Genome Assembly

In order to achieve the best assembly quality possible, three separate assemblers were used to process the raw *F. langsethiae* data, namely Canu v1.8, SMARTdenovo, and Flye v2.4.2. Canu (Koren, 2017) has three phases in its pipeline: correction, trimming and assembly. Since Canu is sensitive to the sequences' genome size, it requires a parameter called 'error rate', which refers to the percentage of difference between the two reads in an alignment. The genome size parameter was set to 37,500,000 for this study, based on the length of the assembly publicly available for this species. According to Canu's user guide, the parameters of error rate should be adjusted according to the coverage and data type of the raw data. Therefore, the error rates were set to 0.045, 0.055, 0.065, 0.075, 0.080, 0.085 and 0.095; this kept the correction, trimming and assembling stages identical. The second assembler used was SMARTdenovo (available at <https://github.com/ruanjue/smartdenovo>). This tool directly utilises reads from raw read alignments without correction or trimming phases, and it provides its own polishing methods to generate accurate consensus sequences. For the present study, the raw data were directly processed by the SMARTdenovo.pl script with all parameters set to default. The last, and the best performing assembler of the three was Flye (Kolmogorov, 2019), previously called A-Bruijn. It uses a repeat graph as its core data structure and utilises raw data in FASTA or FASTQ format from PacBio®. Flye outputs polished contigs with an error rate less than 30% by default. To run the assembly, the genome size was set to 37,500,000 and other parameters were set to default.

A total of five draft assemblies were generated using all three assemblers. QUAST, a quality assessment tool for genome assemblies (Gurevich, 2013), was used to examine the basic quality among the contig assemblies by comparing their total length, longest contig and N50 number. The best assembly output was identified according to completeness through orthologs comparison versus the *Saccharomyceta* OrthoDB (Kriventseva et al. 2019) data set using the Benchmarking Universal Single-Copy Orthologs (BUSCO) tool (Robert, 2017). BUSCO results were judged based on the number of complete BUSCO genes, and then the best assembly was used in the next stage.

To further improve the assembly, polishing was performed using Pilon (Walker, 2014). Since short reads were absent for this sample, the WGS sequence reads of sample FI201059 (Lysoe, 2016) were downloaded from the European Nucleotide Archive and aligned to our assembly using BWA-MEM (Li, 2009). This resulted in an alignment file in BAM format, which was then used by Pilon to perform error correction.

2.3 Gene prediction and functional annotation

Two gene prediction methods, GeneID (Blanco, 2002) and Augustus (Stanke, 2008), were used with four separate procedures, GeneID *ab-initio*, Augustus *ab-initio*, Augustus with hints and Augustus with training. For *ab-initio*, Augustus predicted genes using *F. graminearum*, and GeneID predicted genes using *F. oxysporum*. Only Augustus had settings that could be used for hints prediction and training. The hints were creat-

ed from the cDNA file from FI201059. For training, the BRAKER pipeline (Gremme, 2013) with the protein coding file from FI201059 was used; this pipeline contains an automatic training and prediction pathway using GenomeThreader and Augustus.

A basic local alignment search tool (BLAST; Boratyn, 2019) was used to find regions of local similarity between sequences. Gene sequences that had been extracted from prediction tools were output in a GFF format and then made into a FASTA file. Using the blastx command, the FASTA file was compared with the NR nucleotide database. The number of threads was set to 50. The output format of this command was set as BLAST archive (ASN.1). Functional annotation was performed using OmicsBox (available at <https://www.biobam.com/>). The BLAST hits were imported into OmicsBox to perform Gene Ontology mapping and annotation. InterPro protein signatures and domain hits were obtained using InterProScan5. The output was then imported in OmicsBox and merged with the GO annotation and mapping results.

3 Results

Since this species had not been previously assembled using long-reads, different assemblers were compared to achieve a higher quality overall assembly. After comparing each assembly's statistical information and BUSCO results, the best assembly was polished as the final assembly. Assembly quality metrics for all five chosen are shown in Table 1, which outlines some basic statistical data from the different assemblies. The total assembly lengths obtained from Canu and Flye were relatively closer to each other compared to SMARTdenovo, which had the longest genome length and the largest number of contigs, indicating fragmentation. Most of the GC content results in this study were around 48.4%; the SMARTdenovo assembly was slightly lower. The highest N50 value was achieved by Flye, which was nearly five times longer than the best Canu assembly and 22 times longer than the SMARTdenovo assembly.

Assembly method	Canu			SMART denovo	Flye
Corrected error rate	0.045	0.065	0.085	-	-
Contigs	301	301	280	277,164	177
Contigs (>=50kb)	174	174	168	170	60
Total contigs length	62,453k	62,451k	62,870k	3,850,300k	59,663k
Total contigs length (>=50kb)	59,418k	59,416k	59,802k	9,340k	59,004k
Longest contig	2,495k	2,495k	2,723k	77k	11,601k
GC content (%)	48.46	48.48	48.44	48.21	48.43
N50	512k	521k	614k	16k	3,513k

Table 1: Basic statistic information of draft assembly from Canu, SMARTdenovo and Flye

De novo genome assembly for *Fusarium langsethiae*

The draft assembly from Flye achieved 98.3% completeness, as assessed by BUSCO, and had fewer fragmented and missing hits compared with Canu (See Table 2). On the other hand, the SMARTdenovo assembly only had 11 complete hits, probably due to the fragmented state of this assembly. Considering the contigs length, N50 and BUSCO results, Flye output was considered the best assembly which was then carried forward for error correction.

Assembler	Canu	SMARTdenovo	Flye
Corrected error rate	0.045	0.065	0.085
complete hits (%)	90.8	90.9	91.7
Complete hits no.	3,384	3,387	3,417
Complete single copy	3,361	3,364	3,393
Complete duplicated	23	23	24
Fragmented hits	151	150	148
Missing hits	190	188	160

Table 2: BUSCOs statistic result from Canu, SMARTdenovo and Flye

In order to improve the assembly quality further, error correction was performed with Pilon based on the previously mentioned publicly available Illumina short reads. Comparing the statistical sequence information and BUSCO results between draft assembly and polished assembly, the length of contigs did not change after hybrid polishing, but the BUSCO results improved from 98.3% to 98.8% (See Table 3). Moreover, 28 of the fragmented genes became either complete (20) or missing (8), which support the fact that Pilon actually removed mis-assemblies from the raw draft. Compared with the previous assembly of another *F. langsethiae* strain (F1201059), the draft assembly of this study has better contiguity and slightly higher BUSCO results.

To finish the polishing, 23 contigs were detected as mitochondrial hits and were removed from the assembly.

Draft assembly	Previous assembly (Lysoe, 2016)	Flye Before polishing	Flye After polishing
Contigs	1,586	177	154
Total length	37,543,021	59,662,685	59,637,819
Largest contig	829,859	11,601,651	11,601,651
N50	86,515	3,513,144	3,513,144
Complete BUSCO hits (%)	98.2	98.3	98.8
Complete	3659	3659	3679
Complete single copy	3,647	3,642	3,661
Complete duplicated	12	17	18
Fragmented	24	48	20

Missing	42	18	26
---------	----	----	----

Table 3: Basic sequence statistic information and BUSCO result comparison between before polishing and after polishing

A series of gene prediction approaches were followed using different settings or alignment models from related species as shown in Table 4. A hints file was created for Augustus using the cDNA sequence of sample F1201059 downloaded from EMBL-EBI (Lysoe, 2016). Firstly, the cDNA contigs were aligned to the assembled genome in order to confirm its suitability. A total of 15,280 alignments were obtained, representing 96% mapping results which confirms its suitability to guide the gene prediction process. GeneID in ab-initio mode predicted 21,848 genes, while Augustus predicted a total 16,900 genes. Then, after training, Augustus predicted 19,336 genes.

Tools	GeneID	Augustus
Setting	Ab Initio	With hints
Reference species	<i>F. oxysporum</i>	<i>F. langsethiae</i> (hints), <i>F. graminearum</i> (model)
Predicted genes	21,848	19,336

Table 4: Number of predicted genes from different tools, settings, and reference species

Following the gene prediction step, a BLAST search was performed of predicted coding genes against the NR database, and the hits were further annotated with GO terms and protein signatures. A total of 19,139 out of 19,336 predicted genes (98.98%) had more than five hits against the NR database. This meant that more than 99% of genes predicted by Augustus were reliable for protein analysis to some extent. Among the hits, most had a similarity percentage higher than 80%; some of them reached 100% similarity with the sequence in the NR database.

Species	Top hits
<i>Fusarium poae</i>	50,396
<i>Fusarium oxysporum</i>	33,516
<i>Fusarium oxysporum f. sp. cepae</i>	25,207
<i>Fusarium graminearum</i>	24,162
<i>Fusarium langsethiae</i>	15,412
<i>Fusarium oxysporum f. sp. cubense</i>	14,995
<i>Fusarium fujikuroi</i>	11,795
<i>Fusarium sporotrichioides</i>	11,391
<i>Fusarium oxysporum f. sp. narcissi</i>	11,259
<i>Fusarium venenatum</i>	11,110

Table 7: Top 20 blast top hits distribution among predicted genes

More than 50% of the genes in the assembly had top hits compared with *F. langsethiae* itself. Almost all of these were within the *Fusarium* genus. Table 7 shows the top 10 BLAST hits distribution, all of them coming from the *Fusarium* genus. The assembly showed a linkage between *F. langsethiae* and *F. poae*, *F. oxysporum*, *F. graminearum*, *F. sporotrichioides* therefore indicating a similarity in the metabolic pathways and/or mycotoxin production. As the *F. poae* had the highest number of hits amongst all the *Fusarium* species, it suggests a close relevance between *F. langsethiae* and *F. poae*. Indeed, it should be considered that in papers published before 2004, *F. langsethiae* was considered as “powdery *F. poae*”. It could be inferred that, *F. poae* should thus have the closest linkage to *F. langsethiae*, amongst all the *Fusarium* species. The full chart of mycotoxins and their related proteins is provided in the Supplementary Materials SF1. However, some proteins related to trichothecene and HC-Toxin are shown in Tables 8 & 9 respectively. These tables list the contigs in which each protein was located, as well as the similarity to genes in other species as found via BLAST. The ontology term and ontology ID give a basic description of the protein and its functions.

3.1 TRI genes:

Table 8 lists gene hits related to the TRI genes cluster. In the BLAST stage, 33 sequences were found. Most of them were gathered at the thirteenth contig of our assembly, while some were gathered at the fifty-third and fifty-seventh contigs amongst others.

Contig	Position	Description of protein activity	Gene
contig_13	4474108-4475891	3-acetyltrichothecene 15-o-acetyltransferase	TRI3
	4482462-4483646	aChain A, Trichodiene Synthase	TRI5
	4496454-4498287	cytochrome P450 monooxygenase	TRI1
	6073702-6075258	ddtp-glucose 4,6-dehydratase	TRI13
	4476651-447839	P450 monooxygenase	TRI4
	6196131-619846	protein phosphatase 2 (formerly 2A)	TRI4
	4479511-4480165	regulatory protein	TRI6
	4484262-4485602	regulatory protein	TRI10
	4498937-4500113	TRI14	TRI14
	5087905-5088817	TRI5-2 protein	TRI5-2
	4488333-4488662	TRI9	TRI9
	4471969-4473348	trichothecene-4-O-acetyltransferase	TRI7
	4491419-4493141	trichothecene c-15 hydroxylase	TRI11
	4469927-4471271	trichothecene c-3 deacetylase	TRI8
	4493646-4495682	trichothecene efflux pump	TRI12
	contig_53	7949550-7951359	c6 transcription factor
413927-415700		cytochrome p450 monooxygenas	TRI1

	8051646-8052534	related to TRI15-putative transcription factor	2146
	7113258-7114437	related to TRI7-trichothecene biosynthesis gene cluster	FIE12Z_415
	428032-429514	trichothecene C-8 acyl transferase	Tri16
contig_3	2237847-223899	AChain A, Trichodiene Synthase	Tri5
	582527-584300	cytochrome p450 monooxygenase	Tri1
	2006343-2010297	related to TRI13-cytochrome P450	Focb16_v015293
contig_57	4993092-4994992	CRAL-TRIO domain-containing protein C3H8.02	FVEG_12149
	5234170-5235550	richothecene 3-o-acetyltransferase	Tri101
contig_6	30398-32815	cytochrome p450	Tri13
	contig_3	582527-584300	cytochrome p450 monooxygenase
	2006343-2010297	related to TRI13-cytochrome P450	abd-A
contig_12	674307-676300	cytochrome P450 monooxygenase	Tri1
	contig_74	539307-539832	cytochrome P450 monooxygenase
	563537-564065	related to TRI13-cytochrome P450	abd-A
contig_37	225032-226812	related to TRI13-cytochrome P450	BFJ72_g11013
	contig_92	972537-973648	related to TRI15-putative transcription factor

Table 8: Gene position, for the TRI gene cluster

3.2 HC-toxin related genes

Five proteins were found to be related to HC-toxin; all of them are listed in Table 9. Unlike the previous assembly, we identified two copies of HC-toxin synthetase located on the same contig and only 25 bases apart. Three proteins acted as an HC-toxin efflux carrier TOXA,

Contig	Position	Description	Gene
contig_11	265115-271850	hc-toxin synthetase (non-ribosomal peptide synthetase)	KPA36315
	271875-272769	hc-toxin synthetase (non-ribosomal peptide synthetase)	KPA36315
contig_13	989078-990512	hc-toxin efflux carrier	Efflux pump roqT
contig_34	770120-771587	hc-toxin efflux carrier	Efflux pump roqT
contig_93	1020288-1023864	hc-toxin efflux carrier	Efflux pump roqT

Table 9: Gene position, similarity gene and description of genes related with the keyword “hc-toxin” in assembly annotation

3.3 Related secondary metabolite genes

In addition, some contigs were linked to specific genes which are involved in global secondary metabolite biosynthesis. These include aldehyde reductase member 3 (Contig 13) and an efflux pump (Contig 2, 13, 57), ketose reductase (Contig 6, 13).

De novo genome assembly for *Fusarium langsethiae*

4 Discussions

With third-generation sequencing, the long reads and high depth generated using the PacBio® SMRT sequencing led to a very high-quality assembly. The quality was not determined according to the contig length or N50 alone; BUSCO was another parameter used to examine the quality of the assembly. Compared to the publicly available assembly of FI201059 which did not show good quality in the scaffold statistical data, but it had a high BUSCO compared to the *Saccharomyceta* dataset. Our assembly had both high-quality contigs and a high BUSCO rate, which means that more coding genes could be predicted by Augustus and GeneID. This provided hints to improve the accuracy of prediction. However, although the hints file could improve the accuracy, the core model in Augustus still came from *F. graminearum*, a related species. However, training Augustus produced a model based on the *F. langsethiae* sequence file. With the absence of an aligned BAM file or gene bank structure file from EnsembleFungi FI201059, a FASTA file containing protein in the same sample was used to train the Augustus model. While the *ab initio* approach predicted genes with a model from a different species, it has less accuracy compared to the hints and training. Although the hints file improved the accuracy from *ab initio* to some extent, for some new species such as *F. langsethiae*, the accuracy from the hints file did not achieve the expectations of the analysis. Training Augustus with data that had 96% similarity with the assembly predicted more genes than *ab initio* and the hints file.

Mycotoxin production pathways

The species in the *Fusarium* family are known to produce a wide range of secondary metabolites including type A and type B trichothecenes and zearalenone. Most significant strides have been made in relation to an understanding of the role of gene clusters involved in type B trichothecenes and fumonisins, which are produced predominantly by *F. graminearum*/*F. culmorum* and *F. verticillioides*/*F. proliferatum*, respec-

produced by *F. graminearum* and *F. pseudograminearum* and *F. culmorum*. Type A compounds are generally more toxic than type B trichothecenes while the latter are usually produced in higher concentration. T-2 and HT-2 toxins are type A trichothecenes which mainly accumulate in oats and can cause immunological or haematological defects in

Fig. 1 TRI-genes cluster identified on Contig 13 of the assembly

animals and potentially humans (CONTAM, 2011).

Almost all proteins related to trichothecene biosynthesis are located on the 13th contig of the assembly, with an additional copy of *TRI5* on contig 3 (See Fig. 1).

Other proteins located in other contigs did not seem to have a core function in T-2 and HT-2 biosynthesis, and most of them encoded the transformation or production of TRI-proteins. HC-toxin synthetase was identified on contig 11 as two copies only 25 bps apart. Both copies were identical to the previous *F. langsethiae* assembly, but with 14 a.a. changes compared to the closest blast hit of the RGP60017 gene of *F. sporotrichioides* (See Fig. 2). It could be therefore inferred that contig 13 contains the main functional proteins regrouped in a cluster. However, one of the genes encoding a protein involved in HT-2 and T-2 mechanism called TRI1, were found in the contig 3, 12, 53 and 74.

Fig. 2 Blast hits of the hc-toxin gene highlighting 14 a.a. changes compared to the closest blast hit of *F. sporotrichioides*.

tively (Munkvold, 2016). In addition, the former species also produces other mycotoxins, such as Zearalenone, Fusaric acids and Moniliformin especially in temperate cereals (Garcia-Cela et al., 2018). However, legislation on contamination of temperate cereals is predominantly focused on type B trichothecenes such as deoxynivalenol and zearalenone. Based on substitutions at C-8 and other positions around the core structure, trichothecenes have been classified into four groups (types A, B, C and D) (Munkvold, 2016), *Fusarium* species only produce type A and type B. Type A, which have an ester or hydroxyl or no oxygen substitution in the position of C-8, are usually produced by *F. armeniacum*, *F. langsethiae*, *F. poae*, *F. sambucinum*, *F. sporotrichioides*, and *F. venenatum*. Type B, which have a carbonyl group at C-8, are mainly

In 2011, McCormick described trichothecenes biosynthesis in the *Fusarium* species. Based on these findings, the biosynthesis process in *F. langsethiae* with different proteins could be inferred with the gene BLAST results and descriptions of proteins in *F. langsethiae* based on previous studies of other *Fusarium* species and the previous draft genome of *F. langsethiae* (Lysoe, 2016).

The first step consists in the cyclisation of farnesyl pyrophosphate, which is a primary metabolic intermediate (McCormick, 2011) and is mediated by a trichodiene synthase encoded in the gene *TRI5* in the 13th contig as well as another hit identified on contig 3, suggesting its presence in two copies. *TRI5* is the core gene that mediates the biosynthesis

of different trichothecenes (gi | 136010 | sp | P13513.1 | TRI5_FUSSP), including T-2 toxin. The ontology term of this gene (GO:0045482) indicates its molecular function as a trichodiene synthase. Tri5 is involved in the catalysis of the following reaction: 2-trans, 6-trans-farnesyl diphosphate = diphosphate + trichodiene. The *TRI5* gene was first characterised in a *F. sporotrichioides* strain that produced T-2 toxin (Hohn, 1986).

The trichodiene then goes through an oxygenation series catalysed by cytochrome P450 monooxygenase encoded by *TRI4* (McCormick, 2011). The *TRI4* gene (gi | 927758023 | gb | KPA41245.1) encodes a monooxygenase molecular function (GO:0004497) leading to the addition of four oxygens at C-2, C-3, C-11 and C-12, using C-13-epoxide to form the intermediate isotrichotriol (McCormick, 2006).

Subsequently, isotrichodermol (C-3-OH) is converted to isotrichodermin (C-3-OR) via an acetyltransferase encoded by *TRI101* (gi | 927756670 | gb | KPA40029.1) in the fifty-seventh contig (McCormick, 1999). The toxicity of *Fusarium* trichothecenes should be effectively reduced with this step, which serves as a mechanism for the fungal self-protection other trichothecene-producing organisms (Kimura, 1998). TRI101 (gene located in contig 57) acts as part of the transferase activity (GO:0016747) that transfers an acyl group, other than aminoacyl, from one compound to another. Then, a second hydroxyl group is added to C-15, which is controlled by TRI11 in the contig 13 (Alexander, 1998). TRI11 (gi | 927758018 | gb | KPA41240.1) works with a molecular function (GO:0016705)—an oxidation-reduction reaction in which hydrogen or electrons are transferred from each of two donors as well as an oxidation-reduction process (GO:0055114). After this TRI3 (gi | 927758024 | gb | KPA 41246.1) (GO:0043386) catalyses the acetylation of the 4-hydroxyl to form trichodermin and then, TRI13 protein (gi | 927758016 | gb | KPA41238.1) perform the same oxidation-reduction reaction as TRI11 in C-4 (Lee, 2002), followed by another acetylation process by TRI7 (gi | 927758025 | gb | KPA41247.1).

The next step of this process in *F. sporotrichioides* is the addition of a fourth hydroxyl group to C-8 by TRI1, followed by an addition of an isovaleryl moiety thanks to TRI16. Finally, the C-3 position loses the acetyl group via a TRI8-esterase step to produce T-2 toxin (McCormick, 2002).

In this study, the *TRI1* gene was found in the 3rd and 53rd contigs, and there was about 15% sequence dissimilarity with the *TRI1* sequence (gi | 927755786 | gb | KPA39264.1) in the database. Since *TRI16* was not found, it might have been either mis-matched and mis-labeled by BLAST or missing from the assembly. The gene *TRI8* was also found (gi | 927758026 | gb | KPA41248.1) and was identified as encoding a triglyceride lipase activity (GO:0004806) and is involved in a reaction in which triacylglycerol + H₂O = diacylglycerol + a carboxylate.

Most of the genes linked to T-2 mycotoxin production were found in this assembly, but the *TRI9* gene only had sequences for which no literature could be found to describe their function. However, *TRI9* has been found not only in *F. langsethiae* but also other *Fusarium* species, such as *F. sporotrichioides* and *F. graminearum*. It acts upon the integral component of the membrane (GO:0016021) and might be linked with the T-2 mycotoxin transport mechanism.

With regard to the genes identified in relation with the HC-toxin, not enough information was found to support the production pathways in *F. langsethiae*. However, some genes encoded global regulators such as the efflux pump and carrier of HC-toxin produced by *Cochliobolus* species might have an evolving relationship with *F. langsethiae*.

Availability

This Whole Genome Shotgun project has been deposited at GenBank under the accession JAFFKB000000000 under BioProject PRJNA701381. The version described in this paper is version JAFFKB010000000.

Funding

This research is supported by a BBSRC-SFI research grant (BB/P001432/1) between Cranfield University, UK and the University College Dublin, Ireland.

Conflict of Interest: none declared.

References

- Abildgren, M.P., Lund, F., Thrane, U., Elmholt, S., 1987. Czapek–Dox agar containing iprodione and dichloran as a selective medium for the isolation of *Fusarium* species. *Letters in Applied Microbiology* 5, 83–86
- Adam Phillippy, Sergey Koren, Brian Walenz Revision a95437e1, 2015, available at: <https://canu.readthedocs.io/en/latest/quick-start.html>
- Alexander N.J., Hohn T.M., McCormick S.P. The TRI11 gene of *Fusarium sporotrichioides* encodes a cytochrome P450 monooxygenase required for C-15 hydroxylation in trichothecene biosynthesis. *Appl. Environ. Microbiol.* 1998; 64:221–225.
- Alexey Gurevich, Vladislav Saveliev, Nikolay Vyahhi and Glenn Tesler, QUASt: quality assessment tool for genome assemblies, *Bioinformatics* (2013) 29 (8): 1072–1075. doi: 10.1093/bioinformatics/btt086 First published online: February 19, 2013
- Arcella, D., Gergelova, P., Innocenti, M. and Steinkellner, H. (2017). Human and animal dietary exposure to T-2 and HT-2 toxin. *EFSA Journal*, 15(8).
- Bacha N, Echarki Z, Mathieu F, Lebrihi A, (2015) Development of a novel quantitative PCR assay as a measurement for the presence of geosmin-producing fungi *J. Appl. Microbiol.*, 118(5) 1144–1151.
- Booth, C., 1971. The genus *Fusarium*. Commonwealth Mycological Institute, Kew, Surrey, England. Eastern Press, London and Reading.
- Boratyn GM, Thierry-Mieg J, Thierry-Mieg D, Busby B, Madden T.L. (2019) “Magic-BLAST, an accurate RNA-seq aligner for long and short reads.” *BMC Bioinformatics*. 2019 Jul 25;20(1):405.
- Bruce J. Walker, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A. Cuomo, Qiangdong Zeng, Jennifer Wortman, Sarah K. Young, Ashlee M. Earl (2014) Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement. *PLoS ONE* 9(11): e112963. doi: 10.1371/journal.pone.0112963
- E. Blanco, G. Parra and R. Guigó, “Using geneid to Identify Genes.”, In A. Baxevanis, editor: *Current Protocols in Bioinformatics*. Unit 4.3. John Wiley & Sons Inc., New York (2002) (in press)
- CONTAM (2011) Scientific Opinion on the risks for animal and public health related to the presence of T-2 and HT-2 toxin in food and feed,” *EFSA J.* 9(12) 2481.
- Ensembl.org. (2019). Chromosomes, scaffolds and contigs. [online] Available at: https://www.ensembl.org/info/genome/genebuild/chromosomes_scaffolds_contigs.html [Accessed 5 Aug. 2019].
- Gardiner DM, et al , 2012, Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. *PLoS Pathog* 2012 Sep
- Gardiner DM, et al. 2014, Genome Sequence of *Fusarium graminearum* Isolate CS3005. *Genome Announc* 2014 Apr 17
- Giordano F, Aigrain L, Quail M A, et al. De novo yeast genome assemblies from MinION, PacBio and MiSeq platforms[J]. *Scientific reports*, 2017, 7(1): 3935.
- Gregory, S. Contig Assembly. *Encyclopedia of Life Sciences*, 2005.
- Gremme, G. (2013). Computational Gene Structure Prediction. PhD thesis, Universität Hamburg.
- Guides.library.yale.edu. (2019). Yale University Library Research Guides: Bioinformatics Tools for Research and Discovery at Yale University: Gene Prediction/ Annotation. [online] Available at: <https://guides.library.yale.edu/c.php?g=296274&p=1974247> [Accessed 24 Aug. 2019].

De novo genome assembly for *Fusarium langsethiae*

- Hohn T.M., VanMiddlesworth F. Purification and characterization of the sesquiterpene cyclase trichodiene synthetase from *Fusarium sporotrichioides*. Arch. Biochem. Biophys. 1986; 251:756–761. doi: 10.1016/0003-9861(86)90386-3.
- Jue Ruan, 2015, Ultra-fast de novo assembler using long noisy reads, *ruanjue/smartdenovo*. [online] Available at: <https://github.com/ruanjue/smartdenovo> [Accessed 28 Jul. 2019].
- Kalantari H, M. M., REVIEW ON T-2 TOXIN. Jundishapur Journal of Natural Pharmaceutical Products 2010, 5 (1), 26-38.
- Kimura M., Kaneko I., Komiyama M., Takatsuki A., Koshino H., Yoneyama K., Yamaguchi I. Trichothecene 3-O-acetyltransferase protects both the producing organism and transformed yeast from related mycotoxins. J. Biol. Chem. 1998; 273:1654–1661
- King R., Urban M., Hammond-Kosack M. C. U., Hassani-Pak K., and Hammond-Kosack K. E., (2015) The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*,” *BMC Genomics*, 16 (1), 1–21.
- Kokkonen M, Medina A, Magan N, (2012) Comparative study of water and temperature relations of growth and T-2/HT-2 toxin production by strains of *Fusarium sporotrichioides* and *Fusarium langsethiae*,” *World Mycotoxin J.*, 5 (4) 365–372.
- Kolmogorov, M., Yuan, J., Lin, Y. and Pevzner, P. (2018). Assembly of Long Error-Prone Reads Using Repeat Graphs.
- Koren and Walenz et al. (2017) Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. (2017).
- Koren S, Walenz BP, Berlin K, Miller JR, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*. (2017).
- Kosiak, B., Torp, M., Thrane, U., 1997. The occurrence of *Fusarium* spp. in Norwegian grain—a survey. *Cereal Research Communication* 25, 595–596.
- Kriventseva, E. V., et al. (2019), 'OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs', *Nucleic Acids Res*, 47 (D1), D807–D11
- Langseth, W., Rundberget, T., 1999. The occurrence of HT-2 toxin and other trichothecenes in Norwegian cereals. *Mycopathologia* 147, 157–165.
- Lee T., Han Y.-K., Kim K.-H., Yun S.-H., Lee Y.-W. Tri13 and Tri7 determine deoxynivalenol- and nivalenol-producing chemotypes of *Gibberella zeae*. *Appl. Environ. Microbiol.* 2002; 68:2148–2154.
- Li H. and Durbin R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25, 1754–1760. [PMID: 19451168]
- Li H.*, Handsaker B.*, Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R. and 1000 Genome Project Data Processing Subgroup (2009) The Sequence alignment/map (SAM) format and SAMtools. *Bioinformatics*, 25, 2078–9.
- Lysøe, E., Frandsen, R., Divon, H., Terzi, V., Orrù, L., Lamontanara, A., Kolseth, A., Nielsen, K. and Thrane, U. (2016). Draft genome sequence and chemical profiling of *Fusarium langsethiae*, an emerging producer of type A trichothecenes. *International Journal of Food Microbiology*, 221, pp.29–36.
- Lysøe, E., Divon, H. H., Terzi, V., Orru, L., Lamontanara, A., Kolseth, A.-K., Frandsen, R. J., Nielsen, K. and Thrane, U. 2015, The draft genome sequence of *Fusarium langsethiae*, a T-2/HT-2 mycotoxin producer, Submitted (15-APR-2015) Department of Plant Health and Plant Protection, Bioforsk - Norwegian Institute of Agricultural and Environmental Research, Høgskoleveien 7, Aas 1430, Norge
- Ma LJ, et al. 2014, Genome Sequence of *Fusarium oxysporum* f. sp. melonis Strain NRRL 26406, a Fungus Causing Wilt Disease on Melon. *Genome Announc* 2014 Jul 31
- Mario Stanke, Mark Diekhans, Robert Baertsch, David Haussler (2008), Using native and syntenically mapped cDNA alignments to improve de novo gene finding, *Bioinformatics*, doi:10.1093/bioinformatics/btn013
- McCormick S.P., Alexander N.J. *Fusarium* Tri8 encodes a trichothecene C-3 esterase. *Appl. Environ. Microbiol.* 2002; 68:2959–2964. doi: 10.1128/AEM.68.6.2959-2964.2002.
- McCormick S.P., Alexander N.J., Proctor R.H. *Fusarium* Tri4 encodes a multifunctional oxygenase required for trichothecene biosynthesis. *Can. J. Microbiol.* 2006; 52:636–642.
- McCormick S.P., Alexander N.J., Trapp S.E., Hohn T.M. Disruption of TRI101, the gene encoding trichothecene 3-O-acetyltransferase, from *Fusarium sporotrichioides*. *Appl. Environ. Microbiol.* 1999; 65:5252–5256.
- McCormick S.P., Hohn T.M., Desjardins A.E. Isolation and characterization of Tri3, a gene encoding 15-O-acetyltransferase from *Fusarium sporotrichioides*. *Appl. Environ. Microbiol.* 1996; 62:353–359.
- McCormick S.P., Taylor S.L., Plattner R.D., Beremand M.N. Bioconversion of possible T-2 toxin precursors by a mutant strain of *Fusarium sporotrichioides*. *Appl. Environ. Microbiol.* 1990; 56:702–706.
- McCormick, S., Stanley, A., Stover, N. and Alexander, N. (2011). Trichothecenes: From Simple to Complex Mycotoxins. *Toxins*, 3(7), pp.802–814.
- Merhej J, Richard-Forget F, Barreau C, (2011) Regulation of trichothecene biosynthesis in *Fusarium*: Recent advances and new insights,” *Applied Microbiol. Biotech.* 91(3) 519–528.
- Mikhail Kolmogorov, Jeffrey Yuan, Yu Lin and Pavel Pevzner, “Assembly of Long Error-Prone Reads Using Repeat Graphs”, *Nature Biotechnology*, 2019 doi:10.1038/s41587-019-0072-8
- Munkvold, G. P. (2016). *Fusarium* Species and Their Associated Mycotoxins. *Mycotoxigenic Fungi*, 51–106. doi:10.1007/978-1-4939-6707-0_4
- Myers, E. (2000). A Whole-Genome Assembly of *Drosophila*. *Science*, 287(5461), pp.2196–2204.
- Nathanail, A. V., Varga, E., Meng-Reiterer, J., Bueschl, C., Michlmayr, H., Malachova, A., ... Berthiller, F. (2015). Metabolism of the *Fusarium* Mycotoxins T-2 Toxin and HT-2 Toxin in Wheat. *Journal of agricultural and food chemistry*, 63(35), 7862–7872. doi: 10.1021/acs.jafc.5b02697
- Pacific Biosciences of California (2010 - 2014). Template Preparation and Sequencing Guide, © Copyright 2010 - 2014, Inc. All rights reserved.
- Ray Wu Faculty Profile (1970). [online] Available at: <https://web.archive.org/web/20090304121126/http://www.mbg.cornell.edu/faculty-staff/faculty/wu.cfm> [Accessed 15 Jun. 2019].
- Rep M, Kistler C H, (2010), The genomic organization of plant pathogenicity in *Fusarium* species. *Current Opinion in Plant Biol.* 13(4) 420–426.
- Robert M. Waterhouse, Mathieu Seppey, Felipe A. Simão, Mose Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov, *Mol Biol Evol*, published online Dec 6, 2017, BUSCO applications from quality assessments to gene prediction and phylogenomics. doi: 10.1093/molbev/msx319
- Robert M. Waterhouse, Mathieu Seppey, Felipe A. Simão, Mose Manni, Panagiotis Ioannidis, Guennadi Klioutchnikov, Evgenia V. Kriventseva, and Evgeny M. Zdobnov *Mol Biol Evol*, published online Dec 6, 2017 | Abstract | Full Text PDF | doi: 10.1093/molbev/msx319
- Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes CA, Hutchison CA, Slocombe PM, Smith M (February 1977). “Nucleotide sequence of bacteriophage phi X174 DNA”. *Nature*. 265 (5596): 687–95
- Sanger F, Nicklen S, Coulson AR (December 1977). “DNA sequencing with chain-terminating inhibitors”. *Proc. Natl. Acad. Sci. USA*. 74 (12): 5463–77.
- Summerell B A, (2019) Resolving *Fusarium*: Current Status of the Genus,” *Annu. Rev. Phytopathol.*, 57(1) 323–339.
- Torp, M. and Nirenberg, H. (2004). *Fusarium langsethiae* sp. nov. on cereals in Europe. *International Journal of Food Microbiology*, 95(3), pp.247–256.
- Vanheule, A., Audenaert, K., Warris, S., Van De Geest, H., Schijlen, E., Hofte, M., De Saeger, S., Haesaert, G., Waalwijk, C. and Van Der Lee, T. 2016, Living apart together: crosstalk between the core and supernumerary genomes in a fungal plant pathogen, Submitted (02-JUN-2016) *Applied Bioinformatics*,
- Verhecke-Vaessen C, Garcia-Cela E, Lopez-Pietro A, Osk Jonsdottira I, Medina A and Magan N (2020) Water and temperature relations of *Fusarium langsethiae* strains and modelling of growth and T-2 and HT-2 mycotoxin production on oat-based matrices. *Int J Food Microbiol* (In press).
- Wageningen UR, Droevendaalsesteeg 4, Wageningen, Gelderland 6708PB, Netherlands
- Yli-Mattila T, Mach R L, Alekhina I A, et al. Phylogenetic relationship of *Fusarium langsethiae* to *Fusarium poae* and *Fusarium sporotrichioides* as inferred by IGS, ITS, β -tubulin sequences and UP-PCR hybridization analysis[J]. *International Journal of Food Microbiology*, 2004, 95(3): 267–285.