A peer-reviewed version of this preprint was published in PeerJ on 19 December 2018.

View the peer-reviewed version (peerj.com/articles/5963), which is the preferred citable publication unless you specifically need to cite this preprint.

Brankovics B, Kulik T, Sawicki J, Bilska K, Zhang H, de Hoog GS, van der Lee TA, Waalwijk C, van Diepeningen AD. 2018. First steps towards mitochondrial pan-genomics: detailed analysis of *Fusarium graminearum* mitogenomes. PeerJ 6:e5963
https://doi.org/10.7717/peerj.5963

First steps towards mitochondrial pan-genomics: Detailed analysis of *Fusarium graminearum* mitogenomes

Balázs Brankovics [Corresp., 1, 2, 3], **Tomasz Kulik** [4], **Jakub Sawicki** [4], **Katarzyna Bilska** [4], **Hao Zhang** [5], **G Sybren de Hoog** [2, 3], **Theo AJ van der Lee** [1], **Cees Waalwijk** [1], **Anne D van Diepeningen** [1, 2]

[1] Wageningen Plant Research, Wageningen University & Research, Wageningen, Netherlands

[2] Westerdijk Fungal Biodiversity Institute, Utrecht, Netherlands

[3] Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands

[4] Department of Botany and Nature Protection, University of Warmia and Mazury, Olsztyn, Poland

[5] State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agriculture Sciences, Beijing, China P.R.

Corresponding Author: Balázs Brankovics
Email address: balazs.brankovics@wur.nl

There is a gradual shift from representing a species' genome by a single reference genome sequence to a pan-genome representation. Pan-genomes are the abstract representations of the genomes of all the strains that are present in the population or species. In this study, we employed a pan-genomic approach to analyze the intraspecific mitochondrial genome diversity of *Fusarium graminearum*. We present an improved reference mitochondrial genome for *F. graminearum* with an intron-exon annotation that was verified using RNA-seq data. Each of the 24 studied isolates had a distinct mitochondrial sequence. Length variation in the *F. graminearum* mitogenome was found to be largely due to variation of intron regions (99.98%). The "intronless" mitogenome length was found to be quite stable and could be informative when comparing species. The coding regions showed high conservation, while the variability of intergenic regions was highest. However, the most important variable parts are the intron regions, because they contain approximately half of the variable sites, make up more than half of the mitogenome, and show presence/absence variation. Furthermore, our analyses show that the mitogenome of *F. graminearum* is recombining, as was previously shown in *F. oxysporum*, indicating that mitogenome recombination is a common phenomenon in *Fusarium*. The majority of mitochondrial introns in *F. graminearum* belongs to group I introns, which are associated with homing endonuclease genes (HEGs). Mitochondrial introns containing HE genes may spread within populations through homing, where the endonuclease recognizes and cleaves the recognition site in the target gene. After cleavage of the "host" gene, it is replaced by the gene copy containing the intron with HEG. We propose to use introns unique to a population for tracking the spread of the given population, because introns can spread through vertical inheritance, recombination as well

as via horizontal transfer. We demonstrated how pooled sequencing of strains can be used for mining mitogenome data. The usage of pooled sequencing offers a scalable solution for population analysis and for species level comparisons studies. This study may serve as a basis for future mitochondrial genome variability studies and representations.

First steps towards mitochondrial pan-genomics: Detailed analysis of *Fusarium graminearum* mitogenomes

Balázs Brankovics[1,2,3], Tomasz Kulik[4], Jakub Sawicki[4], Katarzyna Bilska[4], Hao Zhang[5], G Sybren de Hoog[2,3], Theo AJ van der Lee[1], Cees Waalwijk[1], and Anne D van Diepeningen[1,2]

[1]**B.U. Biointeractions and Plant Health, Wageningen University and Research, Wageningen, Netherlands**
[2]**Westerdijk Fungal Biodiversity Institute, Royal Netherlands Academy of Arts and Sciences, Utrecht, Netherlands**
[3]**Institute for Biodiversity and Ecosystem Dynamics, University of Amsterdam, Amsterdam, Netherlands**
[4]**Department of Botany and Nature Protection, University of Warmia and Mazury, Olsztyn, Poland**
[5]**State Key Laboratory for Biology of Plant Diseases and Insect Pests, Institute of Plant Protection, Chinese Academy of Agriculture Sciences, Beijing, China P.R.**

Corresponding author:
Balázs Brankovics[1,2,3]

Email address: balazs.brankovics@wur.nl

ABSTRACT

There is a gradual shift from representing a species' genome by a single reference genome sequence to a pan-genome representation. Pan-genomes are the abstract representations of the genomes of all the strains that are present in the population or species. In this study, we employed a pan-genomic approach to analyze the intraspecific mitochondrial genome diversity of *Fusarium graminearum*. We present an improved reference mitochondrial genome for *F. graminearum* with an intron-exon annotation that was verified using RNA-seq data. Each of the 24 studied isolates had a distinct mitochondrial sequence. Length variation in the *F. graminearum* mitogenome was found to be largely due to variation of intron regions (99.98%). The "intronless" mitogenome length was found to be quite stable and could be informative when comparing species. The coding regions showed high conservation, while the variability of intergenic regions was highest. However, the most important variable parts are the intron regions, because they contain approximately half of the variable sites, make up more than half of the mitogenome, and show presence/absence variation. Furthermore, our analyses show that the mitogenome of *F. graminearum* is recombining, as was previously shown in *F. oxysporum*, indicating that mitogenome recombination is a common phenomenon in *Fusarium*. The majority of mitochondrial introns in *F. graminearum* belongs to group I introns, which are associated with homing endonuclease genes (HEGs). Mitochondrial introns containing HE genes may spread within populations through homing, where the endonuclease recognizes and cleaves the recognition site in the target gene. After cleavage of the "host" gene, it is replaced by the gene copy containing the intron with HEG. We propose to use introns unique to a population for tracking the spread of the given population, because introns can spread through vertical inheritance, recombination as well as via horizontal transfer. We demonstrated how pooled sequencing of strains can be used for mining mitogenome data. The usage of pooled sequencing offers a scalable solution for population analysis and for species level comparisons studies. This study may serve as a basis for future mitochondrial genome variability studies and representations.

INTRODUCTION

One of the most ideal markers for monitoring the distribution and spread of populations is the mitochondrial genome (Harrison, 1989). Due to its high copy number within individual cells, the mitochondrial genome is easy to access. Furthermore, it is mostly maternally inherited and it is less likely to recombine than the nuclear genome. In fungi gender is not genetically determined, and since maternal structures and meiosis require resources, the better adapted genotype is more likely to act as the maternal strain. This means that the mitochondrial genotype has the potential to be used to track the successful nuclear genotypes.

Mitochondrial sequences have been used for resolving phylogenetic and evolutionary relationships between fungi at all taxonomic levels (Liu et al., 2009; Avila-Adame et al., 2006; Fourie et al., 2013). In 2003, the DNA barcoding initiative started, aiming at using a single marker for taxon identification. The marker that was selected was a mitochondrial gene, cytochrome c oxidase I – COI or *cox1* (Hebert et al., 2003). In *Fusarium* however, the use of *cox1* was abandoned as a barcoding region, because of the frequent presence of introns in the gene made this region impractical for PCR amplification (Gilmore et al., 2009). Next generation sequencing (NGS) and new analysis methods have resolved this issue (Brankovics et al., 2016).

Fusarium graminearum is the major causative agent of Fusarium head blight (FHB), a devastating disease of small grain cereals. Besides reducing yield, the fungus contaminates crops with mycotoxins such as trichothecenes and zearalenone, which pose a serious threat to food and feed safety (Desjardins, 2006). Population studies of *F. graminearum* showed that the populations are highly dynamic and several displacements have been reported (Gale et al., 2007; Ward et al., 2008). Monitoring these population shifts is important, as they may differ in virulence, fungicide resistance and/or mycotoxin profile (Gale et al., 2007; Zhang et al., 2012).

The mitochondrial genome of *F. graminearum* encodes all genes typically associated with mtDNAs of filamentous fungi: two rRNA coding genes, 14 protein coding genes and a large set of tRNA coding genes (Al-Reedy et al., 2012). In addition, a large open reading frame with unknown function (LV-uORF) was found, flanked by tRNA genes. The first comparative studies of mitochondrial genomes of *Fusarium* spp. have revealed that *F. graminearum* has a significantly larger mitogenome than *Fusarium* spp. belonging to other species complexes analyzed so far (Fourie et al., 2013; Al-Reedy et al., 2012). Intron variation within the FGSC has not yet been analyzed, but the mitogenomes of different species within the *F. fujikuroi* species complex showed diversity in intron content based on the sequences of *F. circinatum*, *F. fujikuroi* and *F. verticillioides* (Fourie et al., 2013).

Most mitochondrial introns found in *Fusarium* are group I introns. These introns are self-splicing ribozymes, which frequently contain homing endonuclease genes (HEGs) (Haugen et al., 2005). The combination of intron and HEG forms a mobile element that is able to invade intronless copies of the "host" gene (Haugen et al., 2005), thereby enabling horizontal spread of the mobile element through the population. This mechanism is called homing, since the homing endonuclease recognizes a target site of 15–45 bp, which makes the insertion highly sequence specific (Haugen et al., 2005). A functional homing endonuclease is needed for the homing of the intron, but the intron may be retained as long as the self-splicing function of the intron is intact. Since the mitochondrial genes are crucial for the proper functioning of the cell, if an intron loses its ability to self-splice, then the intron is lost through precise excision (Goddard and Burt, 1999). This mechanism allows an intron to spread in populations to strains that do not possess the given intron. This dispersion does not require further recombination. The mechanism does not allow one haplotype of an intron to replace another one, since the horizontal transfer is mediated only by the cleavage of an intronless copy. Hence, the replacement of one haplotype by another one can only be explained either by recombination or by loss of the original intron and insertion of the new haplotype.

Pan-genomes are the abstract representation of the genomes of all the strains that are present in the population. The idea of pan-genome or supra-genome comes from bacterial genomics, and originated from the distributed genome hypothesis (DGH) (Ehrlich, 2001; Tettelin et al., 2005). According to the DGH, each strain within a population/species contains a subset of contingency genes from within the supra-genome (pan-genome), i.e., the supra-genome is distributed among many individual strains (Ehrlich, 2001; Ehrlich et al., 2004). Pan-genome based analysis can be used to identify conserved, variable and strain specific regions within a group of genomes. Pan-genomes can be also employed to contrast two populations or two species.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

In order to create a pan-genome for the mitogenome of *F. graminearum*, we have to better understand the nature and dynamics of the diversity in the mitochondrial genome of this organism. To accomplish this, a reliable reference has to be established as a basis for all comparative analyses. To this end, we resequenced the reference strain of *F. graminearum*, PH-1, assembled its mitochondrial genome, improved its annotation and validated the annotation using RNA-seq. Subsequently, this reference was used to study the SNP frequencies, intron distribution and sequence variability of the different regions of the mitogenome within the species, by analyzing a total of 24 strains, which were individually sequenced, representing a wide range of hosts and geographic origins. Finally, we evaluated the efficacy of using pooled sequencing in assessing the mitogenome sequence diversity within a sample. Pooled sequencing offers the possibility of analyzing populations directly from field samples.

MATERIALS & METHODS

Strains

Thirteen *F. graminearum* strains were sequenced individually on the Illumina Miseq platform (Table 1). In addition, *F. graminearum* strain PH-1 (CBS 123657, NRRL 31084) was sequenced on the Illumina Hiseq platform both as a single strain and as part of pooled set of five *F. graminearum* strains (Table 1). Besides the newly sequenced strains, the whole genome sequencing reads of ten *F. graminearum* were downloaded from the SRA database of NCBI that were produced by other research groups (Laurent et al., 2017; Wang et al., 2017). The outgroup, *F. gerlachii* strain was sequenced for an earlier publication (Kulik et al., 2016). A detailed description of the fungal strains is given in Table 1.

Sequencing

Illumina Miseq

Whole genome libraries were prepared using the Nextera XT kit (Illumina, San Diego, CA, USA) from gDNA extracted from mycelium. The constructed libraries were sequenced on the Illumina Miseq platform with 250 bp paired-end read, version 2. The fungal genomes were sequenced in a multiplexed format (6-7 samples per run), where an oligonucleotide index barcode was embedded within adapter sequences that were ligated to DNA fragments (Smith et al., 2010). Next, the sequence reads were de-multiplexed and filtered for low quality base calls, trimming all bases from 5' and 3' read ends with Phred scores <Q30.

Illumina Hiseq

For *F. graminearum* strain PH-1 (CBS 123657, NRRL 31084) a random sheared shotgun library was prepared using the NEXTflex ChIP-seq Library prep kit with adaptations for low input gDNA according to manufacturer's protocol (Bioscientific). The library was loaded as (part of) one lane of an Illumina paired-end flowcell for cluster generation using a cBot. Sequencing was done on an Illumina HiSeq2000 instrument using 101, 7, 101 flow cycles for forward, index and reverse reads respectively. De-multiplexing of resulting data was carried out using the Casava 1.8 software. Sequencing reads have been uploaded to the European Nucleotide Archive (ENA) with the accession number PRJEB18592.

The same method was applied for the pooled sequencing with the adjustment that random sheared shotgun library was prepared by using equal amounts of genomic DNA extract from all five strains (Table 1). Sequencing reads have been uploaded to the European Nucleotide Archive (ENA) with the accession number PRJEB18596.

Third party sequencing data

Besides the sequencing data that we have generated, we also made use of sequencing data produced by other research groups that had been submitted to SRA (Sequencing Read Archive) databases. This included a dataset of SRA data of 6 strains isolated from France (PRJNA295638; Laurent et al., 2017), 3 strains from China (PRJNA296400; Wang et al., 2017) and one strain from Australia (PRJNA235346; Gardiner et al., 2014). The mitochondrial genome sequences for the strains sequenced by third party are available in the Third Party Annotation Section of the DDBJ/ENA/GenBank databases under the accession numbers TPA: BK010538-BK010547

Assembly

GRAbB was used with SPAdes assembler to reconstruct the mitogenome of the strains. GRAbB (Brankovics et al., 2016) was chosen because it is a wrapper program for iterative *de novo* assembly based on a reference sequence. SPAdes 3.8.1 (Bankevich et al., 2012; Nurk et al., 2013) assembler was used, since

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

it offers good insight for the user into the relationship between nodes in the assembly graph and the relationship between nodes, contigs and scaffolds. The mitochondrial genomes were assembled from NGS reads using GRAbB by specifying the mitogenome sequence of PH-1 strain (HG970331) as query sequence.

For each individually sequenced strain it was possible to resolve the assembly to a single circular sequence. When the GRAbB run finished for the strains that were pooled for sequencing, the final assembly graph was visualized using Bandage (Wick et al., 2015) and the assembly was resolved to two circular sequence variants to capture all the variation within the dataset (Supplementary Text 1). For the first circular sequence, referred to as "short", the shorter alternative contigs were included in the path at each position where continuity was ambiguous. While for the other sequence, referred to as "long", the longer alternatives were included. In this way, all the different sequence regions were represented at least once in the two sequences.

Sequence annotation

The initial mitogenome annotations were done using MFannot (`http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl`) and were manually improved: annotation of tRNA genes was improved using tRNAscan-SE (Pavesi et al., 1994), annotation of protein-coding genes and the *rnl* gene was corrected by aligning intronless homologs to the genome. Intron encoded proteins were identified using NCBI's ORF Finder (`http://www.ncbi.nlm.nih.gov/gorf/gorf.html`) and annotated using InterPro (Mitchell et al., 2015) and CD-Search (Marchler-Bauer and Bryant, 2004). The annotated mitochondrial genome sequences are available under the following GenBank accession numbers: BK010538-BK010547, KP966550-KP966561, KR011238 and MH412632.

Read mapping and SNP discovery

The mitogenome of *F. graminearum* strain PH-1 and the two mitogenome sequences obtained from the assembly of the pooled dataset were used as reference sequences for the read mapping and SNP discovery. The read mapping was done using *aln* and *sampe* subcommands of the Burrows-Wheeler Alignment tool (BWA-0.7.12-r1034) (Li and Durbin, 2009). SNP calling was done using SAMtools mpileup (1.3.1) with *-g* and *-f* flag and BCFtools call (1.3.1) with *-mv* flag (Li et al., 2009).

Coverage analysis

Coverage of different regions was estimated by, first, mapping reads of the pooled dataset to the reference sequence using the *sampe* subcommand of the Burrows-Wheeler Alignment tool (BWA-0.7.12-r1034) (Li and Durbin, 2009). Then, read coverage was calculated using the *genomecov* command of bedtools v2.26.0. The following single copy nuclear protein coding genes were used to represent single copy nuclear regions: γ-actin (*act*), β-tubulin II (*tub2*), calmodulin (*cal*), 60S ribosomal protein L10 (*rpl10a*), the second largest subunit of DNA-dependent RNA polymerase II (*rpb2*), translation elongation factor 1α (*tef1a*), translation elongation factor 3 (*tef3*) and topoisomerase I (*top1*). The reference sequences were extracted from the genome of PH-1 (4 chromosomes: HG970332, HG970333, HG970334, and HG970335). The nuclear mitochondrial DNA segment (NUMT) used for coverage comparison was identified during the assembly of the pooled data (see Supplementary Text 1).

Intron validation

The RNA-seq data for *F. graminearum* PH-1 was downloaded from NCBI's SRA database, accession number PRJNA239711 (Zhao et al., 2014). Read mapping was done by *subjunc* command of the Subread aligner (Liao et al., 2013). Intron positions were identified from the CIGAR string of the SAM file produced by the aligner.

Linear model

R was used for linear model analysis to test whether the intron variation is the main reason of mitochondrial genome length variation within the species. The linear model was the following:

$$y = x + c$$

where y was the total length of the mitochondrial genome, x was the length of the intron sequences and c was the y-intercept (average intronless length of the mitochondrial genomes). The R^2 value obtained

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

from linear model analysis specifies what percentage of the variation of the dependent value (mitogenome length) is explained by the variation in the independent value (intron length).

$$R^2 = 1 - \frac{SS_{residual}}{SS_{total}}$$

Residual sums of squares ($SS_{residual}$) and total sums of squares (SS_{total}) were calculated using the *deviance* function of R.

Comparative sequence analysis

The nucleotide sequences were aligned using MUSCLE (Edgar, 2004b,a). Sequence variability of given regions was calculated by aligning the sequences. Then the number of characters with multiple character states was calculated and divided by the total number of characters in the alignment. This step was done using fasta_variability from the fasta_tools package (`https://github.com/b-brankovics/fasta_tools`).

Phylogenetic analysis

The most appropriate substitution evolution model was determined using jModelTest 2 (Darriba et al., 2012) for each of the regions analyzed. Phylogenetic trees were calculated using RAxML version 8.2.4 (Stamatakis, 2014). Two measures of clade support were used in this study: i) bootstrap (BS) values calculated by 1000 bootstrap runs using RAxML and ii) Bayesian posterior probability (BPP). In order to obtain BPP values, phylogenetic reconstruction has been conducted using MrBayes 3.2.5 (Ronquist et al., 2012). The MCMC algorithm was run for 4,000,000 generations with four incrementally-heated chains, starting from random trees and sampling one out every 1000 generations. Burn-in was set to relative burn-in of 0.25. The generated tree-space was used to calculate the BPP.

Detecting the presence of recombination

The intergenic regions were analyzed using the Φ_w-test implemented in SplitsTree (Bruen et al., 2006) to detect whether there is recombination in the mitochondrial genome.

RESULTS

Mitochondrial genome of *F. graminearum*

The mitochondrial genomes of all 24 strains sequenced individually were assembled into single circular contigs. The re-sequencing of the mitochondrial genome of *F. graminearum* strain PH-1 revealed two SNPs compared to the most recent published mitogenome assembly (HG970331.1) of the strain that was based on next generation sequencing reads (King et al., 2015). The correction of these SNPs was supported by the fact that all the other strains contained the same two SNPs obtained in the new assembly of PH-1. The newly assembled mitochondrial genome of the PH-1 strain as well as the other mitochondrial genomes were annotated. The mitochondrial genomes of all strains contained the same set of genes in the same order and orientation (Fig. 1). To test whether the intron-exon models were predicted correctly, RNA-seq reads were mapped against the mitogenome of *F. graminearum* strain PH-1. The results of the read mapping supported all of the predicted intron-exon boundaries.

Mitogenome variability in *F. graminearum*

The mitogenomes of *F. graminearum* strains analyzed showed variation in size, ranging from 93,560 bp to 101,424 bp (Table 2). To test whether intron variation is the main reason of mitochondrial genome length variation within the species, linear model analysis was used. The linear model that assumed that mitochondrial length variation is due only to variation of the length of intron regions explained 99.98% of intraspecific length variation observed in the data, showing that intron variation is the main reason behind intraspecific mitochondrial genome length variation. The standard deviation of the mitogenome length was 1818 bp, which is 1.87% of the average mitochondrial genome length.

The coding regions (tRNA, rRNA and conserved protein coding genes) showed low levels of variation both within *F. graminearum* (0.02%) and when compared to *F. gerlachii* (0.02%). In addition, none of the SNPs found in protein coding regions caused amino acid substitution.

The large ORF with unknown function (LV-uORF) located in the large variable region of the mitogenome contained five SNPs within *F. graminearum* and the sequence in the *F. gerlachii* strain was

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

identical to the most frequent haplotype within *F. graminearum*. All five SNPs resulted in amino acid substitution in the putative peptide sequences. The variability of the conserved protein coding regions was 0.02%, while the variability of the LV-uORF region was 0.09% within *F. graminearum*. The difference in variability was even more striking on the protein sequence level, where the conserved protein genes showed no variation, while the LV-uORF showed 0.26% variability.

The variability of the intergenic regions was 1.63% and 2.30% for intraspecies and interspecies, respectively. The overall sequence variability of intron sequences was 0.68% and 0.71% for intraspecies and interspecies, respectively. Although the variability of intron regions was significantly less than that of intergenic regions, both regions contained approximately equal numbers of variable sites (Table 3) due to the large length difference between the two regions. The intron regions were the most variable part of the mitochondrial genomes, because approximately half of the variable sites was located in introns, and introns were the only regions showing presence/absence variation within *F. graminearum*.

Interestingly, strains CBS 128539 and CBS 138561 had identical intergenic sequences, while strains CBS 104.09 and CBS 119800 (isolated 81 years apart) had identical intron sequences. However, all the strains had a unique mitochondrial genome sequence.

Intron patterns and phylogeny

A total of 39 intron sites were found in the individually sequenced dataset (Supplementary Table). Out of the 39 introns, 32 were present in all strains and 21 of these showed no variation at the intraspecies level and 14 at the interspecies level. The introns that showed presence/absence variation within the dataset were cob-i159, cob-i201, cox1-i1287, cox2-i228, cox2-i318, cox2-i552 and nad2-i1632 (Fig. 2 and Supplementary Table). The intron names contain the gene name where they are located and the coding nucleotide position of the host gene after which they were inserted.

It was not possible to group the strains based on their intron patterns (presence/absence for each intron) without allowing for multiple gain or loss of introns (Supplementary Table). This could be the result of recombination of the mitochondrial genome or the horizontal transfer of introns. Recombination would affect intergenic regions, while the horizontal transfer of the intron by homing would not affect the intergenic regions. Recombination of the intergenic regions was well supported ($p = 2.26 * 10^{-6}$) by the Φ_w-test.

Strategies to analyze pooled mitochondrial NGS data

Two approaches were used to explore the mitogenome variability in the pooled dataset: i) assembling the reads *de novo* and ii) mapping the reads to a reference sequence.

De novo assembly approach

The assembly resulted in a graph that contained five ambiguous sites that represented four insertion/deletion variations (three intron presence/absence variation cob-i201, cox1-i1287, cox2-i318, and a large insertion inside the cob-i490 intron) in the dataset, and one site (located in nad4L-i239) where two different alleles were found in the strain set (Supplementary Text 1). These polymorphic sites were too far apart to establish linkage between them, so two alternative assemblies were extracted from the assembly graph: one with the shorter allele at all of the positions and one with the longer allele at all of the positions (Supplementary Text 1). The assembly method did not reveal SNP variations, only intron presence/absence variations and one replacement variation.

Mapping approach

To assess the influence of the reference sequence on the mapping and SNP calling results, both of the sequences obtained from the assembly approach of the pooled dataset were used as references, beside the curated mitogenome of the PH-1 strain. Besides giving an insight into the influence of the reference sequence to the downstream analysis, this also makes it possible to detect variation within intron sequences that may be absent in some of the reference sequences.

The lowest coverage detected for a single nucleotide allele was 21% of the reads that mapped to the given position. This is close to the expected value (20%) for an allele present in a single strain in a pool of five strains. This result shows that the method was sensitive enough to detect a SNP present in a single strain. Furthermore, the results of all three analyses identified the same polymorphic sites. This means that the choice of reference sequence did not influence the SNP detection results.

The three runs of read mapping and SNP calling revealed a total of fifteen SNPs (Table 4). The allele ratios were identical even when the reference sequence used for the mapping was different, with one

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

exception: position 90636. At this position both PH-1 and the pooled assembly analysis showed 70% for the nucleotide present in the given reference and 30% for the alternative, despite the fact that the two references had different nucleotides at the given location (Table 4). Examination of the alignment of the reference sequences revealed that the sequence difference was not only a single nucleotide polymorphism at position 90636, but there was a 8 bp long indel at position 90627-90634. This nearby indel influenced the mapping of reads containing the allele differing from the reference sequence. This was the reason why the SNP calling skewed in favor of the reference allele in both mappings.

Coverage analysis

Coverage values were calculated for different genomic regions in order to determine whether coverage cutoffs could be used to differentiate between mitochondrial and nuclear regions. The coverage of single copy nuclear regions that were present in all of the pooled strains was 290x. The coverage of the nuclear mitochondrial DNA segment (NUMT) sequence was 230x, which suggests that it was present in four of the five pooled strains. The coverage of mitogenome regions that were present in all strains was 4000x. While, the coverage of singleton mitochondrial regions, present only in a single strain, was 475x. The coverage gap was sufficiently high between shared single copy nuclear regions (290x) and singleton mitochondrial sequence (475x) to allow clear differentiation between them.

DISCUSSION

Comparative genomics analyses are traditionally reference (Laurent et al., 2017) or pairwise based (Fourie et al., 2013). Reference based methods are efficient at identifying regions that are present in the reference, but absent in other individuals, or detecting smaller variations, like SNPs. This method does not identify regions that are absent from the single reference, while these regions might be valuable for clustering the non-reference individuals. Pairwise comparison is able to identify unique regions for both individuals; however, it is difficult to scale to a larger sample size, because every individual has to be compared to every other individual, then the results have to be brought to the same scale.

To take full advantage of next generation sequencing data, a paradigm shift is needed: from focusing on a single reference genome to using a pan-genome, that is, a representation of all genomic content in a certain population, species or phylogenetic clade (Computational Pan-Genomics Consortium, 2018). In this study, we used an *ad hoc* pan-genomic analysis of the mitochondrial genomes of *Fusarium graminearum*. The reason for using an *ad hoc* approach is that pan-genomics is still a young field of research, and as such, there are no clear standards developed yet for analysis, for files or for data sharing. The goal of the analysis was to understand the nature and the dynamics of mitogenome variability, then to identify the implications of these results for mitogenome based population studies or track & trace implementations. The results of this study can be utilized for the development of suitable data structures and file formats for capturing the variability of mitochondrial pan-genomes.

In this study, we improved the mitochondrial genome reference for *F. graminearum* strain PH-1, which is recognized as the reference strain of this species for genomic studies (Al-Reedy et al., 2012; King et al., 2015; Cuomo et al., 2007). The first mitochondrial genome sequence was produced using Sanger sequencing and primer walking by Al-Reedy et al. (2012). The assembly was improved by King et al. (2015) using NGS reads. This assembly corrected 15 SNPs and 30 indels in the sequence. Here, we present a new assembly, which corrected 2 more SNPs, complete with a detailed annotation. The introns that were predicted during the annotation process were all verified by RNA-seq data.

The mitochondrial genomes of *F. graminearum* and *F. gerlachii* contained the same genes and ORFs in the same orientation. The coding sequences showed high levels of conservations, and all SNPs found in protein coding genes were synonymous substitution. The genetic variation in the mitochondrial genome could be classified in two groups: small sequence variations (SNPs and short indels) and intron gain and loss. Although, variations resulting from SNPs and short indels were twice as frequent in intergenic regions as in intron regions, about half of the variable sites was located in intron regions. The second type of variation, the presence/absence of introns, accounted for 99.98% of the length variation between the mitochondrial genomes. In conclusion, the majority of the sequence variation within the species was related to intron regions: either SNPs and short indels or the presence/absence of complete introns. Thus, in mitogenome comparative analysis or pan-genomic studies, special attention should be given to accurately capturing the intron variation, since it is the most informative fraction of the mitogenome.

An alternative way to sequencing strains individually is sequencing them in a pool. The pooled

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

sequencing approach is more cost efficient than sequencing the strains separately. The data produced by pooled sequencing of strains from a given population could be viewed as the pan-genomic sequencing reads of that population. In this study, we have demonstrated how sequencing data from pools of strains can be mined for mitochondrial genome variation. Sequencing in pools has already been used to discover rare alleles of nuclear loci (Raineri et al., 2012). This method can be used for finding rare alleles, but it also allows a scalable solution for analyzing complete populations. So far, the application of pooled sequencing data has been used for SNP discovery in nuclear loci from samples (Raineri et al., 2012). However, analyzing mitochondrial genome data of fungi possesses some additional challenges. We have demonstrated that besides SNPs, intron presence/absence variation is a major element of the mitogenome variation. To assess what kind of approach can detect intron presence/absence variation and SNP variation, we analyzed the data using a *de novo* assembly approach followed by a read mapping and SNP-calling approach. The results show that the assembly approach is able to identify sequence differences affecting sequence regions longer than individual sequencing reads, such as, insertions and deletions of intron sequence or long polymorphic sequences, while it is unable to identify SNPs or short indels. Read mapping and SNP calling analysis has to be performed to identify SNPs. This method in turn is unable to identify sequence differences affecting longer sequence regions. For optimal results, a sequential approach is needed for analyzing pooled samples: first, an assembly step to identify introns or larger indels absent from the reference genome, then using both the reference and the newly identified extra regions for read mapping and SNP-calling.

The drawbacks of pooled data are that short indel variation might be missed and linkage between markers is lost when using short read sequencing technologies, although linkage information is not crucial when comparing pan-genomes with each other. Furthermore, pooling large amount of strains could mean the loss of the coverage gap between mitochondrial copies and nuclear copies, which makes NGS analysis of mitochondrial genomes more advantageous to PCR methods. This means that nuclear mitochondrial sequences (NUMTs) might affect the results. With sufficient caution the effects of NUMTs can be minimized, since they can be identified in the assembly step. In the assembly step, NUMTs may appear as separate contigs, as in our example, or as new paths similar to introns with the significant difference that intron segments are joined to the rest of the mitochondrial assembly on both termini, while the flanking nuclear regions of NUMTs would only be joined on one of the termini of the segment. Despite these concerns, the benefits of pooled sequencing of large numbers of strains offers a scalable solution for population or species level comparisons. After a reference sequence is established, each population or species could then be represented by pools of multiple strains.

Most of the introns in *F. graminearum* are group I introns, and contain homing endonuclease genes (HEGs). Group I introns harboring a functional HEG can spread in a population through homing. Homing is facilitated by the homing endonuclease that cleaves the target gene at a 15-45 bp recognition site. The resulting double strand break stimulates homologous recombination based DNA repair. Since all copies of the mitochondrial genome that contain the recognition site are susceptible to the homing endonuclease, the only viable template for homologous repair is a genome that contains a copy of the intron. The insertion of the intron into the recognition site modifies the sequence, and it will no longer be recognized by the homing endonuclease.

The mitochondrial genome of *F. graminearum* shows evidence of recombination. We recently showed that mitochondrial recombination does also happen in the *F. oxysporum* species complex (Brankovics et al., 2017). Recombination of the mitochondrial genome in *Fusarium* appears to be a common phenomenon, since both *F. oxysporum* and *F. graminearum* show signs of mitochondrial recombinations, despite the fact that *F. oxysporum* is an asexual fungus with a putative parasexual cycle, while *F. graminearum* is a homothallic species that has an active sexual cycle (Yun et al., 2000). Due to the recombination of the mitogenome, it cannot directly be used as a marker to track successful nuclear genotypes as was proposed. However, based on the spreading mechanism of introns, introns could be used for track and trace implementations. The intron sequences spread through clonal & sexual reproductions, and through horizontal transfer. Due to the effect of the homing endonuclease, all offspring of a sexual cross would be tagged by all the introns that are specific to either parent. The appearance of a new intron in a population could be a sign of migration or gene flow.

The annotation of strain CBS 119173 revealed a putative nested intron in cox1-i906. All other strains contain a haplotype that is 1006 bp long, while this strain contains a haplotype that is 2084 bp long. The sequence comparison indicates that the additional 1078 bp region is an intron that was integrated inside

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

the homing endonuclease of the acceptor intron. This putative intron contains an additional HEG, but the annotation pipeline did not identify the sequence as an intron. This indicates that introns and intron encoded genes themselves are susceptible for intron invasions. The question is whether the invading intron has to retain its self-splicing function or the "host" (or acceptor) intron can splice the complete nested construct with its own self-splicing activity.

The intron regions contain most of the variation within *F. graminearum* and population specific introns promise to be valuable markers for tracking.

CONCLUSIONS

We have improved the reference mitochondrial genome reference sequence for *F. graminearum*. Intraspecific mitochondrial genome length variations are mainly due to intron presence/absence variation, thus using "intronless" length—subtracting the length of the intron regions from the total mitogenome length—could be a valuable information when comparing species. Mitogenomes are also subject to recombination in both *F. graminearum* and in *F. oxysporum*, indicating that it is a common phenomenon in *Fusarium*. We proposed that introns unique to a single population could be used to track the spread of the given population, because introns can spread through vertical inheritance, recombination and horizontal transfer. We also demonstrated how pooled sequencing of strains can be used for the mitogenome. The usage of pooled sequencing offers a scalable solution for population analysis and for species level comparisons studies. The results of this study represent an important step towards establishing pan-genomics for mitochondrial genomes.

REFERENCES

Al-Reedy, R. M., Malireddy, R., Dillman, C. B., and Kennell, J. C. (2012). Comparative analysis of *Fusarium* mitochondrial genomes reveals a highly variable region that encodes an exceptionally large open reading frame. *Fungal Genetics and Biology*, 49(1):2–14.

Avila-Adame, C., Gómez-Alpizar, L., Zismann, V., Jones, K. M., Buell, C. R., and Ristaino, J. B. (2006). Mitochondrial genome sequences and molecular evolution of the Irish potato famine pathogen, *Phytophthora infestans*. *Current Genetics*, 49(1):39–46.

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2012). SPAdes: A new genome assembly algorithm and its applications to single-cell sequencing. *Journal of Computational Biology*, 19(5):455–477.

Brankovics, B., van Dam, P., Rep, M., de Hoog, G. S., van der Lee, T. A. J., Waalwijk, C., and van Diepeningen, A. D. (2017). Mitochondrial genomes reveal recombination in the presumed asexual *Fusarium oxysporum* species complex. *BMC Genomics*, 18(1):735.

Brankovics, B., Zhang, H., van Diepeningen, A. D., van der Lee, T. A. J., Waalwijk, C., and de Hoog, G. S. (2016). GRAbB: Selective assembly of genomic regions, a new niche for genomic research. *PLoS Comput Biol.*, 12(6):e1004753.

Bruen, T. C., Philippe, H., and Bryant, D. (2006). A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4):2665–2681.

Computational Pan-Genomics Consortium (2018). Computational pan-genomics: status, promises and challenges. *Briefings in bioinformatics*, 19(1):118–135.

Cuomo, C. A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B. G., Di Pietro, A., Walton, J. D., Ma, L.-J., Baker, S. E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y.-L., Decaprio, D., Gale, L. R., Gnerre, S., Goswami, R. S., Hammond-Kosack, K., Harris, L. J., Hilburn, K., Kennell, J. C., Kroken, S., Magnuson, J. K., Mannhaupt, G., Mauceli, E., Mewes, H.-W., Mitterbauer, R., Muehlbauer, G., Münsterkötter, M., Nelson, D., O'Donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M. I. G., Seong, K.-Y., Tetko, I. V., Urban, M., Waalwijk, C., Ward, T. J., Yao, J., Birren, B. W., Kistler, H. C., O'donnell, K., Ouellet, T., Qi, W., Quesneville, H., Roncero, M. I. G., Seong, K.-Y., Tetko, I. V., Urban, M., Waalwijk, C., Ward, T. J., Yao, J., Birren, B. W., and Kistler, H. C. (2007). The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, 317(5843):1400–1402.

Darriba, D., Taboada, G. L., Doallo, R., and Posada, D. (2012). jModelTest 2: more models, new heuristics and parallel computing. *Nature Methods*, 9(8):772–772.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Desjardins, A. E. (2006). *Fusarium Mycotoxins: Chemistry, Genetics, and Biology*. The American Phytopathological Society, St. Paul, MN, USA.

Edgar, R. C. (2004a). MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics.*, 19(5):113.

Edgar, R. C. (2004b). MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, 32(5):1792–1797.

Ehrlich, G. D. (2001). The biofilm and distributed genome paradigms provide a new theoretical structure for understanding chronic bacterial infections. In *Interscience Conference on Antimicrobials Agents and Chemotherapy (ICAAC)*, Chicago, IL, USA.

Ehrlich, G. D., Hu, F. Z., and Post, J. C. (2004). Role for biofilms in infectious disease. In Ghannoum, M. and O'Toole, G. A., editors, *Microbial biofilms*, chapter 18, pages 332–358. ASM Press, Washington, DC.

Fourie, G., van der Merwe, N. A., Wingfield, B. D., Bogale, M., Tudzynski, B., Wingfield, M. J., and Steenkamp, E. T. (2013). Evidence for inter-specific recombination among the mitochondrial genomes of *Fusarium* species in the *Gibberella fujikuroi* complex. *BMC genomics*, 14(1):605.

Gale, L. R., Ward, T. J., Balmas, V., and Kistler, H. C. (2007). Population subdivision of *Fusarium graminearum sensu stricto* in the Upper Midwestern United States. *Phytopathology*, 97(11):1434–1439.

Gardiner, D. M., Stiller, J., and Kazan, K. (2014). Genome sequence of *Fusarium graminearum* isolate CS3005. *Genome Announcements*, 2(2):e00227–14.

Gilmore, S. R., Gräfenhan, T., Louis-Seize, G., and Seifert, K. A. (2009). Multiple copies of cytochrome oxidase 1 in species of the fungal genus *Fusarium*. *Molecular Ecology Resources*, 9(Suppl. 1):90–98.

Goddard, M. R. and Burt, A. (1999). Recurrent invasion and extinction of a selfish gene. *Proceedings of the National Academy of Sciences of the United States of America*, 96(24):13880–13885.

Harrison, R. G. (1989). Animal mitochondrial DNA as a genetic marker in population and evolutionary biology. *Trends in Ecology and Evolution*, 4(1):6–11.

Haugen, P., Simon, D. M., and Bhattacharya, D. (2005). The natural history of group I introns. *Trends in Genetics*, 21(2):111–119.

Hebert, P. D. N., Cywinska, A., Ball, S. L., and DeWaard, J. R. (2003). Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences*, 270(1512):313–321.

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*, 16:544.

Kulik, T., Brankovics, B., Sawicki, J., and van Diepeningen, A. D. (2016). The complete mitogenome of *Fusarium gerlachii*. *Mitochondrial DNA. Part A, DNA mapping, sequencing, and analysis*, 27(3):1895–6.

Laurent, B., Moinard, M., Spataro, C., Ponts, N., Barreau, C., and Foulongne-Oriol, M. (2017). Landscape of genomic diversity and host adaptation in *Fusarium graminearum*. *BMC genomics*, 18(203):1–19.

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

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., and Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16):2078–2079.

Liao, Y., Smyth, G. K., and Shi, W. (2013). The Subread aligner: Fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research*, 41(10):e108.

Liu, Y., Steenkamp, E. T., Brinkmann, H., Forget, L., Philippe, H., and Lang, B. F. (2009). Phylogenomic analyses predict sistergroup relationship of nucleariids and Fungi and paraphyly of zygomycetes with significant support. *BMC evolutionary biology*, 9:272.

Marchler-Bauer, A. and Bryant, S. H. (2004). CD-Search: Protein domain annotations on the fly. *Nucleic Acids Research*, 32(Web Server issue):327–331.

Mitchell, A., Chang, H.-Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.-Y., Bateman, A., Punta, M., Attwood, T. K., Sigrist, C. J. A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D. A., Wu, C. H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P. D., and Finn, R. D. (2015). The InterPro protein families database: The classification resource after 15 years. *Nucleic Acids Research*, 43(Database issue):D213–D221.

Nurk, S., Bankevich, A., Antipov, D., Gurevich, A. A., Korobeynikov, A., Lapidus, A., Prjibelski, A. D.,

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Pyshkin, A., Sirotkin, A., Sirotkin, Y., Stepanauskas, R., Clingenpeel, S. R., Woyke, T., McLean, J. S., Lasken, R., Tesler, G., Alekseyev, M. A., and Pevzner, P. A. (2013). Assembling single-cell genomes and mini-metagenomes from chimeric MDA products. *Journal of Computational Biology*, 20(10):714–37.

Pavesi, A., Conterio, F., Bolchi, A., Dieci, G., and Ottonello, S. (1994). Identification of new eukaryotic tRNA genes in genomic DNA databases by a multistep weight matrix analysis of transcriptional control regions. *Nucleic Acids Res.*, 22(7):1247–1256.

Raineri, E., Ferretti, L., Esteve-Codina, A., Nevado, B., Heath, S., and Pérez-Enciso, M. (2012). SNP calling by sequencing pooled samples. *BMC Bioinformatics*, 13:239.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D. L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M. A., and Huelsenbeck, J. P. (2012). MrBayes 3.2: Efficient bayesian phylogenetic inference and model choice across a large model space. *Syst. Biol.*, 61(3):539–542.

Smith, A. M., Heisler, L. E., St.Onge, R. P., Farias-Hesson, E., Wallace, I. M., Bodeau, J., Harris, A. N., Perry, K. M., Giaever, G., Pourmand, N., and Nislow, C. (2010). Highly-multiplexed barcode sequencing: an efficient method for parallel analysis of pooled samples. *Nucleic Acids Research*, 38(13):e142.

Stamatakis, A. (2014). RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, 30(9):1312–1313.

Tettelin, H., Masignani, V., Cieslewicz, M. J., Donati, C., Medini, D., Ward, N. L., Angiuoli, S. V., Crabtree, J., Jones, A. L., Durkin, A. S., DeBoy, R. T., Davidsen, T. M., Mora, M., Scarselli, M., Margarit y Ros, I., Peterson, J. D., Hauser, C. R., Sundaram, J. P., Nelson, W. C., Madupu, R., Brinkac, L. M., Dodson, R. J., Rosovitz, M. J., Sullivan, S. A., Daugherty, S. C., Haft, D. H., Selengut, J., Gwinn, M. L., Zhou, L., Zafar, N., Khouri, H., Radune, D., Dimitrov, G., Watkins, K., O'Connor, K. J. B., Smith, S., Utterback, T. R., White, O., Rubens, C. E., Grandi, G., Madoff, L. C., Kasper, D. L., Telford, J. L., Wessels, M. R., Rappuoli, R., and Fraser, C. M. (2005). Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial "pan-genome". *Proceedings of the National Academy of Sciences*, 102(39):13950–13955.

Wang, Q., Jiang, C., Wang, C., Chen, C., Xu, J.-R., and Liu, H. (2017). Characterization of the two-speed subgenomes of *Fusarium graminearum* reveals the fast-speed subgenome specialized for adaption and infection. *Frontiers in Plant Science*, 8:140.

Ward, T. J., Clear, R. M., Rooney, A. P., O'Donnell, K., Gaba, D., Patrick, S., Starkey, D. E., Gilbert, J., Geiser, D. M., and Nowicki, T. W. (2008). An adaptive evolutionary shift in Fusarium head blight pathogen populations is driving the rapid spread of more toxigenic *Fusarium graminearum* in North America. *Fungal Genetics and Biology*, 45:473–484.

Wick, R. R., Schultz, M. B., Zobel, J., and Holt, K. E. (2015). Bandage: Interactive visualization of *de novo* genome assemblies. *Bioinformatics*, 31(20):3350–3352.

Yun, S. H., Arie, T., Kaneko, I., Yoder, O. C., and Turgeon, B. G. (2000). Molecular organization of mating type loci in heterothallic, homothallic, and asexual Gibberella/Fusarium species. *Fungal genetics and biology*, 31(1):7–20.

Zhang, H., van der Lee, T. A. J., Waalwijk, C., Chen, W., Xu, J., Xu, J., Zhang, Y., and Feng, J. (2012). Population analysis of the species complex from wheat in China show a shift to *Fusarium graminearum* more aggressive isolates. *PLoS One*, 7(2):e31722.

Zhao, C., Waalwijk, C., de Wit, P., Tang, D., and van der Lee, T. (2014). Relocation of genes generates non-conserved chromosomal segments in *Fusarium graminearum* that show distinct and co-regulated gene expression patterns. *BMC Genomics*, 15(1):191.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

ADDITIONAL FILES

Supplementary Table 1

Intron locations, lengths and haplotypes within standard mitochondrial genes of *Fusarium graminearum* and *F. gerlachii* strains. The different haplotypes are displayed with different background color per intron site. Haplotypes that have identical length are differentiated from each other by using ', ^ or *, corresponding to the number of SNPs differentiating the haplotypes (haplotypes 1159' and 1159"' differ by 2 SNPs, while 1159 and 1159"' differ by 3 SNPs).

Supplementary Text 1

Assembling the pooled data set.

TABLES AND FIGURES

Table 1. List of *Fusarium* strains analysed in this study

Species	Strain	Origin	Host	Year of isolation	Sequenced individually or in a pool
F. graminearum	CBS123657 (PH-1) NRRL31084	USA	maize	1996	both
F. graminearum	CBS119173	USA	wheat head	2005	individually
F. graminearum	CBS139513	Argentina	barley	2011	individually
F. graminearum	CBS139514	Argentina	barley	2010	individually
F. graminearum	CBS119799	South Africa	wheat kernel	1987	individually
F. graminearum	CBS119800	South Africa	maize	1990	individually
F. graminearum	CBS110263	Iran	maize	1968	individually
F. graminearum	CBS123688	Sweden	oats	unknown	individually
F. graminearum	CBS128539	Belgium	wheat kernel	2007	individually
F. graminearum	CBS138561	Poland	wheat kernel	2010	individually
F. graminearum	CBS138562	Poland	wheat kernel	2010	individually
F. graminearum	CBS138563	Poland	wheat kernel	2003	individually
F. graminearum	CBS104.09	unknown	unknown	1909	individually
F. graminearum	CBS185.32	unknown	maize	1932	individually
F. graminearum	CS3005	Australia	barley	2001	individually
F. graminearum	HN9-1	China	wheat	2002	individually
F. graminearum	HN-Z6	China	wheat	2012	individually
F. graminearum	INRA-156	France	wheat	2001	individually
F. graminearum	INRA-159	France	wheat	2001	individually
F. graminearum	INRA-164	France	wheat	2002	individually
F. graminearum	INRA-171	France	wheat	2001	individually
F. graminearum	INRA-181	France	wheat	2002	individually
F. graminearum	INRA-195	France	wheat	2002	individually
F. graminearum	YL-1	China	wheat	2012	individually
F. graminearum	bfb0999_1	China	barley	2005	pooled
F. graminearum	68D2	Netherlands	wheat	2001	pooled
F. graminearum	CHG013	China	maize	2005	pooled
F. graminearum	CHG157	China	barley	2005	pooled
F. gerlachii	CBS123666	USA	wheat head	2000	individually

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

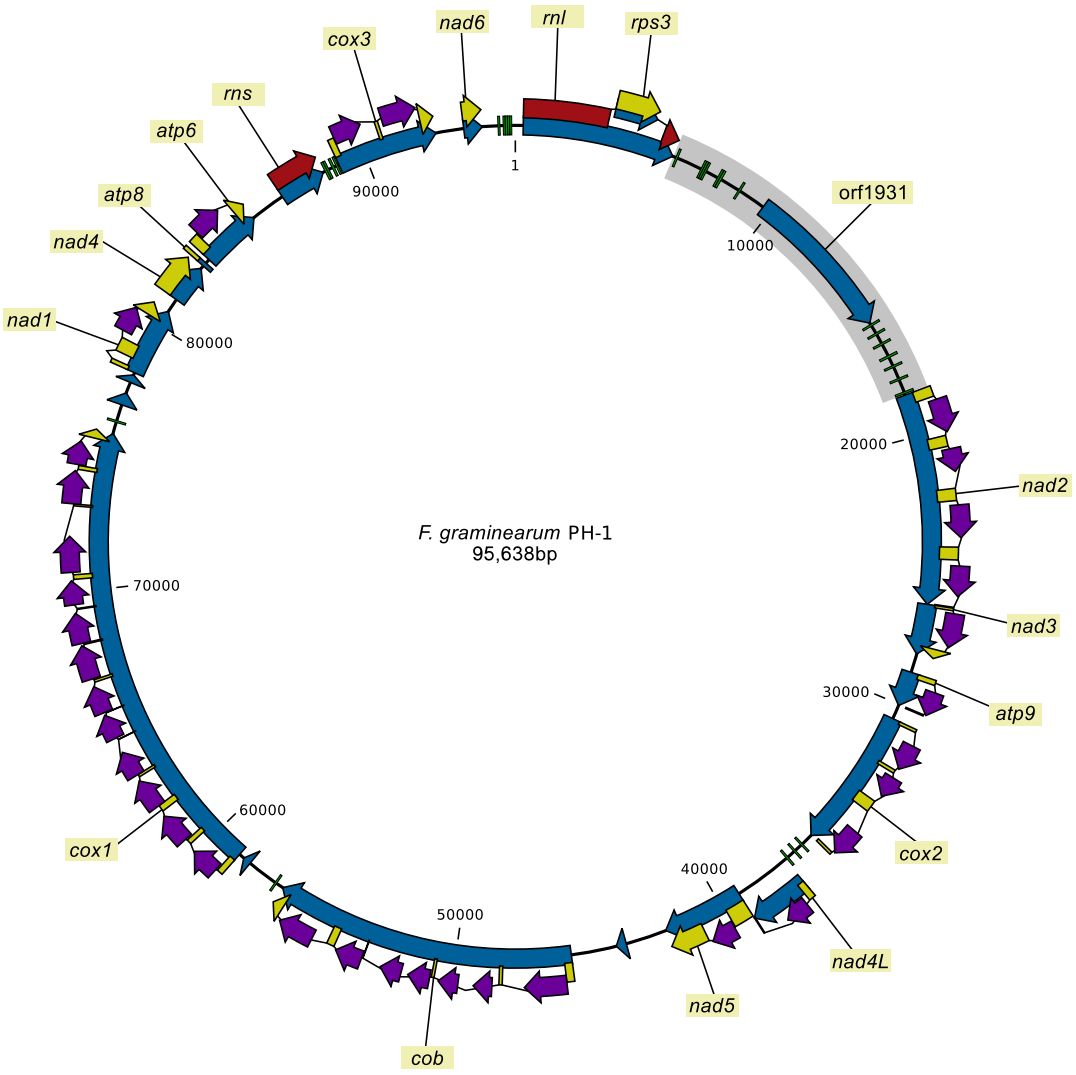

Figure 1. The mitogenome of *F. graminearum* strain PH-1. Green blocks: tRNA coding genes, blue arrows: genes or ORFs (no labels added for short ORFs), yellow arrows: protein coding sequences, red arrows: rDNA coding sequence, purple arrows: intron encoded homing endonuclease genes, gray box: the large variable (LV) region with orf1931 (LV-uORF).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

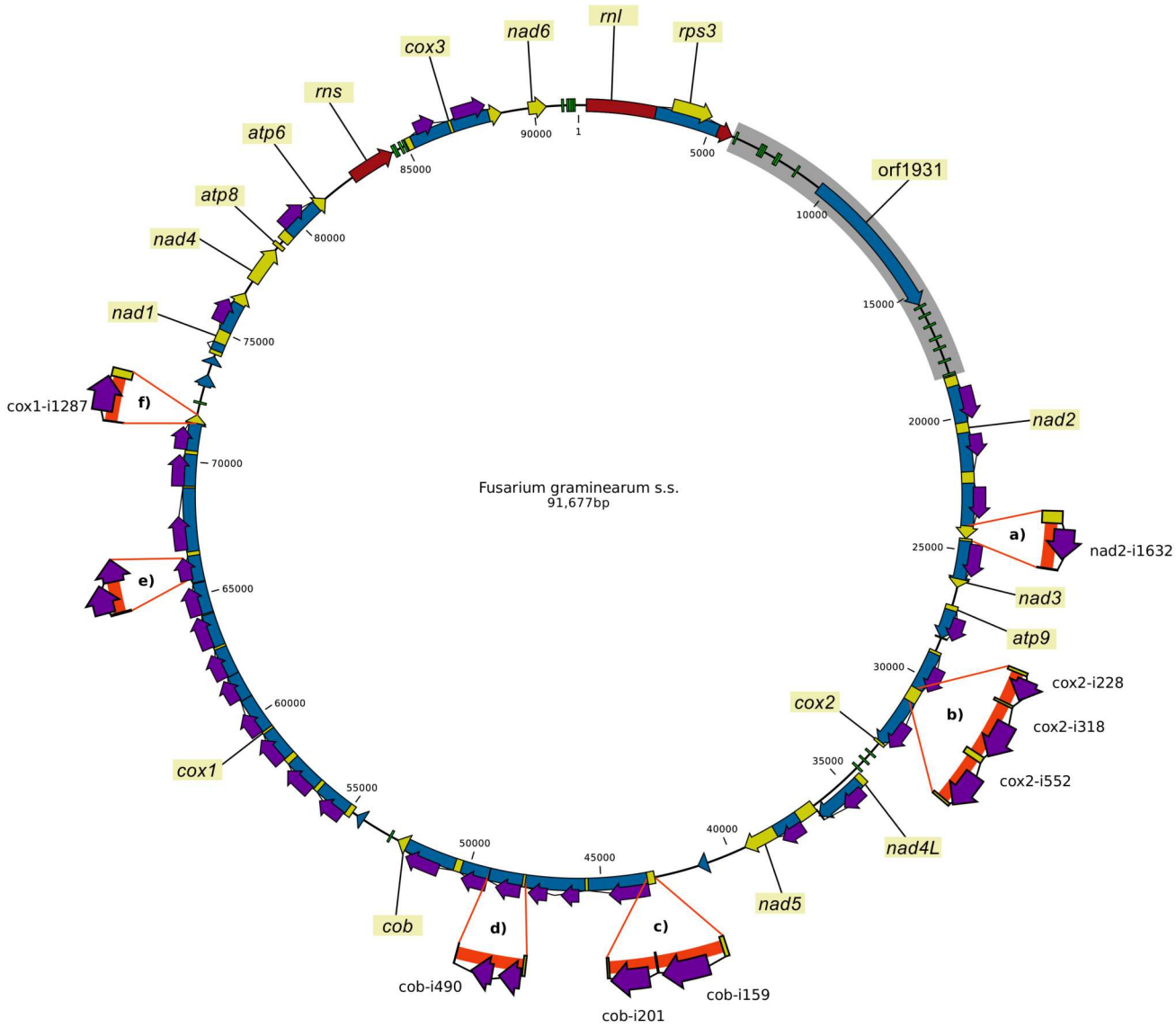

Figure 2. Pan-genomic representation of the presence/absence variation of introns in the mitochondrial genomes of the 24 *F. graminearum* strains.
In the figure, the thick orange lines highlight intron sequences in the alternative sequences. (SNPs and short indels are not indicated.) **a)** The insertion of nad2-i1632; **b)** the insertion of cox2-i228, cox2-i318 and cox2-i552; **c)** the insertion of cob-i159 and cob-i201; **d)** longer variant of cob-i490; **e)** intron insertion in the HEG located in cox1-i906; and **f)** the insertion of cox1-1287.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Table 2. Mitochondrial genome variation of the *Fusarium graminearum* strains

Strain ID	GenBank accession numbers	Size (bp)	Introns	Intronic (bp)	Core (bp)
CBS123657 (PH-1)	MH412632	95638	34	49429	46209
CBS185.32	KP966550	96300	34	50120	46180
CBS110263	KP966551	97364	35	51165	46199
CBS119173	KP966552	100342	37	54130	46212
CBS119799	KP966553	96005	35	49919	46086
CBS119800	KP966554	97462	35	51280	46182
CBS123688	KP966555	95035	34	48837	46198
CBS128539	KP966556	96134	35	49996	46138
CBS138561	KP966557	95034	34	48837	46197
CBS138562	KP966558	99062	36	52865	46197
CBS138563	KP966559	99068	36	52865	46203
CBS139514	KP966560	96167	35	49980	46187
CBS139513	KP966561	95041	34	48837	46204
CBS104.09	KR011238	97460	35	51280	46180
CS3005	BK010538	93560	33	47381	46179
HN9-1	BK010539	96307	35	51567	44740
HN-Z6	BK010540	97767	34	50120	47647
INRA-156	BK010541	101424	37	55243	46181
INRA-159	BK010542	96199	35	49996	46203
INRA-164	BK010543	99678	37	53476	46202
INRA-171	BK010544	96199	35	49996	46203
INRA-181	BK010545	96187	35	49996	46191
INRA-195	BK010546	97358	35	51165	46193
YL-1	BK010547	97996	36	51777	46219

Core stands for the total mitogenome length minus the length of the intron regions.

Table 3. Distribution of variation in the intron and intergenic regions within and between species

	Intraspecies			Interspecies		
	Length (bp)	Variable positions	Variation frequency	Length (bp)	Variable positions	Variation frequency
Coding	21572	4	0.02%	21572	5	0.02%
intron	59091	399	0.68%	59091	419	0.71%
Intergenic	18982	310	1.63%	18982	436	2.30%

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018

Table 4. List of single nucleotide polymorphisms identified in the pooled dataset of *Fusarium graminearum* strains. Positions are aligned positions between the PH-1 reference sequence and the pooled sequences ("short" and "long"). "Reference" refers to the nucleotide found in the given reference sequence used for mapping, while "Alternative" refers to the nucleotide suggested by the mapped reads. Position 90636 shows unusual ratios: in both mappings the reference nucleotide (C or A) has a frequency of 70% and the alternative nucleotide has 30%. This is due to an adjacent indel that affects the mapping results.

	PH-1		Pooled	
Position	Reference	Alternative	Reference	Alternative
2337	A (0.77)	G (0.23)	A (0.77)	G (0.23)
6288	C (0.41)	A (0.59)	A (0.61)	C (0.39)
6355	T (0.42)	C (0.58)	C (0.60)	T (0.40)
13540	C (0.78)	A (0.22)	C (0.78)	A (0.22)
37126	C (0.75)	T (0.25)	C (0.75)	T (0.25)
37773	A (0.75)	G (0.25)	A (0.75)	G (0.25)
44773	A (0.62)	G (0.38)	A (0.62)	G (0.38)
64776	G (0.53)	A (0.47)	G (0.53)	A (0.47)
70827	A (0.62)	G (0.38)	A (0.62)	G (0.38)
89194	G (0.57)	A (0.43)	G (0.57)	A (0.43)
90636	C (0.70)	A (0.30)	A (0.70)	C (0.30)
95918	A (0.43)	C (0.57)	C (0.59)	A (0.41)
99784	A (0.40)	G (0.60)	G (0.62)	A (0.38)
100362	C (0.42)	A (0.58)	A (0.59)	C (0.41)
100538	G (0.42)	A (0.58)	A (0.61)	G (0.39)

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.27017v1 | CC BY 4.0 Open Access | rec: 3 Jul 2018, publ: 3 Jul 2018