

RESEARCH ARTICLE

Open Access



The completed genome sequence of the pathogenic ascomycete fungus *Fusarium graminearum*

Robert King^{1*} , Martin Urban², Michael C. U. Hammond-Kosack², Keywan Hassani-Pak¹ and Kim E. Hammond-Kosack^{2*}

Abstract

Background: Accurate genome assembly and gene model annotation are critical for comparative species and gene functional analyses. Here we present the completed genome sequence and annotation of the reference strain PH-1 of *Fusarium graminearum*, the causal agent of head scab disease of small grain cereals which threatens global food security. Completion was achieved by combining (a) the BROAD Sanger sequenced draft, with (b) the gene predictions from Munich Information Services for Protein Sequences (MIPS) v3.2, with (c) de novo whole-genome shotgun re-sequencing, (d) re-annotation of the gene models using RNA-seq evidence and Fgenesh, Snap, GeneMark and Augustus prediction algorithms, followed by (e) manual curation.

Results: We have comprehensively completed the genomic 36,563,796 bp sequence by replacing unknown bases, placing supercontigs within their correct loci, correcting assembly errors, and inserting new sequences which include for the first time complete AT rich sequences such as centromere sequences, subtelomeric regions and the telomeres. Each of the four *F. graminearum* chromosomes was found to be submetacentric with respect to centromere positioning. The position of a potential neocentromere was also defined. A preferentially higher frequency of genetic recombination was observed at the end of the longer arm of each chromosome. Within the genome 1529 gene models have been modified and 412 new gene models predicted, with a total gene call of 14,164. The re-annotation impacts upon 69 entries held within the Pathogen-Host Interactions database (PHI-base) which stores information on genes for which mutant phenotypes in pathogen-host interactions have been experimentally tested, of which 59 are putative transcription factors, 8 kinases, 1 ATP citrate lyase (ACL1), and 1 syntaxin-like SNARE gene (GzSYN1). Although the completed *F. graminearum* contains very few transposon sequences, a previously unrecognised and potentially active gypsy-type long-terminal-repeat (LTR) retrotransposon was identified. In addition, each of the sub-telomeres and centromeres contained either a LTR or MarCry-1_FO element. The full content of the proposed ancient chromosome fusion sites has also been revealed and investigated. Regions with high recombination previously noted to be rich in secretome encoding genes were also found to be rich in tRNA sequences. This study has identified 741 *F. graminearum* species specific genes and provides the first complete genome assembly for a Sordariomycetes species.

Conclusions: This fully completed *F. graminearum* PH-1 genome and manually curated annotation, available at Ensembl Fungi, provides the optimum resource to perform interspecies comparative analyses and gene function studies.

Keywords: *Gibberella zeae*, *Fusarium* head blight disease, Fungal centromere sequences, Fungal telomere sequences, Fungal virulence gene hotspots, Completed Sordariomycetes genome

* Correspondence:

robert.king@rothamsted.ac.uk; kim.hammond-kosack@rothamsted.ac.uk

¹Department of Computational and Systems Biology, Rothamsted Research, Harpenden, Herts AL5 2JQ, UK

²Department of Plant Biology and Crop Science, Rothamsted Research, Harpenden, Herts AL5 2JQ, UK



© 2015 King et al. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly credited. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated.

Background

The acquisition of a completed genomic reference sequence has numerous benefits over a near or partial completed sequence. Fully completed genomes have higher impact in terms of the comparative power with other species to evaluate the divergence of pathways and the expansion and contraction of gene families. Completed genomes can also be used to align syntenic regions between related species and thereby improve the accuracy and efficacy of other genome assembly projects [1]. In addition, comparative gene finding tools can be used to annotate other related species [2] and define species specific genes.

The Fusaria are globally important ascomycete fungi of growing agricultural, horticultural, medical and ecological importance. Different species of Fusaria can infect and cause disease on a wide diversity of host species including cereal and non-cereal crops, for example *F. graminearum*, only non-cereal crops (*F. oxysporum* spp and *F. solani*), forestry trees (*F. circinatum*), turf grass (*Microdochium nivale*), animals, fish, shell fish and humans (*F. solani* species complex). Whereas other Fusarium species are a major source of a high value edible protein consumed by humans (*F. venenatum*) or are used to produce specific biomolecules, for example plant growth promoting gibberellins (*F. fujikuroi*). As a consequence over the past decade, multiple Fusarium sequencing projects have taken place [3, 4] or are currently in progress [5]. Although the haploid Fusarium genomes are small in size, typically 35-50 Mbp, for no Fusarium species is the complete genome sequence available. This is counter-intuitive to providing high quality genomic resources for research however a full spectrum of genomic resources currently exists. This is problematic for the international Fusarium community. The most complete genomic reference sequence of the Fusaria clade is available for *F. graminearum*. The assembled contigs/supercontigs for this species have been assigned to four large chromosomes by using information from a genetic map [4]. But the assembled genome 10 years later still contains > 212,843 unknown bases (N's), and lacks telomeric and centromeric sequences. Although incomplete this genome has provided the basis to align and assemble genomes of four other plant pathogenic Fusarium species, namely *F. fujikuroi* B14 [6], *F. verticillioides*, *F. oxysporum*, *f. sp lycopersici* [7], *F. pseudograminearum* CS3096, and isolates of *F. pseudograminearum* [8] including the highly virulent *F. pseudograminearum* CS3005 isolate [9]. Unfortunately, each of these other genomes also consists of multiple scaffolds which are heavily interspersed with unknown bases. Consequently their gene annotations may be incomplete. The value of a complete genomic reference sequence is complemented by high-quality annotation of gene and non-coding RNA annotation because all targets

of interest can be identified and tested. Gene deletion studies to probe the function of proteins/metabolites potentially linked to virulence or other phenotypic traits of interest [10] are dependent upon the quality of the reference genome and the annotation provided. Therefore an increased confidence in research findings based upon genome and annotation use is observed using a complete reference/annotation versus an incomplete. For example, deletion or silencing of a class of enzymes that are implicated in virulence is not hindered by unknown genes of the same class which maintain the virulence effect.

Some Fusarium species, for example *F. graminearum*, are ideal targets for full genome completion given the predicted low repetitive sequence content and small genome size. Completion of the *F. graminearum* reference genome would be the first completed within the class Sordariomycetes, which contains fungal species with very diverse lifestyles ranging from free living saprophytes, plant invading endophytes, to pathogens of microbes, plants and/or animals. Ongoing comparative analysis between and within Fusarium species is important to food security due to their role in infecting hosts which include; wheat and barley (*F. culmorum* and *F. graminearum*), maize and sorghum (*F. verticillioides*), tomato, tobacco, legumes, cucurbits, sweet potatoes and banana (*F. oxysporum* numerous *formae specialis*), and rice (*F. fujikuroi*).

Fusarium graminearum is a fungus of world-wide economic importance because it causes Fusarium head blight (FHB) disease on wheat and barley, also known as head scab disease, and Fusarium stalk and ear rot disease on maize [11]. The infection results in reduced yield affecting grain/kernel mass and quality, and mycotoxin contamination [12, 13], which is a threat to human health [14, 15]. Consequently mycotoxin content in the harvested grain has to be expensively monitored due to maximum tolerated levels of toxins permitted in human foods and animal feeds [16]. Research to reduce FHB disease in cereal species has used the PH-1 strain as a model to identify and understand virulence mechanisms and pathways, with the ultimate aim of developing novel species specific fungicides and/or resistant crop plants [17].

The *F. graminearum* PH-1 strain was originally sequenced using Sanger technology, assembled into 433 contigs and 31 supercontigs, anchored to a genetic map, and annotated by the BROAD institute [4, 18]. Later on the sequence was improved by BROAD, and the gene model annotations were refined by Munich Information Services for Protein Sequences (MIPS) [19, 20], resulting in 13,826 mRNA transcripts in the MIPS annotation (version 3.2). However there remain >397 gaps of unknown sequences within the current assembly represented either by sequential N bases or completely lacking sequence as indicated by the absence of the eight telomeres. In addition, no sequence is available for the

four centromeres and 99,079 bp remained unassembled in small contigs. For comparative purposes both the reference and annotation are both referred to in the text as MIPS v3.2.

In this study, to complete the *F. graminearum* genome, whole shotgun re-sequencing of the strain PH-1 to 85-fold coverage was done using short paired end reads followed by *de novo* assembly using various k-mer values. The accompanying annotation was produced by transferring the MIPS v3.2 annotation, then refining the gene model annotation integrating multiple gene prediction algorithms within the Maker2 software [21] as well as by using available RNA-seq from two wild type *F. graminearum* isolates (PH-1 and Gz-3639) to provide additional gene model support. Finally, manual inspection was done on all the gene models. In addition, the PH-1 mitochondrial draft [GenBank:NC_009493] which was 95,676 bp in length was re-evaluated, corrected and reduced to 95,638 bp. To complete this project, the RRes v4.0 genome sequence and the mitochondrial sequence were submitted to European Nuclear Archive (ENA) [EMBL:HG970330, EMBL:HG970331, EMBL:HG970332, EMBL:HG970333, EMBL:HG970334, EMBL:HG970335]. The completed *F. graminearum* genome is available within ENSEMBL fungi [22] and displayed on the plant pathogen specific PhytoPath database [23].

Results

Genome closure

Assembly

The *de novo* assembly using a sequencing depth of 85x coverage created a *F. graminearum* genome of 36,563,796 bp in length, assembled into four chromosomes and a mitochondrial genome of 95,638 bp in length. There was high accordance between the alignments of the RRes v4.0 genome to the previous MIPS v3.2 reference genome. The alignment of the *de novo* contigs identified a previous mis-assembly at the amino end of chromosome 1 with the 1-306,000 bp sequence observed to be the reverse complement of the true sequence at this position. This was confirmed by the extension at the amino end to a telomere sequence and the manual inspection of mapped reads throughout the genome to detect soft clipping of reads which are indicators of misassemblies. The *de novo* assembly also provided extension of the amino and carboxyl sequence of chromosome 1-4 and identified a common telomere sequence (TAACCC) and reverse complement (GGGTTA) respectively [24] which is found in all vertebrates [25]. This new data completes the four chromosomes at the amino and carboxyl end and confirms the new orientation of the amino terminus of chromosome 1. In addition, a total of 25,707 N bases have been removed to create the new assembly.

Three previously constructed supercontigs namely 3.26, 3.31, and 3.15 had a similar depth of coverage (~85) to the reads that mapped to the remainder of the genomic reference (Table 1). Therefore each of these sequences became the targets for amalgamation into chromosomes 1 to 4 sequences. After the alignment of contigs from the *de novo* assemblies and replacement of unknown bases these three supercontigs were successfully placed into the genome. Supercontig 3.15 was placed at the carboxyl end of chromosome 2 (Fig. 1) to provide the subtelomeric AT-rich sequence and the missing telomere sequence. Supercontigs 3.26 and 3.31 were placed respectively in chromosomes 3 at position 2,466,775-2,469,779 in which a new gene prediction is found (FGRRES_20327), and chromosome 1 at position 4,541,855-4,543,915.

Chromosome 4 was found to lack a telomere sequence at the carboxyl terminus but was found instead to possess a region of highly repetitive sequences solely annotated as rRNA coding. This large feature of the genome was found to extend for at least 1.38 Mbp and was identified using the following manual curation procedure. Mapping the reads of the re-sequenced PH-1 isolate to the MIPS reference sequence revealed a disproportionately higher read coverage for 12 of the 15 supercontigs not yet placed within the genome compared to the rest of the *de novo* assembled genome (Table 1). This suggested that some or all of these supercontig sequences were represented in multiple copies. The high copy number of each supercontig may have contributed to the difficulty experienced in prior efforts to place these sequences within chromosomes 1-4 [4]. Following manual inspection, ten of these repetitive sequences containing supercontigs could be aligned to the carboxyl end of chromosome 4. Each of these supercontigs contained rRNA encoding repetitive sequence. This manual procedure permitted the chromosome 4 to be extended by 1.38 Mbp at the carboxyl terminus by further extrapolation of rRNA sequence by division of the total mapped read coverage of this region by the expected 85 x coverage. Supercontig 3.13 was identified as mitochondrial DNA and was transferred into the mitochondrial assembly pipeline. The unknown sequence within supercontig 3.12 (MIPS length 15,604 bp) was corrected and the only supercontig remaining which had multiple copy coverage (>85) but did not align partially or completely to chromosome 1-4 sequences or the mitochondria. Manual curation of the mapped reads suggests this supercontig to be either a repeating unit or a circular sequence of 5846 bp with no known genomic loci. Consequently this supercontig was reduced to one single repeating unit. The identification of genomic loci of the remaining supercontigs from the original Sanger sequencing and unknown sequence replacement completes chromosomes 1-4 sequences.

Table 1 Chromosome and Supercontig sequence coverage of the *F. graminearum* PH-1 isolate and MIPS reference. Identification of underrepresented repeating sequences represented in the MIPS reference sequence could be calculated by dividing the total corrected average observed coverage by the original sequencing depth of 85.

MIPS Chromosome/ Supercontig number	Length (bp)	Raw average coverage ^b	Bases with coverage (%) ^c	Corrected average coverage ^d	Calculated multiples of supercontig sequences in genome ^e
1	11,694,295	79.5	100.0	79.5	1
2	8,911,601	90.7	99.6	90.3	1
3	7,711,129	90.4	95.5	86.3	1
4	8,029,942	70.6	91.8	64.8	1
4: 7,953,943+ bp ^a	75,973	1,000.2	77.8	778.16	9
3.13	12,585	1,083	96	1,039.7	12
3.18	8,774	930.3	99	921	11
3.2	7,037	1,000.6	91.3	913.5	11
3.29	2,326	731.9	95	695.3	8
3.28	2,628	687.8	98.1	674.7	8
3.3	2,172	573.5	99.2	568.9	7
3.22	6,062	573.5	75.5	433	5
3.24	6,000	557.4	73.5	409.7	5
3.23	6,331	541.4	74.2	401.7	5
3.21	7,489	551.3	78.2	431.1	5
3.25	6,524	368.1	63	231.9	3
3.15	10,471	96.7	100	96.7	1
3.31	2,080	85.8	100	85.8	1
3.26	2,996	95.7	100	95.7	1
3.12	15,604	438.9	99.4	436.3	5

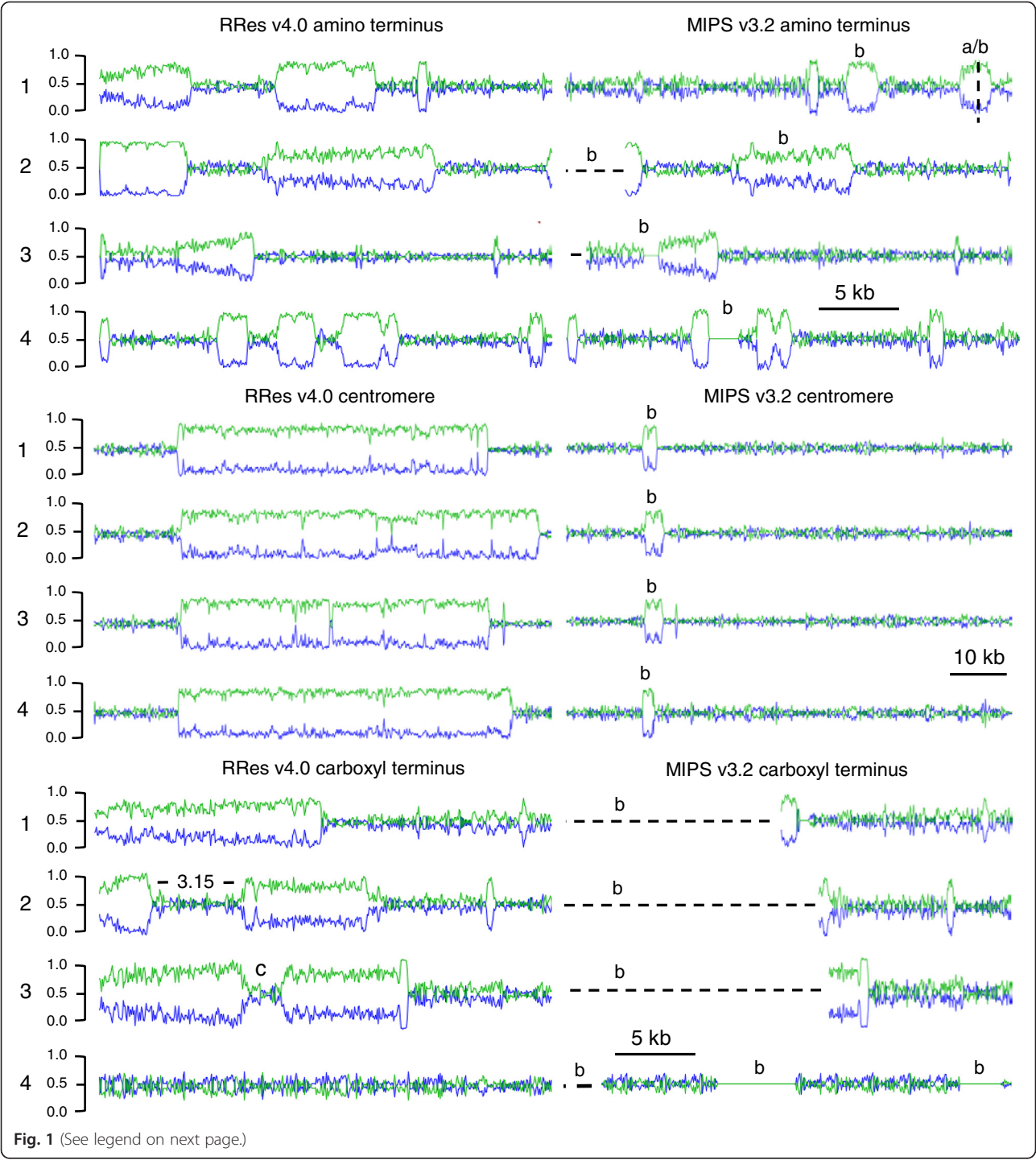
^aChromosome 4 from 7,953,943 bp onwards, this is the start of the repeating RNA annotated sequence^bThe read depth observed across the length of the sequence^cPercentage of bases that are A,C,T,G but not N^dThe calculated N bp corrected coverage^eThe number of copies of the sequence that should be represented in the reference sequence

The remaining 1.28 million re-sequenced un-mapped reads underwent *de novo* assembly of which the resulting 777,064 remaining reads and 79,424 contigs were analysed using BLAST with a 0.01 significance value followed by MEGAN [26]. These reads and contigs were shown to represent contamination (99.1 % data not shown) with 74 sequences that had fungal hits. The length of these 74 individual contigs did not exceed 400 bp and the concatenated length was 13,850 bp. The 55,093 unplaced contigs from the original Sanger sequencing used in the MIPS v3.2 genome were mapped to the new reference and mitochondrial sequence, RRes v4.0. The majority mapped to these sequences using BWA MEM with only 3131 sequences remaining to be placed of a combined length of 460,949 bp. These remaining sequences were assigned using BLAST with a 0.01 significance value and MEGAN to Mouse 55,653 bp (90 sequences), low complexity 186,753 bp (404 sequences), no hits 193,456 bp (2375 sequences), and not assigned 25,087 bp (262 sequences). The BLAST and MEGAN analysis of the remaining sequencing data from

both sequencing platforms suggests that there are no additional chromosomes to the four represented in RRes v4.0.

The completed *F. graminearum* genome

The revised RRes v4.0 of the *F. graminearum* PH-1 strain genomic reference, deposited to ENA, contains chromosomes 1 through 4 and comprises 36,563,796 bp with 12 'N' bases. This number excludes the significantly extended repetitive rRNA encoding region at the carboxyl end of chromosome-4 (Table 2). In contrast in the MIPS v3.2 genome, chromosomes 1-4, comprises 36,261,856 bp with 25,707 'N' bases whilst the BROAD FG3.0 version is in the form of multiple large supercontigs 3.1-3.10 and comprises 36,073,610 bp (supercontigs above 25 kb) with 210,520 'N' bases [5]. The genomic length of RRes v4.0 chromosomes 1-4 is 354,665 bp longer than that of MIPS v3.2. In order that this figure was not distorted, the calculations excluded the repetitive rRNA annotated supercontigs within MIPS and complementary sequence at the carboxyl end of chromosome 4



(See figure on previous page.)

Fig. 1 Visual representation of AT rich and repeat rich regions in the RRes and MIPS references. All values are approximations to the nearest 1 kb and images have been scaled to equivalency. The green line represents the AT percentage and the blue the CG percentage to give a combined total of 100 %. The labelling is in the format of a number which represents the chromosome i.e. 1 = chromosome 1, a = the end of a reverse complement section, and b denotes regions in the MIPS version that were extended in the RRes version. The 3.15 label represents where the supercontig 3.15 was placed and "c" represents where FGRRES_20327 is located. The amino terminus set (5') (upper panel) reveals the AT rich extensions of Chromosomes 2-4 and the sequence that was reverse complemented in Chromosome 1. The centromere set (middle panel) shows chromosome 1 AT rich region extension from 3 kbp to 57 kbp MIPS to RRes respectively within 8.97 Mbp region, Chromosome 2 extension from 3 kbp to 65 kbp MIPS to RRes respectively within 3.27 Mbp region, Chromosome 3 extension from 4 kbp to 56 kbp MIPS to RRes respectively within 5.35 Mbp region, and Chromosome 4 extension from 2 kbp to 61 kbp MIPS to RRes respectively within 5.0 Mbp region. The carboxyl terminus set (lower panel) shows Chromosome 1 extension at the carboxyl end of 14 kbp, Chromosome 2 extension at the carboxyl end of 17 kbp, Chromosome 3 extension at the carboxyl end of 19 kbp, Chromosome 4 extension at the carboxyl end of a repetitive sequence 1.3 Mbp (complete sequence length not shown)

in both the MIPS and RRes v4.0 reference because of significant extension in RRes v4.0. The 12 unknown bases in the RRes v4.0 are found at a single poly-A locus within the centromere of chromosome 1. There were 2,433 homozygous mutations observed when mapping the re-sequenced PH-1 to the MIPS v3.2 genomic reference consisting of: 2,025 snps with a transition/transversion ratio of 0.5555, 226 insertions, and 182 deletions. These mutations may have arisen due to strain passaging or a consequence of sequencing errors from the original sequencing but the observation of these mutations clustering at loci suggests due to passaging. Readers are reminded that all genomic reference sequences are snapshots of an individual or a subset of individuals of a population. The size of the four chromosomes was found to be; Chromosome 1: 11,760,938 bp (excluding 12 'N' bases), Chromosome 2: 8,997,558 bp, Chromosome 3: 7,792,947 bp, and Chromosome 4: 9,407,501 bp.

To complete the *F. graminearum* mitochondrial genome, 15 SNPs were corrected, 28 indels deleted and 2 indels were inserted. As a result, the new mitochondrial reference genome deposited at ENA has a length that is slightly lesser in size at 95,638 bp, compared to the original size of 95,676 bp. Both the original Sanger sequence reads and the new mitochondrial genome obtained by paired end reads exhibited no polymorphisms. This result not only confirms the reliability of the new version of the

genome, but clearly demonstrates that there has been no change in the mitochondrial sequence of strain PH-1 in over a decade of repeated fungal growth *in vitro*, *in planta* and long term storage at -80C.

Resolving two specific genome features: telomeres and centromeres

The placement and extension of AT rich sequence is normally difficult to resolve during genome assembly due to repetitive DNA content. These sequences predominate at the amino and carboxyl ends of each chromosome and represent the subtelomeric regions. In most eukaryotic species there are multiple such regions within the relative vicinity of the repeating telomere sequence (TAACCC). This was found to be the case in the seven assembled subtelomeric regions. In Fig. 1, the original and de novo extended AT rich sequences at the end of each chromosome are depicted. To achieve the subtelomeric assembly, a multi k-mer approach was used in which a range of values from 61 to 85 was sequentially aligned to extend the sequence. However, this difficult part of the assembly was most likely achieved because of the previously noted intrinsically low abundance of repeat sequences in *F. graminearum* PH-1 [4].

The extension of an AT rich region within each chromosome permitted the four centromere regions to be defined and accurately positioned. In chromosome 1,

Table 2 Comparative statistics on the length and nucleotide composition of the four *F. graminearum* chromosomes. Comparison of the MIPS (old) and RRes (new) versions of the reference genome

Reference	MIPS			RRes			
	Number of total bases (bp)	Number of 'N' bases (bp)	Total number of bases minus 'N' bases (bp)	Number of total bases (bp)	Number of 'N' bases (bp)	Total number of bases minus 'N' bases (bp)	Length difference RR vs. MIPS (bp) ^a
Chromosome 1	11,694,295	65,691	11,628,604	11,760,950	12	11,760,938	+132,334
2	8,911,601	49,430	8,862,171	8,997,558	0	8,997,558	+135,387
3	7,711,129	32,005	7,679,124	7,792,947	0	7,792,947	+113,823
4	8,029,942 (85,155 ^b)	65,717	7,964,225	9,397,177 (1,384,836 ^b)	0	9,407,501	+1,432,952
Totals	36,346,967	212,843	36,134,124	37,948,632	12	37,958,944	+1,824,820

^aNumber of total bases minus 'N' bases (bp) used in difference comparison

^bApproximate total length of repeating sequence at the carboxyl end of chromosome 4

an AT region starting at 8.97 Mbp was extended from 3 kbp to 57 kbp, in chromosome 2 at 3.27 Mbp a 3 kbp region was extended to 65 kbp, in chromosome 3 at 5.35 Mbp a 4 kbp region was extended to 56 kbp, and in chromosome 4 at 5.0 Mbp a 2 kbp region was to 61 kbp. The centromere position quoted is directly related to the AT rich positions originally noted in the MIPS version, see Fig. 1. These results indicate that each centromere has an asymmetric position in each chromosome. Therefore each of the four *F. graminearum* chromosomes is submetacentric with respect to centromere positioning. The evidence supporting these regions are centromere sequences was tested by the Connolly et al. study [27]. Where there was centromeric sequence present, which was only in chromosome 3 in the MIPS version of the reference, then H3K9me3 binding was observed but no H3K27me3, see Additional file 1.

Annotation of the completed *Fusarium graminearum*

The revised gene call

From a total of 13,826 mRNA transcripts presented in the MIPS v3.2 genome most were transferred to the RRes v4.0 genome. However, sixteen were deleted due to sequence changes, two were redundant isoforms and removed (see Additional file 2), 63 gene models were removed due to merging and five gene models were split. In addition, gene models that had alternative model predictions which were longer were explored using InterProScan5 and manually corrected where applicable. The gene models that either did not 100 % match after transfer or had been revised by extension, splitting or merging have been given the ID extension “_M”, representing modification. Some of the most common types of changes made are described in the following three generic examples: (1) An extended or reduced gene, for example, FGSG_17172 (hereafter referred to only as 17172 would become 17172_M, (2) splitting of a gene into two ORFs results in the new gene designations 17172_A_M and 17172_B_M, and (3) joining of a gene with other gene models i.e. 17172 and 17173, was designated as 17172_3_M keeping only the unique digit for the second gene if sequential ID's or if not sequential then the full gene ID for each was used i.e. 17172 and 16222 becomes 17172_16222_M. These changes to the gene names were done to signify that these gene models had been changed from the previous MIPS v3.2 annotation but this approach also ensured the retention of the original gene ID. In total, 1529 modified gene models are presented in the RRes v4.0 genome annotation. Also, 412 new gene models have been predicted of which 152 have supporting RNA-seq evidence from cultured mycelium or spore transcriptome assemblies. These have arisen because of the replacement of the ‘N’ sequences or alternative gene prediction software. The new gene

models have been given sequential numbers from 20,000, beginning at the amino end of chromosome-1 and ending at the carboxyl end of chromosome-4. The Maker annotation software was used to transfer gene annotation ID's, predict gene models to correct MIPS gene models v3.2 and add additional new genes using as evidence, RNA-seq for the mycelium and spores of PH-1 [EMBL:SRP039087], Z-3639 mycelia harvested 24 h after inoculation of wild-type conidia in complete medium RNA-seq [EMBL: SAMN02400310], proteins from Uniprot from *Fusarium* with evidence, and Cogeme EST transcripts [28]. All evidence and final gene model predictions are available through the Ensembl fungi tracks as “*Fusarium graminearum*”. The total number of gene models identified in the completed *F. graminearum* genome is 14,164 which is in contrast to 13,322 and 13,826 in the BROAD and MIPS v3.2 annotations respectively (see Table 3). A comparison of the length of proteins between the RRes and MIPS v3.2 annotations shows proportionally the total percentage of proteins with a length below 149 aa, and between 150 and 999 aa to increase and decrease respectively for the RRes set (see Additional file 3). This figure is a reflection of the greater proportion of genes that are predicted as a single exon in the RRes v4.0 vs the MIPS v3.2 set (see Additional file 4).

The gene annotation

An analysis of the RRes v4.0 annotation using Blast2GO with the NCBI nr database and InterProScan5 revealed a total of 12,691 genes that were Blast2GO annotated or had blast hits, whilst 5,442 genes have no GO annotation (38.4 %). There are an additional 386 level-2 GO biological process GO annotations and 159 molecular function GO annotations in the RRes v4.0 annotation compared to the MIPS v3.2 (Table 4 and Table 5). The number of proteins with alpha/beta hydrolase fold-1 or

Table 3 Basic statistics of the different reference and annotation versions of *F. graminearum*. A comparison of *F. graminearum* genome version statistics between BROAD, MIPS, and RRes

	BROAD	MIPS	RRes
Genome size (bp) ^a	36,565,771	36,553,761	38,060,440
Scaffolds ^b	31	19	5
GC (%) content ^c	48.3	48.3	48.2 (48.0 ^d)
Spanned gaps	402	402	1
Predicted Genes	13,322	13,826	14,164
Repetitive (%)	0.24	0.24	0.24
Transposable elements (%)	0.029	0.029	0.060
ENA project accession	PRJNA13839	N/A	PRJEB5475

^aincluding all scaffolds, N bases, and the mitochondria

^bincluding all scaffolds, N bases excluding the mitochondria

^cExcluding N's and mitochondria

^dExcluding N's, mitochondria and large repetitive sequence at the carboxyl end of chromosome

Table 4 GO term (level 2) summary of the 412 new gene annotations. GO terms summaries for biological process and molecular function of the 412 new genes annotated in the RRes V4.0 gene set

Biological process		
GO-id	GO-term	No.
GO:0044699	single-organism process	18
GO:0050896	response to stimulus	2
GO:0008152	metabolic process	32
GO:0009987	cellular process	22
GO:0032502	developmental process	2
GO:0071840	cellular component organization or biogenesis	2
GO:0065007	biological regulation	10
GO:0051179	localization	4
GO:0032501	multicellular organismal process	2
GO:0023052	signalling	1
Total		99
Molecular function		
GO-id	GO-term	No.
GO:0005215	transporter activity	3
GO:0001071	nucleic acid binding transcription factor activity	8
GO:0004872	receptor activity	1
GO:0003824	catalytic activity	25
GO:0005488	binding	24
GO:0030234	enzyme regulator activity	1
Total		64

Zn(2)-C6 fungal-type DNA-binding domain have increased from 21 to 81 and 316 to 358, respectively (Table 6). The greatest increases for biological processes were in the categories; biological regulation, single-organism process, cellular process, metabolic process, and for molecular functions; catalytic activity, binding, and nucleic acid binding transcription factor activity.

The GO annotation assigned to the new genes are detailed in Table 4 (for full BLAST and GO annotations see Additional file 5). The new genes given enzyme designation codes are listed in Table 7. Some of the new genes potentially have roles in primary metabolism, secondary metabolism or cell signalling. Some newly predicted genes of particular interest due to coding for metabolic pathway enzymes or orthologous in virulence mechanisms are as follows: FGRRES_20327 is predicted to have alpha-N-arabinofuranosidase activity (enzyme code: 3.2.1.55), is located in an average % GC region of approximately 2 kb in length neighboured by two AT rich regions of approximately 9 kb length at the carboxyl end of chromosome 3. FGRRES_20406 and FGRRES_20294 are both serine threonine protein kinases (enzyme code: 2.7.11). FGRRES_20377 is predicted to be a cutinase (enzyme code: 3.1.74).

Table 5 Level 2 GO terms of the modified gene annotations (_M) between RRes and the MIPS counterparts. GO summaries were extracted using Blast2GO

Biological process					
GO-id	GO-term	RRes No.	MIPS No.	Diff No.	
GO:0032501	multicellular organismal process	8	3	5	
GO:0048511	rhythmic process	2	1	1	
GO:0022610	biological adhesion	7	7	0	
GO:0051704	multi-organism process	6	3	3	
GO:0051179	localization	165	150	15	
GO:0008152	metabolic process	617	539	78	
GO:0023052	signalling	30	26	4	
GO:0000003	reproduction	4	4	0	
GO:0009987	cellular process	482	428	54	
GO:0050896	response to stimulus	72	56	16	
GO:0065007	biological regulation	152	112	40	
GO:0002376	immune system process	1	0	1	
GO:0040007	growth	1	1	0	
GO:0044699	single-organism process	408	357	51	
GO:0032502	developmental process	10	5	5	
GO:0022414	reproductive process	3	1	2	
GO:0040011	locomotion	1	0	1	
GO:0071840	cellular component organization or biogenesis	52	41	11	
Total		2021	1734	287	
Molecular function					
GO-id	GO-term	RRes	MIPS	Diff	
GO:0004872	receptor activity	6	3	3	
GO:0060089	molecular transducer activity	7	5	2	
GO:0001071	nucleic acid binding transcription factor activity	79	55	24	
GO:0005198	structural molecule activity	21	20	1	
GO:0005085	guanyl-nucleotide exchange factor activity	2	2	0	
GO:0009055	electron carrier activity	5	5	0	
GO:0016209	antioxidant activity	5	5	0	
GO:0030234	enzyme regulator activity	4	4	0	
GO:0005488	binding	514	491	23	
GO:0003824	catalytic activity	488	449	39	
GO:0005215	transporter activity	75	72	3	
Total		1206	1111	95	

FGRRES_20330 is predicted to be related to the enzyme galactoside o-acetyltransferase (enzyme code: 2.3.1). FGRRES_20264 is predicted to be involved in the inositol phosphate metabolism pathway, (enzyme code: 2.7.1.67). FGRRES_20189 is predicted to be involved in pyruvate metabolism Acetyl-CoA to homocitrate and

Table 6 InterProscan5 domain annotation comparison between RRes v4.0 versus MIPS v3.2. The number of proteins identified with protein domains associated with InterProscan ID's that are linked to virulence genes. Those in bold are increased in the RRes v4.0 set

Parent Inter ID	Child Interpro ID	RRes no. protein	MIPS no. protein
IPR017853	Glycoside hydrolase superfamily	122	122
IPR029058	Alpha/Beta hydrolase fold	286	N/A
	IPR000675 cutinase	13	12
	IPR001031 thioesterase	5	5
	IPR000383 Xaa-Pro dipeptidyl-peptidase-like domain	4	4
	IPR000073 Alpha/beta hydrolase fold-1	81	21
	IPR001375 Peptidase S9, prolyl oligopeptidase, catalytic domain	9	9
	IPR002018 Carboxylesterase, type B	26	26
	IPR002921 Fungal lipase-like domain	8	8
	IPR002925 Dienelactone hydrolase	8	8
	IPR003140 Phospholipase/carboxylesterase/thioesterase	5	5
	IPR013094 Alpha/beta hydrolase fold-3	30	31
	IPR029059 Alpha/beta hydrolase fold-5	12	N/A
IPR003439	ABC transporter-like	62	62
IPR011050	Pectin lyase fold/virulence factor	33	33
	IPR000070 Pectinesterase, catalytic	3	3
IPR004835	Chitin synthase	12	9
IPR001138	Zn(2)-C6 fungal-type DNA-binding domain	358	316
IPR001128	Cytochrome P450	114	114
IPR011701	Major facilitator superfamily	250	248
IPR015433	Phosphatidylinositol kinase	3	2
IPR015500	Peptidase S8, subtilisin-related	28	29
IPR001283/IPR014044	Cysteine-rich secretory protein, allergen V5/Tpx-1-related/CAP domain	5	5 ^a
IPR011329	Killer toxin	4	4

^aOne of these proteins was only partial**Table 7** New enzyme's from the new RRes gene set. New enzyme codes designated to the 412 new RRes V4.0 gene set

Gene ID	Enzyme code	Enzyme Name or description
20327	EC:3.2.1.55	Non-reducing end alpha-L-arabinofuranosidase
20406	EC:2.7.11	Transferring phosphorous-containing groups
20377	EC:3.1.74	Cutinase
20330	EC:2.3.1	Acytransferases
20264	EC:2.7.1.67	1-phosphatidylinositol 4-kinase
20189	EC:2.3.3.14	Homocitrate synthase
20172	EC:1.14.12	Oxidoreductases
20056	EC:1.14.11	Oxidoreductases
20292	EC:3.4.23	Aspartic endopeptidases
20373	EC:1.14.11	Oxidoreductases
20101	EC:1.1.1.158	UDP-N-acetylmuramate dehydrogenase

lysine biosynthesis of 2-oxoglutarate to homocitrate (enzyme code: 2.3.3.14). Finally, FGRRES_20176 has a pathogenicity protein description and GO transporter activity.

An additional 19 new enzymes have been identified from the modified RRes gene model set, see Table 8. In total, 287 new biological process level GO terms have been identified and 95 new molecular function level 2 GO terms were found (Table 5). Of the modified genes, two genes were of particular interest in terms of virulence, FGRRES_17235_M, because this gene originally had no functional annotation but was extended to be identified as coding for a cysteine rich secretory protein, allergen V5/Tpx-1-related with CAP and signal peptide domains, which were previously linked with plant pathogenesis proteins of the PR-1 family [29] and identified in the highly virulent *F. graminearum* strain CS3005 (gene ID: FG05_09548). The second gene 15917_M was identified as an endo-1,4-beta-xylanase enzyme acting on plant cell walls by endohydrolysis of (1->4)-beta-D-xylosidic linkages in xylans, also identified as part of the secretome.

Table 8 New enzyme codes from the modified RRes gene set. Enzymes codes identified in the modified RRes gene subset (_M) not found in the MIPS annotation

Gene ID	Enzyme code	Enzyme Name
06098_M	EC:6.1.1.20	Phenylalanine-tRNA ligase
08061_M	EC:4.1.1.49	Phosphoenolpyruvate carboxykinase (ATP)
01195_M	EC:3.6.3.8	Calcium-transporting ATPase
07202_M, 11727_M	EC:3.6.1.15	Nucleoside-triphosphate phosphatase
15917_M	EC:3.2.1.8	Endo-1,4-beta-xylanase
16359_B_M	EC:3.1.4.11	Phosphoinositide phospholipase C
10980_M	EC:2.7.7.49	RNA-directed DNA polymerase
13343_M	EC:2.7.1.30	Glycerol kinase
09688_M	EC:2.5.1.18	Glutathione transferase
16945_M	EC:2.3.1.225	Protein S-acyltransferase
13161_M	EC:2.2.1.2	Transaldolase
08613_M	EC:2.1.1.71	Phosphatidyl-N-methylethanolamine N-methyltransferase
15680_M	EC:1.3.1.9	Enoyl-[acyl-carrier-protein] reductase (NADH)
06127_M	EC:1.2.1.2	Formate dehydrogenase
15680_M	EC:1.13.12.16	Nitronate monooxygenase
04922_M	EC:1.1.1.9	D-xylulose reductase
03873_M	EC:1.1.1.158	UDP-N-acetylmuramate dehydrogenase
03648_M	EC:1.1.1.184	Carbonyl reductase (NADPH)

For nine genes from the MIPS v3.2 annotation, which were in regions of unknown sequence an improved level of annotation was possible for one gene. FGRRES_16611 previously classified as a hypothetical protein of 329 amino acid (aa) length but was extended to 437 aa and identified as an UAA transporter. The remaining eight corrected gene models could not be assigned additional functionality.

A total of 504 RNA annotations were found across the four fully assembled chromosomes including 319 tRNA,

62 rRNA, 121 ncRNA, and 2 tmRNA see Table 9. In addition a further 538 rRNA were identified in the extensive repetitive region present at the carboxyl of chromosome 4. A non-coding RNA 4.5S RNA that has sequence similarity to a Plant_SRP (RFAM: RF01855) and Fungi_SRP (RFAM:RF01502) [30] of 1.00 E-4 and 2.00 E-31, which is the only one identified in the genome at position 5,836,584 bp on chromosome 3, may be important to the effectiveness of the secretome and pathogenicity. Signal recognition particles (SRP) are universally conserved ribonucleoproteins that direct the traffic of proteins within the cell allowing them to be secreted [31] and therefore are a potentially important mechanism for virulence.

To complete the genome annotation, the remaining RRes v4.0 supercontig 3.12 was found to contain four genes including an additional new gene of 34 aa length (FGRRES_20410). The only gene to have InterProscan or BLASTP ($E < 0.01$) annotation was FGRRES_14025 which contains an intron endonuclease, group 1 domain and is a member of the GIY-YIG nuclease superfamily [32], typically found predominantly in phage and mitochondria. This supercontig may represent phage DNA rather than mitochondrial due to the lack of alignment with the mitochondrial DNA. In addition to annotation using four gene predictors, open reading frame annotation of this contig showed no additional proteins that matched within the NCBI database therefore there are unlikely to be any small genes in addition to these four predicted genes.

Deoxynivalenol (DON) produced by *Fusarium graminearum* is a potent B-type trichothecene mycotoxin [33] and virulence factor to disease formation in wheat floral tissue [34]. Commonly found in harvested cereal grains, DON lowers grain quality and is a serious health concern to humans and animals [35]. DON synthesis has been assigned to 11 genes (Tri1FGRRES_00071, Tri3 FGRRES_03534, Tri4 FGRRES_03535, Tri5 FGRRES_03537, Tri7A FGRRES_02115, Tri7B FGRRES_03533, Tri8

Table 9 The number of coding and non-coding gene models predicted in the RRes and MIPS versions. Summary of coding and non-coding annotations of both the RRes and MIPS annotations across chromosomes 1-4 and supercontigs

Chromosome	RRes gene no.	MIPS gene no.	RNA RRes (total)	tRNA RRes	rRNA RRes	ncRNA RRes	tmRNA RRes
1	4393	4303	118	74	9	34	1
2	3645	3549	156	97	32	27	0
3	3090	2999	121	87	8	25	1
4 ^a	3032	2969	99	61	13	25	0
4 ^b (1,384,836 bp)	0	0	538	0	538	0	0
Supercontig_3.12	4	3	10	0	0	10	0
Supercontig_3.15	n/a	3	n/a	0	0	0	0
Total	14,164	13,826	504 (538 ^a)	319	62 (538)	121	2

^aWithout carboxyl end repeating units with RNA annotation

^bcarboxyl end repeating units with RNA annotation

FGRRES_03532, Tri101 FGRRES_07896, Tri11 FGRRES_03540, Tri13A FGRRES_03860, Tri13B FGRRES_10629 and four regulators Tri6 FGRRES_16251, Tri10 FGRRES_03538, Tri12 FGRRES_15001, and Tri12 FGRRES_11025 [36]. Changes to gene structures of these and other mycotoxin associated sequences may affect previous or future studies involving gene deletion/gene silencing studies. Within the RRes v4.0 gene annotations the Tri-gene cluster situated on chromosome 2, Tri3 FGRRES_03534_M 15-O-acetyltransferase protein domain has increased from 391 aa to 413 aa.

Other less prevalent mycotoxins produced by *F. graminearum* include the estrogenic polyketide zearalenone and nivalenol (NIV), which is structurally related to DON. The *F. graminearum* PH-1 strain lacks functional copies of the Tri13 and Tri7 genes resulting in the inability to produce NIV. Zearalenone induces hyperoestrogenic responses in mammals and can result in reproductive disorders in farm animals, therefore it is of agricultural importance to limit levels in the harvested grain. Zearalenone synthesis requires four genes in a cluster on chromosome 1, namely FGRRES_15980, FGRRES_17745, FGRRES_15982, and FGRRES_02398 [37]. In the Zearalenone gene cluster, the PS-DH domain of the protein 15980_M has increased from 145 aa to 299 aa and the ketoacy-synt-c domain has increase from 115 aa to 117 aa.

Although the DON production pathway is well studied, the regulatory mechanisms are less well characterised [38] and unknowns such as the oxygenation step involving converting the C8 hydroxyl to a carbonyl group and the deacetylation of C15 have not been assigned an enzyme [39]. A greater understanding of genetic factors regulating toxin biosynthesis could lead to the development of new strategies aimed at reducing toxin accumulation in the host tissue. In the MIPS annotation there are eight gene predictions with aflatoxin biosynthesis regulatory protein annotations with two that have transcription factor domains. Although aflatoxin is not known to be produced by *F. graminearum* there role(s) may be associated with another mycotoxin pathway such as DON. Two modified gene predictions from the RRes v4.0 set that are annotated as aflatoxin biosynthesis regulatory proteins and one that also has a transcription factor domain (FGRRES_12018_M and FGRRES_03294_M respectively) are additions to this small clade of functionally unknown regulatory genes.

Analysis of the predicted secretome

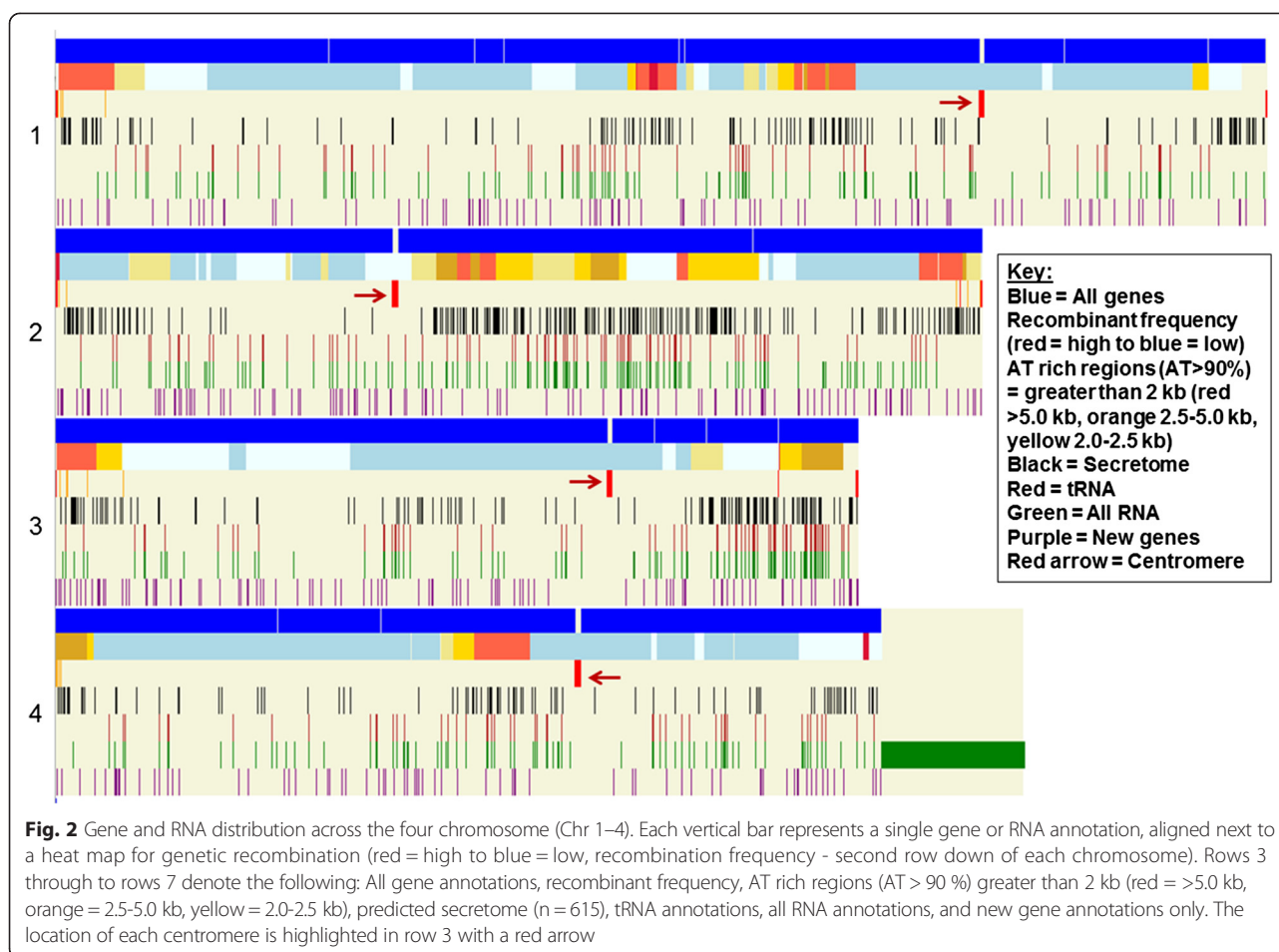
The secretome of a fungus is the set of secreted proteins that frequently defines and facilitates the host infection process [40]. During a fungal pathogens interaction with a host, the invading fungal cells secrete various proteins that can block host responses, i.e. the fungus *Ustilago maydis* secretes Cmu1 which can move through plant cells and redirect plant metabolic pathways to favour

fungal infection [41], or kill host cells i.e. for *F. graminearum*, 109 genes coding for secreted cell wall degrading enzymes have previously been identified [42]. Characterisation of the secretome is fundamental to the understanding of the mechanisms of host infection and pathogen virulence.

Using the MIPS v3.2 assembly and annotation in 2012, Brown et al. [42] predicted the secretome to contain 574 proteins (575 including a protein without an 'M' start codon). Using the new RRes v4.0 predicted protein set and updated versions of the software we obtain 616 proteins using the Brown et al. pipeline to represent the refined secretome (see Additional file 6 for a complete gene ID list and detailed methodology). 144 of the RRes predictions were confirmed using proteomics from the Yang et al. [43] and Paper et al. [44] studies. 30 of the 70 proteins identified in the Yang et al. study from proteins identified in *F. graminearum* culture supernatants after growth in wheat or barley flour-containing medium, and 137 of the 286 from the extracellular proteins *in vitro* and *in planta* from the pathogenic fungus *F. graminearum* identified in the Paper et al. study were confirmed in the *in silico* predictions. Of the 616 proteins predicted, 55 are from modified genes and 12 are from newly predicted genes. In total, 515 protein ID's of the newly refined secretome set match the original predictions. Whereas 59 secreted proteins are unique to Brown et al. and 101 unique to RRes v4.0 (see Additional file 7). Of the complete predicted secretome, 119 proteins have some level of annotation whilst 497 proteins lack annotation and 73 proteins (11.85 %) of the newly refined secretome have a cysteine content of > 5 %. Although genomic distribution of the secretome is not altered, an overview diagram of the secretome genes and tRNA annotated genes shows an overlap of clustering, in particular on chromosome 2 and 3, See Fig. 2.

Analysis of genes predicted to possess nucleic acid binding transcription factor activity

Proteins that bind to specific DNA sequences and regulate transcription of genes are termed transcription factors. They represent an important family of proteins controlling activities of the organism that regulate traits important for growth, development, stress response, pathogenesis, and toxin production, therefore are of importance to the development of fungicides for intervention. Within the RRes v4.0 gene set we have modified 25 gene annotations of the MIPS set and predicted 7 new genes (See Additional file 8). The modified set of gene annotations shows 18 gene models that lacked a Zn(2)-C6 fungal-type DNA-binding domain in the MIPS annotation to have been extended to include it in the RRES v4.0 set, and 15 of these have been identified as having Gal4 protein domain and therefore are potential positive



regulators for the gene expression of galactose-induced genes. Of the new set of gene annotations predicted with transcription factor activity: FGRRES_20168 is annotated as having oxidoreductase activity, FGRRES_20126 is annotated as a component of a protein kinase CK2 complex, and FGRRES_20239 is annotated as having lipid glycosylation activity which is part of a family of transferases involved in the final stages of the biosynthesis of antibiotics vancomycin and related chloroeremomycin.

Genes located in close proximity to the telomere and centromere sequences

The centromeres are the region of the chromosomes that determine kinetochore formations. Each centromere enables pairing of the sister chromatids and interacts with spindle microtubules during mitotic chromatid segregation and homologous chromosome segregation during meiosis. Centromere structure in *F. graminearum* and more generally in eukaryotes is poorly understood [27]. For eukaryotes it was suggested that centromeres may be derived from subtelomeric regions [45]. The four centromeres identified in *F. graminearum* chromosome 1-4 respectively are 56,581, 65,181, 56,387, 60,933 bp

long and are marked by their similar high AT content, 87.2, 86.4, 87.2, and 87.9 % respectively. Chromosome 3 is the only chromosome to have an additional $\sim \frac{1}{4}$ sized AT rich region not located within the proximity of the subtelomeres, at position 7,011,437-7,026,387 bp (possible a neocentromere), of 14,951 bp length and an AT content of 80.9 %. Here we surveyed the functional annotation of neighbouring genes of the newly integrated centromere, neocentromere and subtelomeric sequences in *F. graminearum* (Results are given in Additional file 9). Although no overall pattern was found within 40 kb regions in proximity to the centromeres or the amino and carboxyl telomere sequences, two genes in common between these regions were found on the chromosome 4. These two genes (FGRRES_17709_M and FGRRES_08956_M) possess a NB-ARC signalling domain (PF00931) implicated in programmed cell death [46].

Genes and other features associated with the proposed ancient chromosome fusion sites

In previous studies, *F. graminearum* was shown to contain only four chromosomes [4, 18], whilst in related *Fusarium* species a considerably higher numbers of

chromosomes exist. For example, for *F. verticilloides* the presence of 11 chromosomes, and for *F. oxysporum*, 15 chromosomes were demonstrated using a whole-genome sequencing and comparative genomics approach [7]. Ma et al. [7] used the whole-genome alignment of *F. graminearum* and *F. verticilloides* to show that the 11 sequenced *F. verticilloides* chromosomes could be mapped to the four *F. graminearum* chromosomes and proposed that the smaller number of chromosomes in *F. graminearum* was the result of a series of ancient chromosome fusion events. The identified potential fusion points matched regions of high genomic diversity (Ma et al, Supplementary Fig. S3 [7]). Closer inspection of the centromeres in the RRes v4.0 reference sequence revealed repeat sequences of either MarCry-1_FO or LTR-Gypsy to embody a large proportion of the sequence content and also the AT rich regions within the subtelomeric loci. The fusion event proposed by Ma et al that led to *F. graminearum* centromeres were likely a result of fusion at these AT rich regions, which are found at various loci within each chromosome of *F. verticilloides*. This hypothesis is supported by repeating the alignment by Ma et al. but using the RRes v4.0 genomic reference sequence and focusing upon the centromere regions of *F. graminearum* that were not identified in the MIPS reference. Regions that aligned contained a number of annotated and non-annotated genes sequentially homologous in sequence and context. Different flanking sequences of AT rich regions of *F. verticilloides* from the same chromosomes align to flanks of the centromeres found in *F. graminearum* (see Additional file 10). The observation of different flanking sequences adjacent to AT rich regions from the same chromosome aligning to *F. graminearum* chromosome centromeres flanking sequences, suggests that fusing of multiple AT rich sequences to make the centromeres in *F. graminearum* did not include exclusively ancient centromere sequences either in part or full but incorporated other AT rich regions into the recombination event(s). This analysis provides further evidence in support of the existence of ancient chromosome fusion events in the *F. graminearum* genome which were not a direct result of the maintenance of ancient centromeres.

Genes previously tested for a role in the disease causing ability of *F. graminearum*

The Pathogen-Host Interactions database (PHI-base) stores molecular and biological information on pathogen genes for which mutant phenotypes in pathogen-host interactions have been tested experimentally [47]. PHI-base currently provides information on > 2,800 pathogen genes, involving ~160 pathogenic species. There were *F. graminearum* 69 entries, including 59 putative transcription factors [48] and 8 kinases [49], an ATP citrate lyase

ACL1 [50] and a syntaxin-like SNARE gene GzSYN1 [51] from PHI-base that had modified gene sequences in the RRes v4.0 gene set versus MIPS v3.2 (see Additional file 11). The kinases and transcription factor genes were analysed in high-throughput gene deletion studies for their virulence [48, 49]. Five of these genes were essential genes in *F. graminearum*, four genes when mutated caused reduced virulence, and one gene loss of pathogenicity. The remaining 59 genes were neither implicated in virulence nor essential for life. Due to changes in the gene model a re-evaluation of the function of each gene is recommended.

Recombinant regions

Telomeres are established in many species as sites of high Single Nucleotide Polymorphism (SNP) frequency and structural rearrangement [52–54]. *F. graminearum* was shown to exhibit this trait but was also found to show interstitial regions of high diversity [4]. This was suggested to represent ancient telomeres as a consequence of chromosomal fusion of progenitor chromosomes. The mapping of markers used to establish these recombinant regions to the RRes v4.0 genome confirms these results with no observable changes in marker locations. The identification of the centromeres in the RRes genome and mapping of previously established markers shows that the high recombinant loci at the telomeres is on the long arm of the chromosomes, see Fig. 2. However, it is worth noting that markers are lacking for the carboxyl ends of the short arms of chromosome 1 and 3 due to new sequence inserted for the RRes v4.0 genome and may still show these regions to be highly recombinant in future studies. A comparison of total genes, secretome predicted genes, unique genes, and AT content by chromosomal arm reveals no pattern differences between areas of high or low recombination, (see Additional file 12). However it is clear that the secretome genes are predominantly found at areas of high recombination and within the subtelomeric regions, see Fig. 2.

Repeat motifs and transposons

In the original BROAD analysis, the *F. graminearum* genome was found to contain very few transposon sequences annotated with gene ID's and none were considered to be biologically active [4]. Active elements such as transposons are of interest because these are important contributors to genome evolution, including the creation of novel host genes and disruption of existing genes. The gene annotations identified with transposon protein domains are shown in Additional file 13. Of particular note was the new gene annotation, FGRRES_20409 is a gypsy-type long-terminal-repeat (LTR) retrotransposon, not previously included in the MIPS v3.2 assembly and the only

retroviral transposon gene annotation found. RNA-seq supporting expression data in both the mycelium and spore was found suggesting this transposon is actively transcribed. However it was not found to have been included within any gene annotation leading to a truncation of the protein coding sequence. The sequence has been previously reported [GenBank:XM_389127] as a hypothetical protein partial mRNA in another PH-1 sequencing and annotation project but appears to have been an automated process with no reference in print. A comparison of centromere and subtelomeric AT rich regions repeat content shows a LTR gypsy sequence or MarCry-1_FO represented in each of these regions (see Additional file 14).

This study confirms previous findings in other species demonstrating the presence of transposon sequence in the centromeres [55]. The potential neocentromere identified in chromosome 3 shows a higher proportion of transposon content to the chromosome 3 centromere of 62.5 and 53 % respectively. In contrast to the subtelomeric regions there is a general trend towards a lower transposon content which further supports the hypothesis of a neocentromere at this location.

BLAST comparisons

The original BLASTP analysis of the BROAD *F. graminearum* against *F. asiaticum*, *F. boothii*, *F. culmorum*, and *F. pseudograminearum* identified 704 species specific genes [4]. A similar BLASTP analysis (evaluate 0.1) using high quality *Fusarium* genome sequences found on Ensembl fungi including: *F. culmorum*, *F. fujikuroi*, *F. oxysporum*, *F. verticilloides*, and *Nectria haematococca* using the RRes v4.0 genome annotations identified 741 species specific genes (see Additional file 15). Visualisation of these species specific genes loci shows no observable clustering (see Additional file 16).

BLASTP against the *Fusarium* cereal infecting subset, *F. culmorum*, *F. verticilloides*, and *F. fujikuroi* identified 769 unique genes. A conservative evaluate of 0.1 for excluding proteins as part of the clade was used as little difference was observed as this value decreased. Comparison of the *Fusarium* and cereal clade species specific genes within the predicted RRes v4.0 secretome identified six genes that were *F. graminearum* unique and contain signal peptide sequences. Translated BLASTN of these six secreted *F. graminearum* unique sequences were explored against all *Fusarium* taxon entries in NCBI 12/12/2014 to test if the identifications were a result of missing or incorrect gene annotations. This analysis revealed one, FGRRES_20407, to have a hit below evaluate 0.1 in both *Nectria haematococca* and *F. fujikuroi* and therefore this sequence is not unique to *F. graminearum*. The remaining five secreted sequences identified as unique to *F. graminearum* are; FGRRES_02309,

FGRRES_15048, FGRRES_12652, FGRRES_15251, and FGRRES_20027.

The *F. graminearum* isolate genomes CS3005 and CS3096 are incomplete. This somewhat hinders inter-strain comparative analysis. However, our analysis of *F. graminearum* isolates CS3005 and CS3096 has revealed a total of 263 and 473 proteins, respectively identified as unique compared to *F. graminearum* PH-1. However a combinational analysis using complete genome sequences to identify SNP/Indel polymorphisms, expression analysis differences and unique proteins between isolates and the RRes v4.0 reference is required to identify the underlying reasoning behind these isolates genomic and phenotypic differences.

The original BROAD analysis of the *F. graminearum* genome had identified very little evidence of gene duplication and had discovered an active RIP mechanism that had eliminated duplicated sequences. An integral protein of the RIP mechanism is the presence of a DNA methyltransferase RIP defective gene (RID) [Swiss-Prot: Q8NJW0], reported in *Neurospora crassa* [56]. BLASTP of the proteome of *F. graminearum* versus the RID gene from *N. crassa* reveals one protein, FGRRES_08648, of similar identity (380 aa alignment length of 42.0 %) but with an additional DNA methylase C-5 cytosine-specific active site. This protein is hypothesised to represent the RID gene in *F. graminearum* and further support the activity of RIP. The transposon class TcMar-Fot1, FGRRES_15950, FGRRES_17263, FGRRES_16975, FGRRES_13431_M are all single exon predictions but for FGRRES_13431_M which is multi-exon. A previous analysis by Cuomo et al. has shown RIP induced SNPs in this class of transposons however an alignment was not reported. A BLASTN alignment of the region of FGRRES_13431_M versus the other TcMar-Fot1 class of transposons reveals that the transposon has had two stop codons introduced which explains the multi-exon prediction from the gene software. FGRRES_16975 has also a stop codon introduced causing a truncation. In addition to the induced SNPs between these transposons we can report that this has led to stop codon introduction, likely terminating the functionality of the transposons discussed and additional evidence to the process of RIP in *F. graminearum*.

A reciprocal BLASTP of the *F. graminearum* proteome identified two duplicated protein sequences, one of which had 100 % nucleotide identity. The first is a Tigger transposon (FGRRES_15950 and FGRRES_17263) at position 6,157,128 chromosome 1 and position 5,484,124 chromosome 4 respectively which had identical nucleotides. The second is the histone H4 protein FGRRES_20411 (to be re-designated as FGRRES_04289 in RRes v4.1) and FGRRES_05491_M with 20 SNPs: C:G = 1, C:A = 1, T:G = 6, G:A = 12 (see Additional file 17). The sequence of FGRRES_05491_M is longer than FGRRES_20411 with an

additional 62 aa at the amino end of the protein which includes a signal peptide of 20 aa which is predicted to target the protein to the mitochondria. Within the RRes v4.0 annotation there are two H3 proteins, one H1, one H1/H5, three H2 (A, B, C), but three H4 proteins of which two have the aforementioned 100 % matching sequences present.

In a further analysis, we explored gene families based upon protein sequence similarity which grouped sequences into no more than four using a similar methodology to Klenk et al. and Heidelberg et al. [57, 58]. The average family size is 2.09 and the largest 18 families identified included 6 families of amino acid transporters, three major facilitator families, three transposon families, three tyrosinases (pigment formation) families, one aldehyde/histidinol dehydrogenase family, one unknown function family and a family of nucleoside triphosphate hydrolases. The remaining families were no greater than 2. This suggests that few genes have formed via multi-duplications arising from a single progenitor gene to form these small multigene families.

Discussion

The *de novo* genome assembly using *de Bruijn* assembler software, such as SOAPdenovo2 [59], requires the use of a k-mer value which requires a set sequence length in order to align and join sequenced reads but this is limited to the length of the reads i.e. 100 bp Illumina reads. As the k-mer value rises, this increases the specificity but reduces sensitivity. Therefore contigs of a repetitive nature such as AT rich regions within centromeres, have a higher chance of assembling into a single sequence as the k-mer value increases. There is a limitation to this approach, because as the k-mer value is increased greater sequencing coverage is needed to compensate for the reduced sensitivity. Using 100 bp Illumina reads and a ~85 x genome coverage, a k-mer value of approximately 61 resulted in the highest N50 assembly. The presence of AT rich regions or repetitive transposons loci are likely to cause breaks in the resulting contigs [60]. Using a greater k-mer value can in some cases create longer sequence contigs for these loci and potentially result in contigs that span these repetitive regions along with the non-repetitive flanks. However combining such assemblies can be manually time consuming for genomes with many repetitive loci and genomes larger than 36 Mbp. The multi k-mer approaches like those that are available to SOAPdenovo2 can combine these two scenarios to produce a superior assembly with little manual curation. We were successful in using this approach to fully assemble the *F. graminearum* genome. For genomes of a similar size but contain more repetitive sequences, this strategy might be less successful but is worthwhile investigating. For *F. graminearum* the

multi-kmer approach resulted in the insertion of previously unknown sequences, identification of a significant mis-assembly in chromosome 1 at the amino end of the MIPS v3.2 genome, extending several internal AT rich regions, placing 12 supercontigs within chromosomes, and for the first time identifying for each of the four chromosomes the telomeric sequence (TAACCC) and centromeric DNA.

The analysis of centromeric DNA in filamentous fungi is difficult due to the fact that centromeric regions are comprised of heterogenous, repetitive, AT-rich sequences [61]. Identification of the size and position of the centromeres on each of the four *F. graminearum* chromosomes permitted an examination of the annotations within the immediate vicinity versus annotation within the rest of the chromosome arms. However comparison of total gene content, the secretome gene subset, *F. graminearum* unique genes, and GC content revealed no obvious patterns (see Additional file 18). In *Saccharomyces cerevisiae* short point centromeres of 120 bp exist [62]. In contrast in many other fungi including *Schizosaccharomyces pombe*, the human pathogenic fungus *Candida albicans* and in the filamentous fungus *N. crassa* regional centromeres of considerable length are prevalent [63]. In *C. albicans*, regional centromeres are composed of unique 3–4.5 kb sequences [64]. In the case of *N. crassa*, centromeres are composed of degenerate transposons, retrotransposons and simple sequence repeats, with no gene sequences. Our BLAST analysis ($e < 0.00001$) confirmed the absence of genes or distinctive DNA elements for the four centromeric regions in *F. graminearum*. The centromeric composition, despite its AT rich >90 %, was sufficiently varied to assemble using the four re-sequencing data sets we had available with an insert distance of between 408 and 449 and the use of k-mer values >81, see Additional file 19. The average size of the four centromeres was found to be 59,771 bp, and for chromosome 3 our analysis revealed the existence of an additional smaller AT-rich region of 15 kb. This region is approximately $\frac{1}{4}$ the size of the designated centromeric region and may comprise a neocentromeric region or the remainder of centromere repositioning event. The genes localised immediately adjacent to the four centromeres did not include genes implicated in virulence. Several sequences with predicted functions in programmed cell death and chromosome segregation, replication and chromatin remodelling were found to be located in close proximity to the centromeres. We propose that the availability of centromeric sequence information for *F. graminearum* may be exploited in several ways. (1) Although no evidence exists, they may help construct artificial chromosomes. Currently none are available for filamentous fungi. (2) Assist in cancer research. In several types of human cancers, centromeres

were found to be deleted and kinetochores instead assemble on neocentromeres. Experiments in both *S. pombe* and *C. albicans* were successful in deleting centromeric regions and identifying neocentromeres. (3) To identify fungal specific drug targets. Kinetochore proteins binding to the centromeric regions in filamentous fungi during mitosis were proposed to be highly fungal-specific and provide pertinent specific drug targets [61]. Binding studies between fungal specific proteins and centromeric DNA regions may reveal suitable small-molecule drugs.

The annotation of the genome was comprehensive. First we took the BROAD (FG3) genome and the gene models provided by MIPS (v3.2). Then we included additional annotation from a range of different sources to provide evidence for new gene models and to modify existing gene models. This additional annotation evidence was obtained by using RNA-seq data from the mycelium and spores of strain PH-1 [EMBL:SRP039087] as well as the mycelium of strain Z-3639 harvested 24 h after inoculation of wild-type conidia in complete medium RNA-seq [EMBL:SAMN02400310]. In addition *Fusarium* protein data was taken from Uniprot and we also used the extensive EST libraries generated by Sanger sequencing available from the Cogeme EST transcripts database [27]. As a consequence, a total 412 new gene models were predicted and 1529 gene models have been modified.

The new 412 gene predictions and curated gene models are a significant step towards a fully annotated genome. However further manual curation is required by the community using RNA-seq supporting evidence from multiple biological situations to achieve the high level of novel transcriptome information needed to complete this important activity. Currently, this process is organised using the public Ensembl Fungi websites [22] and the phytopathogenic specific website, PhytoPathdb [23]. Custom tracks can be displayed against the completed *F. graminearum* alongside any new gene model annotations and all changes are shared via a mailing list. Group members can contribute, view and discuss updated gene models prior to being incorporated into future Ensembl Fungi versions (see Additional file 20 for instructions).

The study of the *F. graminearum* gene repertoire predicted to code for secreted proteins is highly significant to research aimed at understanding the pathogenic process due to their hypothesised roles in penetration, host tissue necrosis and host immune subversion [65]. The rapid rearrangement in these regions may promote rapid evolution of secretory genes in response to host responses leading to species specific genes and virulence hot spots. For example, in the *Aspergillus* species, species unique secondary metabolite genes are enriched in subtelomeric regions [66, 67]. The only similarly

secreted and *F. graminearum* unique protein that has a cysteine content of above 5 % (9 %) is FGRRES_15251 and is located in the subtelomeric region of chromosome 2, at 8,527,286 bp. These characteristics are a hallmark of an effector protein in other filamentous pathogenic species and of potential interest to gene deletion studies. The reconfirmation of the identification of secretory genes and the identification of new secretory genes within these subtelomeric and the high recombinant regions previously demonstrated by Ma et al. and Brown et al. supports this hypothesis. The identification of a predominance of the tRNA annotations within the genes predicted to code for secreted proteins, not previously noted, may be hypothesised to represent an evolutionary bias for improved response to stimulus such as host detection resulting in a rapid response to infection and an overall switch to the pathogenic process.

The detailed analysis of the newly predicted *F. graminearum* secretome, revealed two new secretory and one non-secretory proteins found in the RRes v4.0 predictions of potential importance to virulence and also serves to illustrate the significance of the complete genome and high quality annotations to the conclusions of research findings going forward. FGRRES_20327 is annotated as secreted and possessing alpha-N-arabinofuranosidase activity and is located in a new coding sequence island approximately 2 kb within two AT rich regions of approximately 9 kb length at the carboxyl end of chromosome 3. Whilst FGRRES_20176 is annotated as a secreted pathogenicity protein with transporter activity. Both genes were previously only identified in the NGS genome study of the high virulence *F. graminearum* Australia isolate CS3005 (FG05_30019 and FG05_35058, respectively). A third example, is the new cysteine-rich potential virulence gene (FGRRES_17235_M) previously only reported in the highly virulent strain CS3005 of *F. graminearum* which contains several features characteristic of many apoplastic fungal effectors shown to inhibit plant chitinases used to prevent pathogen associated molecular pattern triggered immunity thereby preventing induction of host defences. The previous identification of these proteins in CS3005 but not PH-1 was therefore the result of the incompleteness of the PH-1 genome.

The reanalysis of the well-studied *Tri* and zearalenone gene clusters proved to be interesting. A previous study had observed the co-expression of three flanking genes, namely 03531 (OrfA), 03530 (OrfB) and 3529 with the core *Tri* genes and extended the list of genes involved in trichothecene DON production and maintenance [68]. In the current study, one gene was modified in the *Tri* gene cluster and in the zearalenone gene clusters. However with the publically available RNA-seq data and manual curation of other newly identified genes, then it may be possible to

identify and associate other sequences to these two key pathways and iteratively improve their curation and annotation. These results highlight the need for re-appraisal of existing functional data to inform the current protein function studies on these two important toxin pathways.

The absence of genes that arose from duplications in *F. graminearum* reflects the activity of a process found in many fungi [69] namely, repeat-induced point mutation (RIP) in which mutations are caused in duplicated sequences [55]. This process is hypothesised to be a defence mechanism against transposable elements which are common to viral/bacterial infection. A previous study has shown the introduction of C:G to T:A transition mutations during the sexual cycle in *F. graminearum* transposons [4]. The identification of two histone H4 proteins of identical protein content but different DNA sequences suggests these two sequences may be the consequence of a recent duplication event which is in the process of RIP. However the introns of the proteins are substantially different and SNP C:G or T:A transitions are not predominant in the shared coding regions. This suggests that these H4 proteins containing identical coding sequences but for an additional sequence including a signal peptide that is predicted to target the mitochondria for FGRRES_05491_M, may not be in the process of RIP but is present due to important functional reasons currently unknown.

The large repeating rRNA containing sequence at the carboxyl end of chromosome 4 is similarly found in the human genome on chromosome 21 with adjacent unknown bases. It is highly likely to exist in other genome reference sequences at the end of a chromosome interspersed with N bases and/or as an unplaced contigs. By manual investigation such features were found in the genome reference sequences (Ensembl Fungi v23) of *F. fujikuroi* and *F. oxysporum* f. sp. *lycopersici* at the respective amino and carboxyl ends of chromosome 2. In *F. solani* such sequences were found as unscaffolded contigs which are likely to originate from the amino or carboxyl end of a chromosome sequence yet unplaced. This feature may be common among the *Fusarium* clade and beyond but requires a greater number of genomic reference sequences to be completed before this feature can be thoroughly explored.

The genome and annotation sequences have been deposited into ENA and displayed at Ensembl fungi and with the dedicated plant pathogen PhytoPathdb database to permit public access to the data and tracking of versions. Information can be linked to the genome from other databases, such as the Pathogen-Host Interaction entries [47] to efficiently access phenotypic information from pathogenicity testing involving single and multiple gene deletion isolates generated mostly in the reference strain PH-1. The presentation in Ensembl fungi/PhytoPathdb will aid comparative genomics studies by the

identification of homologous and orthologous genes with other representative fungal genomes which are continually increasing and being updated.

Conclusion

Using whole shotgun re-sequencing of *F. graminearum* we have comprehensively completed the genomic sequence by replacing unknown bases, placing supercontigs within their correct loci, correcting assembly errors, and inserting new sequences which include for the first time complete AT rich sequences such as centromere sequences and subtelomeric regions. The identification of a retroviral transposon on chromosome 4 shows a potential route that *F. graminearum* is continuing to evolve in particular to environmental stresses, such as fungicides, and diverge from related fungal species.

An analysis of 412 new gene models and 1529 curated gene models reveals 7 new and 25 modified transcription regulator protein predictions, three new targets for virulence mechanisms, and includes a range of proteins that may require further re-testing as demonstrated by the 59 entries with modified sequences reported in PHI-base which are reported as not being virulently associated. All these changes will greatly improve the analysis of gene function and the reinterpretation of the considerable number of transcriptomics data sets deposited by the international *Fusarium* community which are available at PLEXdb [70]. The redefined subset of genes predicted to comprise the refined secretome, verifies the previously recognised secretome predominantly within previously identified hotspot location for this gene type in the genome, namely the subtelomeric and highly recombinatorial regions. The *F. graminearum* genome can now be comprehensively studied with the knowledge that a full set of gene models and surrounding genomic sequence can be probed in combination with non-coding RNA annotation to identify pathogenicity mechanisms.

In summary, we present a fully completed *F. graminearum* isolate PH-1 genome and manually curated annotation using RNA-seq evidence which provides the optimum resource to perform interspecies comparative analyses and gene function studies using various reverse genetics approaches. The availability of the RRes v4.0 genome should also permit in the near future a detailed inter-comparison of historic and modern *F. graminearum* isolates collected from diverse geographical locations, plant species as well as fungicide application and crop rotation regimes.

Availability of data and materials

Sequence Read Archive (SRA) accession numbers for the four cells of Illumina sequencing are [EMBL:ERS430784, EMBL:ERS430785, EMBL:ERS430786, EMBL:ERS430787]. The RRes genomic and mitochondrial sequence for the

F. graminearum strain PH-1 is available at the European Nucleotide Archive [EMBL:HG970330, EMBL:HG970331, EMBL:HG970332, EMBL:HG970333, EMBL:HG970334, EMBL:HG970335]. An OmniMap instance of the RRes v4.0 FG genome which includes files to reproduce images contained herein is available for download [71].

Methods

Fungal growth, DNA preparation and sequencing

The fungal strain PH-1 (ATCC MYA-4620/FGSC 9075/NRRL 31084) was cultured as described [72]. Genomic fungal DNA for sequencing was extracted using the CTAB protocol [73] and purified using a Qiagen Kit (Qiagen Ltd, Crawley, West Sussex, UK). High-quality genomic DNA was then submitted to the genome analysis centre (TGAC, Norwich, UK) for generation of a 0.8-kb fragment library. The Illumina HiSeq 2000 sequencing platform (San Diego, CA) was used to produce 100-bp paired-end reads [74].

Assembly and alignment

No pre-processing of reads took place. The software Novoalign3 (version 3.01.02) and BWA (version 0.7.5a-r405) were used to align the sequencing reads, with default parameters, to the *F. graminearum* reference sequence (MIPS version FG3.2 [75]). Alignments were converted from the sequence alignment map (SAM) format to binary alignment map (BAM), and the BAM files were sorted and indexed using SAMtools (version 0.1.19). Visualisations were done using Tablet [76] (version 1.13.07.31).

For *de novo* assembly, the software SOAPdenovo2 (version 2.0.4) was used with different k-mer values: 61, 63, 71, 81, and 91. The resulting assemblies were used to correct and extend MIPS version FG3.2 sequence and orientation where possible. Reference sequence statistics were extracted from Tablet and Geneious (version 6.1 created by Biomatters). LASTZ (version 1.02.00) was used from within Geneious to align genomic sequences.

Genome annotation

The assembled genome was annotated using the MAKER (version 2.30) [21] annotation pipeline with RepeatMasker (version 4.0.5) [77]. A *F. graminearum* specific repeat library was constructed using RepeatModeler (version 1.0.7) and supplied to MAKER for the repeat masking step. Gene calls were generated using FGENESH (version 3.1.2) [78] using the Fusarium matrix, AUGUSTUS (version 2.7) [79] using *F. graminearum* species model, GeneMark [80] and SNAP [81], which was trained using all the Fusarium proteins in UNIPROT with the keyword "Fusarium" which had evidence, ESTs from Cogeme [28], and trinity assemblies using RNA-seq from the mycelium and spores of PH-1, and wild type Z-3639. This evidence was provided to MAKER as hints to the annotation. The final annotation set produced by MAKER and manually curated to include

the MIPS v3.2 annotation is summarized in Table 9. Non-coding RNA were identified using default settings with both tRNAscan-SE-1.3.1 [82] and Infernal-1.1 [83]. InterProscan-5.7-48.0 (analyses: TIGRFAM-13.0, ProDom-2006.1, SMART-6.2, HAMAP-201311.27, SignalP-EUK-4.0, PrositePatterns-20.97, PRINTS-42.0, SuperFamily-1.75, Panther-9.0, Gene3d-3.5.0, SignalP-GRAM_POSITIVE-4.0, PIRSF-2.84, SignalP-GRAM_NEGATIVE-4.0, PfamA-27.0, PrositeProfiles-20.97, Phobius-1.01, TMHMM-2.0c, Coils-2.2) was used to determine whether to correct a gene model based upon new or extended protein domains.

Gene statistics, interproscan domain, GO and enzyme comparisons

Blast2GO V.2.7.2 was used with Decypher BLASTP search's with an E-value of 0.001 against the NCBI nr database from 04/09/14 and 15/09/14 NCBI nr database for the 412 new gene and modified protein sets respectively, filtered using Blast2GO annotation algorithm with settings, E-value filter 0.000001, Annotation CutOff 55, GO weight 5, Hsp-Hit Coverage CutOff 0, and GO and enzyme code annotated using a local GO database from 07/2014 with 41,436 GOs available and 4,098 Enzymes available. Interproscan results were imported in to Blast2GO and the GO annotations merged. Annotation statistics were produced using Eval V.2.2.8 and Geneious.

BLAST comparisons

Unique proteins from the annotations of RRes V4.0, and the *F. graminearum* isolates CS3005 and CS3096 were assigned using BLASTP with an evaluate 0.1 using either: 1) Fusarium subset which included; *F. culmorum*, *F. fujikuroi*, *F. oxysprum f.sp. lycopersici* (4287), *F. verticilloides*, *Gibberella moniliformis*, *Nectria haematococca*, or 2) Fusarium cereal infecting subset including; *F. culmorum*, *F. verticilloides*, *F. fujikuroi*.

Secretome identification

InterProscan-5.7-48.0 was used to identify signal and transmembrane domains. Proteins smaller than 20 amino acids were excluded. ProtComp (Version 9.0) [84] and result columns, LocDB and PotLocDB were used to exclude GPI anchored membrane proteins and other non-extracellular loci proteins. WoLFPSort [85] (extracellular score > 17, WoLFPSort) was used to identify final destination and big-PI [86] to further remove GPI-anchored proteins.

SNP calling

BWA-MEM [87] was used to align the sequencing reads to the BROAD *F. graminearum* (FG3) reference using default parameters, followed by conversion to BAM format and sorting by coordinate using SAMtools (v0.1.19-44428 cd) [88]. SNPs and indels were called using

SAMtools mpileup followed by VarScan (v2.3.7) using default settings.

Additional files

Additional file 1: Figures showing H3K9me3 binding in the MIPS reference with alignments of RRes and MIPS reference to highlight where centromeric sequence was present in the Connolly et al. study.

Additional file 2: A table listing the gene models in RRes v4.0 deleted from the MIPS annotation.

Additional file 3: A table of RRes v4.0 and MIPS v3.2 binned protein lengths.

Additional file 4: A table of RRes v4.0 and MIPS exons per transcript.

Additional file 5: A table listing the BLAST and GO annotations of the RRes v4.0 new gene prediction subset.

Additional file 6: A description and table listing the RRes secretome gene ID's and those unique to the earlier Brown et al. [42] vs. RRes v4.0 secretome.

Additional file 7: A figure of RRes v4.0 secretome set and RRes and Brown et al. unique secretome gene prediction comparison.

Additional file 8: A table of Gene ID's with GO term 0001071 which are identified as having nucleic acid binding transcription factor activity from either altered gene predictions or genes that were not found in the MIPS v3.2 annotation.

Additional file 9: A table of distal centromere, possible neocentromere and telomere gene ID's and annotations.

Additional file 10: A table of the alignment positions of flanking sequences of AT rich regions from the *F. verticillioides* chromosomes to the sequences flanking the centromeres in RRes v4.0 *F. graminearum*.

Additional file 11: A table of RRes v4.0 gene models that have changed since MIPS v3.2 that are present in PHI-base.

Additional file 12: A table of RRes v4.0 *F. graminearum* recombinant and non-recombinant region comparison.

Additional file 13: A description of RRes repeat masking details and a table of gene ID's identified with transposon domains.

Additional file 14: A table of transposon and repeat content of centromeres and subtelomeric AT rich regions.

Additional file 15: A table of *Fusarium graminearum* species specific genes identified in the RRes v4.0 gene annotation.

Additional file 16: A figure representing chromosomal locations of *Fusarium graminearum* species specific gene ID's identified in the RRes v4.0 gene annotation.

Additional file 17: A figure of the nucleotide alignment of genomic sequence of duplicate containing H4 protein sequences FGRRES_20411 (FGSG_04289) and FGRRES_05491_M.

Additional file 18: A table of RRes v4.0 all genes, secretome subset genes, and unique gene sets per chromosome arm.

Additional file 19: Shows the mapped read coverage across the centromeres and the corresponding GC content.

Additional file 20: A set of instructions for how to stay updated and contribute to *F. graminearum* PH-1 gene models changes prior to future Ensembl Fungi version releases.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

MU prepared the DNA for sequencing. RK performed the genome assembly, annotation, and all other bioinformatics. RK performed the manual curation of Chromosome 1-2, 4, while MHK performed the manual curation of Chromosome 3. KHK, MU, RK and KHP contributed to experimental design, data evaluation,

writing of the manuscript and approved the final manuscript. All authors read and approved the final manuscript.

Acknowledgement

We thank Munich Information Services for Protein Sequences (MIPS) for making their gene models of *Fusarium graminearum* publicly available for others to use. We thank the staff at The Genome Analysis Centre (TGAC), Norwich, UK, for their excellent technical assistance. We thank the staff at the European Nucleotide Archive (ENA), Cambridge, UK, in particular Dr Ana M. Cerdeño-Tarraga for her assistance in submission of *F. graminearum* to the public repository. This research was supported by the Biotechnology and Biological Sciences Research Council of the UK (BBSRC) through the Institute Strategic Programme Grant 20:20 Wheat (BB/J/00426X/1). In addition, Kim Hammond-Kosack and Martin Urban are supported by the BBSRC National Capability project grant (BB/J/004383/1) and the BBSRC PhytoPath project grant (BB/K020056/1).

Received: 30 March 2015 Accepted: 7 July 2015

Published online: 22 July 2015

References

- Swain MT, Tsai IJ, Assefa SA, Newbold C, Berriman M, Otto TD. A post-assembly genome-improvement toolkit (PAGIT) to obtain annotated genomes from contigs. *Nat Protoc.* 2012;7(7):1260–84.
- Gross SS, Do CB, Sirota M, Batzoglu S. CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. *Genome Biol.* 2007;8(12):R269.
- Srivastava SK, Huang X, Brar HK, Fakhoury AM, Bluhm BH, Bhattacharyya MK. The genome sequence of the fungal pathogen *Fusarium virguliforme* that causes sudden death syndrome in soybean. *PLoS One.* 2014;9(1):e81832.
- Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, et al. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science.* 2007;317(5843):1400–2.
- BROAD INSTITUTE. Fungal Genomics. 2015. <http://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/fungal-genome-initiative>. Accessed 01 July 2014.
- Jeong H, Lee S, Choi GJ, Lee T, Yun SH. Draft genome sequence of *Fusarium fujikuroi* B14, the causal agent of the Bakanae disease of rice. *Genome Announc.* 2013. 1(1):e00035–00013.
- Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature.* 2010;464(7287):367–73.
- Moolhuijzen PM, Manners JM, Wilcox SA, Bellgard MI, Gardiner DM. Genome sequences of six wheat-infecting *Fusarium* species isolates. *Genome Announc.* 2013. 1(5):e00670-13-e00670-13.
- Gardiner DM, Stiller J, Kazan K. Genome Sequence of *Fusarium graminearum* Isolate CS3005. *Genome Announc.* 2014. 2(2):e00227-14.
- Jiang D, Zhu W, Wang Y, Sun C, Zhang KQ, Yang J. Molecular tools for functional genomics in filamentous fungi: recent advances and new strategies. *Biotechnol Adv.* 2013;31(8):1562–74.
- Goswami RS, Kistler HC. Heading for disaster: *Fusarium graminearum* on cereal crops. *Mol Plant Pathol.* 2004;5(6):515–25.
- Lee T, Oh DW, Kim HS, Lee J, Kim YH, Yun SH, et al. Identification of deoxynivalenol- and nivalenol-producing chemotypes of *Gibberella zeae* by using PCR. *Appl Environ Microbiol.* 2001;67(7):2966–72.
- Lysoe E, Klemsdal SS, Bone KR, Frandsen RJ, Johansen T, Thrane U, et al. The PKS4 gene of *Fusarium graminearum* is essential for zearalenone production. *Appl Environ Microbiol.* 2006;72(6):3924–32.
- Desjardins AE, Hohn TM, McCormick SP. Trichothecene biosynthesis in *Fusarium* species: chemistry, genetics, and significance. *Microbiol Rev.* 1993;57(3):595–604.
- Proctor RH, Hohn TM, McCormick SP. Reduced virulence of *Gibberella zeae* caused by disruption of a trichothecene toxin biosynthetic gene. *Mol Plant Microbe Interact.* 1995;8(4):593–601.
- Wu F, Guclu H. Aflatoxin regulations in a network of global maize trade. *PLoS One.* 2012;7(9):e45151.
- Urban M, Hammond-Kosack K. Molecular genetics and genomic approaches to explore *Fusarium* infection of wheat floral tissue. In: *Fusarium Genomics and Molecular and Cellular Biology* (Proctor, R.H. and Brown, D., eds), Chapter 12 Norwich, Norfolk, UK: Horizon Scientific Press; 2013. p. 43–79.

18. Gale LR, Bryant JD, Calvo S, Giese H, Katan T, O'Donnell K, et al. Chromosome complement of the fungal plant pathogen *Fusarium graminearum* based on genetic and physical mapping and cytological observations. *Genetics*. 2005;171(3):985–1001.
19. Wong P, Walter M, Lee W, Mannhaupt G, Munsterkotter M, Mewes HW, et al. FGDB: revisiting the genome annotation of the plant pathogen *Fusarium graminearum*. *Nucleic Acids Res*. 2011;39(Database issue):D637–9.
20. Guldener U, Mannhaupt G, Munsterkotter M, Haase D, Oesterheld M, Stumpflen V, et al. FGDB: a comprehensive fungal genome resource on the plant pathogen *Fusarium graminearum*. *Nucleic Acids Res*. 2006;34(Database issue):D456–8.
21. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics*. 2011;12:491.
22. Ensembl Fungi. 2015. <http://fungi.ensembl.org/>. Accessed 24 Jan 2015.
23. PytoPath. 2015. <http://www.phytopathdb.org/>. Accessed 24 Jan 2015.
24. Blackburn EH. Structure and function of telomeres. *Nature*. 1991;350(6319):569–73.
25. Meyne J, Ratliff RL, Moyzis RK. Conservation of the human telomere sequence (TTAGGG)_n among vertebrates. *Proc Natl Acad Sci U S A*. 1989;86(18):7049–53.
26. Huson DH, Auch AF, Qi J, Schuster SC. MEGAN analysis of metagenomic data. *Genome Res*. 2007;17(3):377–86.
27. Connolly LR, Smith KM, Freitag M. The *Fusarium graminearum* histone H3 K27 methyltransferase KMT6 regulates development and expression of secondary metabolite gene clusters. *PLoS Genetics*. 2013. 9(10):e1003916.
28. Soanes DM, Skinner W, Keon J, Hargreaves J, Talbot NJ. Genomics of phytopathogenic fungi and the development of bioinformatic resources. *Mol Plant Microbe Interact*. 2002;15(5):421–7.
29. Dixon DC, Cutt JR, Klessig DF. Differential targeting of the tobacco PR-1 pathogenesis-related proteins to the extracellular space and vacuoles of crystal idioblasts. *EMBO J*. 1991;10(6):1317–24.
30. Shen K, Wang Y, Hwang Fu YH, Zhang Q, Feigon J, Shan SO. Molecular mechanism of GTPase activation at the signal recognition particle (SRP) RNA distal end. *J Biol Chem*. 2013;288(51):36385–97.
31. Bovia F, Strub K. The signal recognition particle and related small cytoplasmic ribonucleoprotein particles. *J Cell Sci*. 1996;109(Pt 11):2601–8.
32. Dunin-Horkawicz S, Feder M, Bujnicki JM. Phylogenomic analysis of the GYL-YIG nuclease superfamily. *BMC Genomics*. 2006;7:98.
33. Proctor RH, McCormick SP, Alexander NJ, Desjardins AE. Evidence that a secondary metabolic biosynthetic gene cluster has grown by gene relocation during evolution of the filamentous fungus *Fusarium*. *Mol Microbiol*. 2009;74(5):1128–42.
34. Hallen-Adams HE, Wenner N, Kuldau GA, Trail F. Deoxynivalenol biosynthesis-related gene expression during wheat kernel colonization by *Fusarium graminearum*. *Phytopathology*. 2011;101(9):1091–6.
35. Pestka JJ, Smolinski AT. Deoxynivalenol: toxicology and potential effects on humans. *J Toxicol Environ Health B Crit Rev*. 2005;8(1):39–69.
36. Urban M, Willighagen E, Chichester C, Kutmon M. WikiPathways. DON mycotoxin biosynthesis (*Gibberella zeae*). 2014. <http://www.wikipathways.org/index.php/Pathway:WP2258>. Accessed 01 Jul 2014.
37. Kim YT, Lee YR, Jin J, Han KH, Kim H, Kim JC, et al. Two different polyketide synthase genes are required for synthesis of zearalenone in *Gibberella zeae*. *Mol Microbiol*. 2005;58(4):1102–13.
38. Gardiner DM, Kazan K, Manners JM. Novel genes of *Fusarium graminearum* that negatively regulate deoxynivalenol production and virulence. *Mol Plant Microbe Interact*. 2009;22(12):1588–600.
39. McCormick SP, Harris LJ, Alexander NJ, Ouellet T, Saparno A, Allard S, et al. Tri1 in *Fusarium graminearum* encodes a P450 oxygenase. *Appl Environ Microbiol*. 2004;70(4):2044–51.
40. Brown NA, Hammond-Kosack KE. Secreted biomolecules in fungal plant pathogenesis, in fungal biomolecules: sources, applications and recent developments, V. K. Gupta, S. Sreenivasaprasad, and Robert L. Mach, Editor. John Wiley & Sons, Ltd, Chichester, UK; 2015.
41. Molloy S. Fungal physiology. *Ustilago takes control*. *Nat Rev Microbiol*. 2011;9(12):832–3.
42. Brown NA, Antoniw J, Hammond-Kosack KE. The predicted secretome of the plant pathogenic fungus *Fusarium graminearum*: a refined comparative analysis. *PLoS One*. 2012;7(4):e33731.
43. Yang F, Jensen JD, Svensson B, Jorgensen HJ, Collinge DB, Finnie C. Secretomics identifies *Fusarium graminearum* proteins involved in the interaction with barley and wheat. *Mol Plant Pathol*. 2012;13(5):445–53.
44. Paper JM, Scott-Craig JS, Adhikari ND, Cuomo CA, Walton JD. Comparative proteomics of extracellular proteins in vitro and in planta from the pathogenic fungus *Fusarium graminearum*. *Proteomics*. 2007;7(17):3171–83.
45. Villasante A, Mendez-Lago M, Abad JP, Montejo de Garcini E. The birth of the centromere. *Cell Cycle*. 2007;6(23):2872–6.
46. van der Biezen EA, Jones JD. The NB-ARC domain: a novel signalling motif shared by plant resistance gene products and regulators of cell death in animals. *Curr Biol*. 1998;8(7):R226–7.
47. Urban M, Pant R, Raghunath A, Irvine AG, Pedro H, Hammond-Kosack KE. The Pathogen-Host Interactions database (PHI-base): additions and future developments. *Nucleic Acids Res*. 2015;43(Database issue):D645–55.
48. Son H, Seo YS, Min K, Park AR, Lee J, Jin JM, et al. A phenome-based functional analysis of transcription factors in the cereal head blight fungus, *Fusarium graminearum*. *PLoS Pathog*. 2011;7(10):e1002310.
49. Wang C, Zhang S, Hou R, Zhao Z, Zheng Q, Xu Q, et al. Functional analysis of the kinome of the wheat scab fungus *Fusarium graminearum*. *PLoS Pathog*. 2011;7(12):e1002460.
50. Son H, Lee J, Park AR, Lee YW. ATP citrate lyase is required for normal sexual and asexual development in *Gibberella zeae*. *Fungal Genet Biol*. 2011;48(4):408–17.
51. Hong SY, So J, Lee J, Min K, Son H, Park C, et al. Functional analyses of two syntaxin-like SNARE genes, GzSYN1 and GzSYN2, in the ascomycete *Gibberella zeae*. *Fungal Genet Biol*. 2010;47(4):364–72.
52. Mefford HC, Trask BJ. The complex structure and dynamic evolution of human subtelomeres. *Nat Rev Genet*. 2002;3(2):91–102.
53. Gardner MJ, Hall N, Fung E, White O, Berriman M, Hyman RW, et al. Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature*. 2002;419(6906):498–511.
54. Winzeler EA, Castillo-Davis CI, Oshiro G, Liang D, Richards DR, Zhou Y, et al. Genetic diversity in yeast assessed with whole-genome oligonucleotide arrays. *Genetics*. 2003;163(1):79–89.
55. Cambareri EB, Aisner R, Carbon J. Structure of the chromosome VII centromere region in *Neurospora crassa*: degenerate transposons and simple repeats. *Mol Cell Biol*. 1998;18(9):5465–77.
56. Freitag M, Williams RL, Kothe GO, Selker EU. A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. *Proc Natl Acad Sci U S A*. 2002;99(13):8802–7.
57. Heidelberg JF, Eisen JA, Nelson WC, Clayton RA, Gwinn ML, Dodson RJ, et al. DNA sequence of both chromosomes of the cholera pathogen *Vibrio cholerae*. *Nature*. 2000;406(6795):477–83.
58. Klenk HP, Clayton RA, Tomb JF, White O, Nelson KE, Ketchum KA, et al. The complete genome sequence of the hyperthermophilic, sulphate-reducing archaeon *Archaeoglobus fulgidus*. *Nature*. 1997;390(6658):364–70.
59. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, et al. SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *GigaScience*. 2012;1(1):18.
60. Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2011;8(1):61–5.
61. Smith KM, Galazka JM, Phatale PA, Connolly LR, Freitag M. Centromeres of filamentous fungi. *Chromosome Res*. 2012;20(5):635–56.
62. Cleveland DW, Mao Y, Sullivan KF. Centromeres and kinetochores: from epigenetics to mitotic checkpoint signaling. *Cell*. 2003;112(4):407–21.
63. Meraldi P, McAnish AD, Rheinbay E, Sorger PK. Phylogenetic and structural analysis of centromeric DNA and kinetochore proteins. *Genome Biol*. 2006;7(3):R23.
64. Sanyal K, Baum M, Carbon J. Centromeric DNA sequences in the pathogenic yeast *Candida albicans* are all different and unique. *Proc Natl Acad Sci U S A*. 2004;101(31):11374–9.
65. Brown NA, Antoniw J, Hammond-Kosack KE. The predicted secretome of the plant pathogenic fungus *Fusarium graminearum*: a refined comparative analysis. *Plos One*. 2012. 7(4):e33731.
66. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglu S, et al. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. *Nature*. 2005;438(7071):1105–15.

67. Perrin RM, Fedorova ND, Bok JW, Cramer RA, Wortman JR, Kim HS, et al. Transcriptional regulation of chemical diversity in *Aspergillus fumigatus* by LaeA. *PLoS Pathog.* 2007;3(4):e50.
68. Sieber CM, Lee W, Wong P, Munsterkötter M, Mewes HW, Schmeitzl C, et al. The *Fusarium graminearum* genome reveals more secondary metabolite gene clusters and hints of horizontal gene transfer. *PLoS One.* 2014;9(10):e110311.
69. Clutterbuck AJ. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. *Fungal Genet Biol.* 2011;48(3):306–26.
70. MIAME/Plant Compliant Gene Expression Resources for Plants and Plant Pathogens. 2014. <http://www.plexdb.org/>. Accessed 01 Aug 2014.
71. King R. 2015. <https://rrescloud.rothamsted.ac.uk/public.php?service=files&t=cd70482248d1f0230d3cf6346662f663>. Accessed 01 Jan 2015.
72. Urban M, Mott E, Farley T, Hammond-Kosack K. The *Fusarium graminearum* MAP1 gene is essential for pathogenicity and development of perithecia. *Mol Plant Pathol.* 2003;4(5):347–59.
73. Doyle J. A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem Bull.* 1987;19:11–5.
74. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, et al. Accurate whole human genome sequencing using reversible terminator chemistry. *Nature.* 2008;456(7218):53–9.
75. Güldener U. MIPS version FG3.3. 2011. ftp://ftpmips.gsf.de/fungi/Fusarium/F_graminearum_PH1_v32/p3_p13839_Fus_grami_v32.scaf. Accessed 15 Jul 2014.
76. Milne I, Stephen G, Bayer M, Cock PJ, Pritchard L, Cardle L, et al. Using Tablet for visual exploration of second-generation sequencing data. *Brief Bioinform.* 2013;14(2):193–202.
77. Tempel S. Using and understanding RepeatMasker. *Methods Mol Biol.* 2012;859:29–51.
78. Solovyev V, Kosarev P, Seledsov I, Vorobyev D. Automatic annotation of eukaryotic genes, pseudogenes and promoters. *Genome Biol.* 2006;7 Suppl 1:S10 1–12.
79. Stanke M, Schoffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics.* 2006;7:62.
80. Lukashin AV, Borodovsky M. GeneMark.hmm: new solutions for gene finding. *Nucleic Acids Res.* 1998;26(4):1107–15.
81. Korf I. Gene finding in novel genomes. *BMC Bioinformatics.* 2004;5:59.
82. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 1997;25(5):955–64.
83. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics.* 2013;29(22):2933–5.
84. Softberry. 2015. www.SoftBerry.com. Accessed 24 Aug 2014.
85. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, et al. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 2007;35(Web Server issue):W585–7.
86. Eisenhaber B, Wildpaner M, Schultz CJ, Borner GH, Dupree P, Eisenhaber F. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for Arabidopsis and rice. *Plant Physiol.* 2003;133(4):1691–701.
87. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics.* 2009;25(14):1754–60.
88. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics.* 2009;25(16):2078–9.

Submit your next manuscript to BioMed Central and take full advantage of:

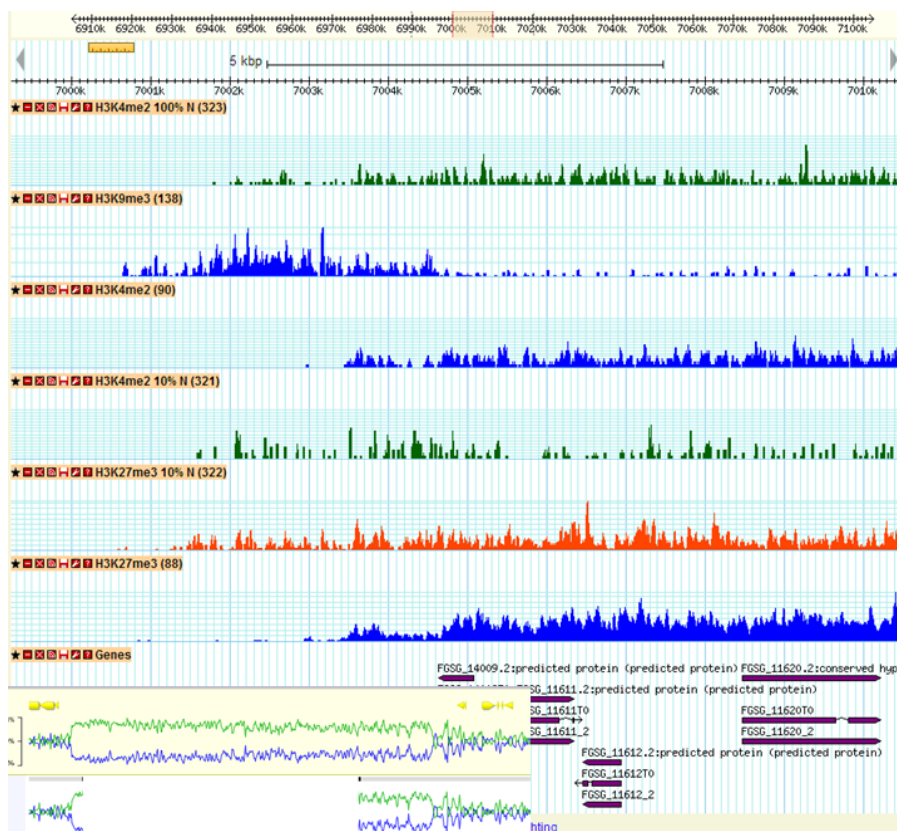
- Convenient online submission
- Thorough peer review
- No space constraints or color figure charges
- Immediate publication on acceptance
- Inclusion in PubMed, CAS, Scopus and Google Scholar
- Research which is freely available for redistribution

Submit your manuscript at
www.biomedcentral.com/submit

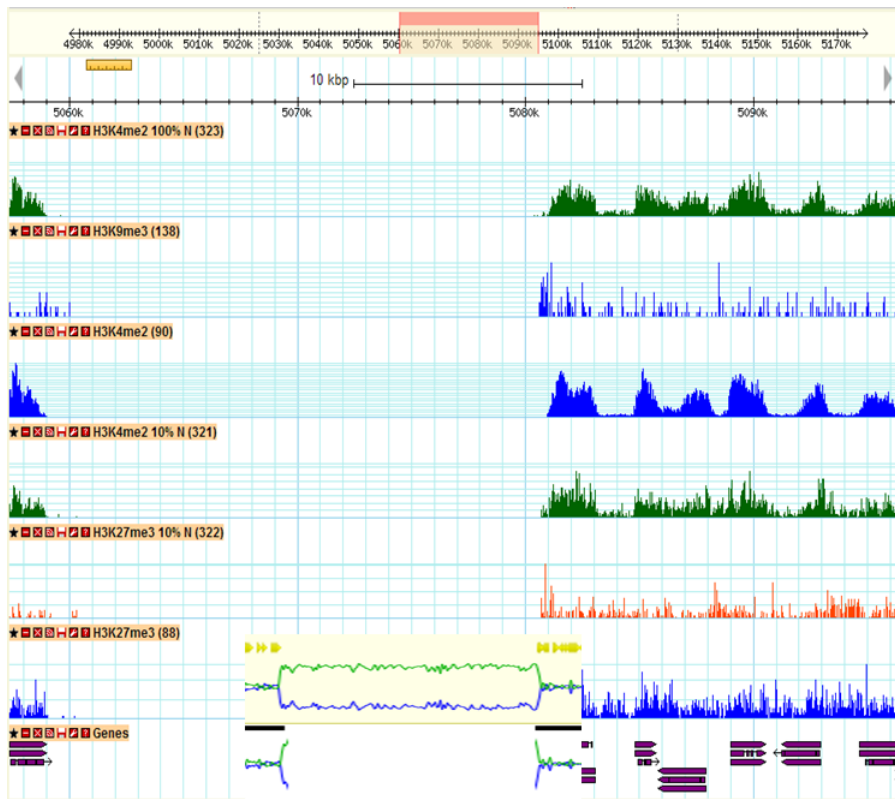


Additional file 1. Comparisons of H3K9me3 and H3K27me3 binding to the MIPS reference with an alignment superimposed of MIPS vs RRes. The only chromosome centromere to show H3K9me3 binding (normally associated with chromatin) in the MIPS reference from the *Connolly et al* study is chromosome 3 but this is explained due to chromosome 3 being the only sequence in the MIPS version that had part of the centromere sequence complete. Further studies are likely to show this.

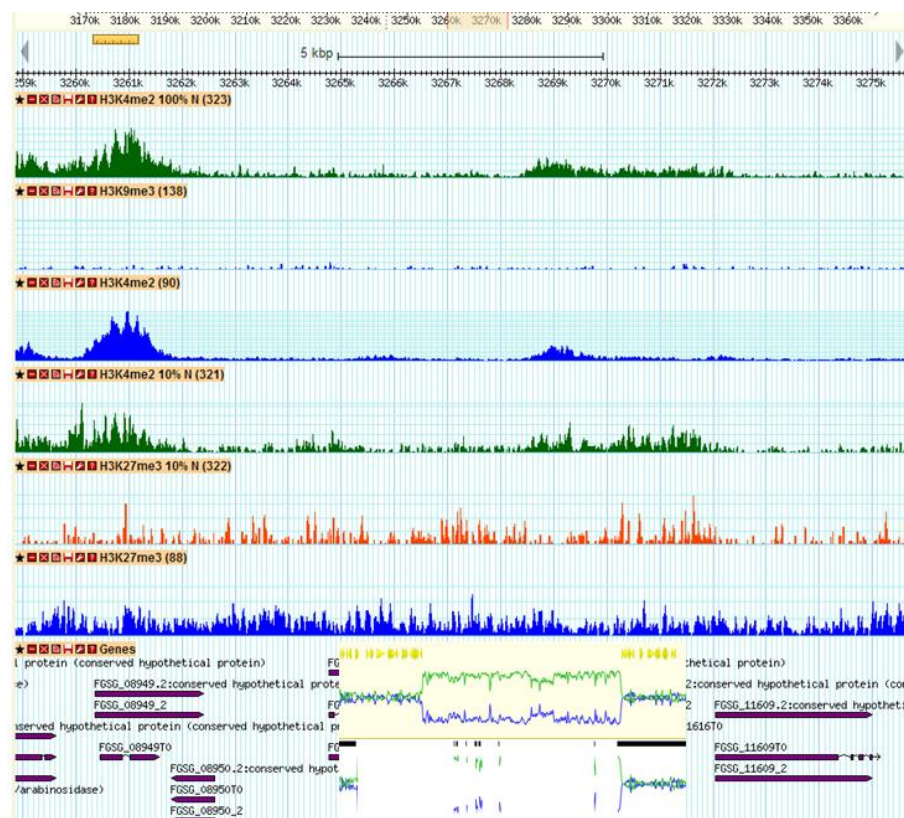
Chromosome3, alignment shows the MIPS version had part of the centromere sequence hence why see some H3K9me3 binding.



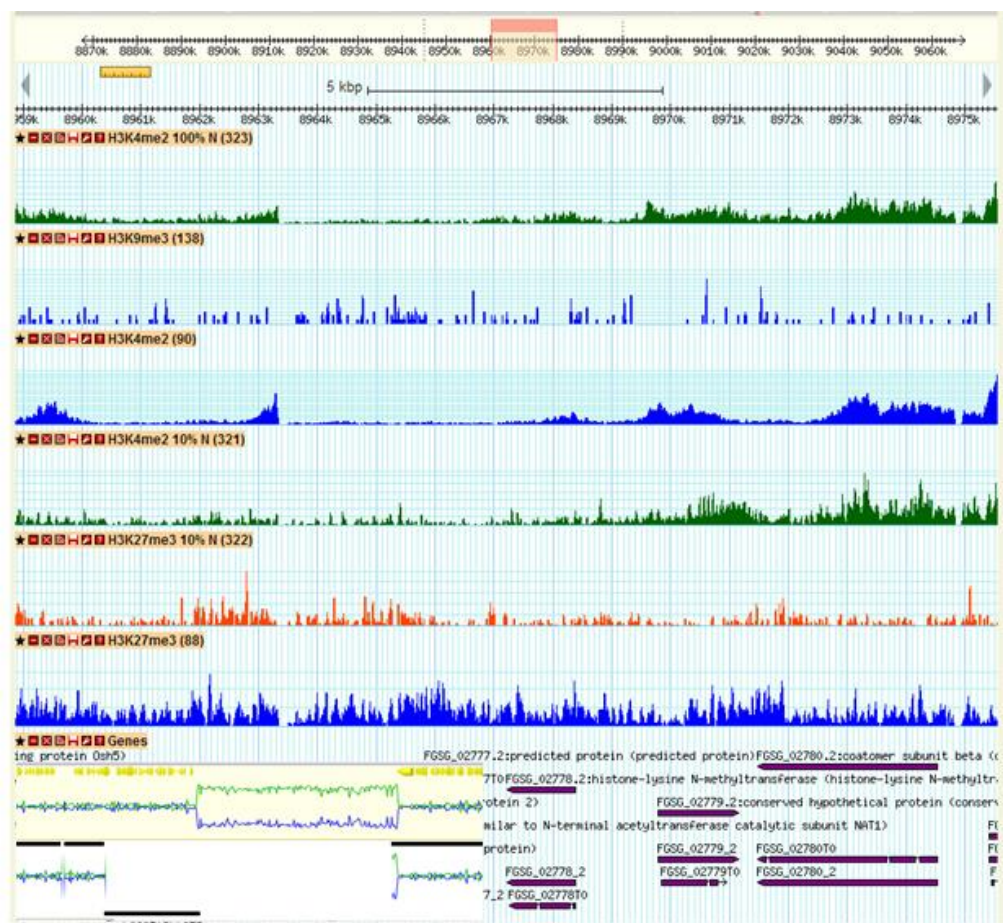
Chromosome4, alignment shows the MIPS version did not have part of the centromere sequence hence why no H3K9me3 binding is observed.



Chromosome2, alignment shows the MIPS version did not have part of the centromere sequence
hence why no H3K9me3 binding is observed.



Chromosome1, alignment shows the MIPS version did not have part of the centromere sequence
hence why no H3K9me3 binding is observed.



Additional file 2. Deleted gene models (bp length in brackets, *isoform).

Chr 1	Chr 2	Chr 3	Chr 4
FGSG_11749 (32)	FGSG_15223 (58)	FGSG_12844 (113)	FGSG_15195 (57)
FGSG_15045 (66)	FGSG_15495 (78)		FGSG_16834 (21)
FGSG_15100 (55)	FGSG_13453 (151)		FGSG_15417 (62)
FGSG_15125 (59)	FGSG_13448 (133)		
FGSG_16203 (79)	FGSG_15191 (51)		
FGSG_01500 (105)			
FGSG_17612 (1039)*			
FGSG_00674 (167)*			
FGSG_15616 (86)			

Additional file 3. Length of proteins of RRes and MIPS v3.2 sets

Length	MIPS		RRes	
	No. of Proteins	% Proteins	No. of Proteins	% Proteins
Proteins < 50 aa	35	0.30%	68	0.48%
Proteins 50-149 aa	2143	15.50%	2353	16.61%
Proteins 150-299 aa	2980	21.60%	2920	20.62%
Proteins 300-999 aa	7783	56.30%	7899	55.77%
Proteins 1000-3000 aa	861	6.20%	899	6.35%
Proteins > 3000 aa	24	0.20%	25	0.18%

Additional file 4. Exons per transcript distribution using Eval.

Total number of exons /gene model	RRes	MIPS v3.2	RRes diff	RRes % of total gene models	MIPS % of total gene models
1	3536	3250	+286	24.96	23.51
2	4206	4130	+76	29.70	29.87
3	2961	2957	+4	20.91	21.39
4	1616	1646	-30	11.41	11.91
5	857	866	-9	6.05	6.26
6	440	448	-8	3.11	3.24
7	203	210	-7	1.43	1.52
8	144	127	+17	1.02	0.92
9	90	85	+5	0.64	0.61
10	44	46	-2	0.31	0.33
11	30	27	+3	0.21	0.20
12	17	15	+2	0.12	0.11
13	6	3	+3	0.04	0.02
14	5	10	-5	0.04	0.07
15	4	3	+1	0.03	0.02
16	1	1	0	0.01	0.01
17	2	0	+2	0.01	0.00
18	0	0	0	0.00	0.00
19	1	1	0	0.01	0.01
20	0	0	0	0.00	0.00
21	1	1	0	0.01	0.01

Additional file 5. New gene predictions blast and GO ID's annotations using Blast2Go. DOCX format.

SeqName FGRRES_	Hit-Description	GO- Group	GO-ID	Term
20004	hypothetical protein FG00057.1	F	GO:0016491	oxidoreductase activity
20006	related to micromolar calcium activated neutral protease 1	C	GO:0005622	intracellular
20006	related to micromolar calcium activated neutral protease 1	P	GO:0006508	proteolysis
20006	related to micromolar calcium activated neutral protease 1	F	GO:0004198	calcium-dependent cysteine-type endopeptidase activity
20036	hypothetical protein FOVG_01478	F	GO:0004872	receptor activity
20036	hypothetical protein FOVG_01478	P	GO:0007275	multicellular organismal development
20036	hypothetical protein FOVG_01478	C	GO:0016020	membrane
20048	hypothetical protein FG05_30608	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20048	hypothetical protein FG05_30608	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20048	hypothetical protein FG05_30608	C	GO:0005634	nucleus
20048	hypothetical protein FG05_30608	F	GO:0008270	zinc ion binding
20056	epoxide hydrolase	F	GO:0016787	hydrolase activity
20056	epoxide hydrolase	P	GO:0055114	oxidation-reduction process
20056	epoxide hydrolase	F	GO:0016706	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2- oxoglutarate as one donor, and incorporation of one atom each of oxygen into both donors
20063	v-type h+- transporting atpase subunit h	F	GO:0015078	hydrogen ion transmembrane transporter activity
20063	v-type h+- transporting atpase subunit h	F	GO:0042625	ATPase activity, coupled to transmembrane movement of ions
20063	v-type h+- transporting atpase subunit h	C	GO:0000329	fungus-type vacuole membrane

20063	v-type h ⁺ -transporting atpase subunit h	P	GO:0000032	cell wall mannoprotein biosynthetic process
20063	v-type h ⁺ -transporting atpase subunit h	P	GO:0006461	protein complex assembly
20063	v-type h ⁺ -transporting atpase subunit h	P	GO:0007035	vacuolar acidification
20063	v-type h ⁺ -transporting atpase subunit h	C	GO:0000220	vacuolar proton-transporting V-type ATPase, V0 domain
20063	v-type h ⁺ -transporting atpase subunit h	P	GO:0015991	ATP hydrolysis coupled proton transport
20064	hypothetical protein FG01998.1	F	GO:0008270	zinc ion binding
20064	hypothetical protein FG01998.1	F	GO:0016832	aldehyde-lyase activity
20064	hypothetical protein FG01998.1	P	GO:0005975	carbohydrate metabolic process
20067	related to permease of the major facilitator superfamily	C	GO:0016021	integral component of membrane
20067	related to permease of the major facilitator superfamily	F	GO:0003723	RNA binding
20067	related to permease of the major facilitator superfamily	F	GO:0008270	zinc ion binding
20067	related to permease of the major facilitator superfamily	P	GO:0055085	transmembrane transport
20078	hypothetical protein FG05_30155	C	GO:0016021	integral component of membrane
20101	hypothetical protein FG10427.1	F	GO:0050660	2ydrol adenine dinucleotide binding
20101	hypothetical protein FG10427.1	F	GO:0008762	UDP-N-acetylmuramate dehydrogenase activity
20101	hypothetical protein FG10427.1	P	GO:0055114	oxidation-reduction process
20105	arylsulfatase precursor	F	GO:0016787	hydrolase activity
20105	arylsulfatase precursor	P	GO:0008152	metabolic process
20107	hypothetical protein FOTG_17375	F	GO:0003824	catalytic activity

20107	hypothetical protein FOTG_17375	P	GO:0009116	nucleoside metabolic process
20116	glucose transporter rco-3	C	GO:0016021	integral component of membrane
20116	glucose transporter rco-3	F	GO:0022891	substrate-specific transmembrane transporter activity
20116	glucose transporter rco-3	P	GO:0055085	transmembrane transport
20116	glucose transporter rco-3	P	GO:0008643	carbohydrate transport
20126	hypothetical protein FG05_30313	C	GO:0005956	protein kinase CK2 complex
20126	hypothetical protein FG05_30313	P	GO:0045859	regulation of protein kinase activity
20126	hypothetical protein FG05_30313	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20126	hypothetical protein FG05_30313	F	GO:0008270	zinc ion binding
20126	hypothetical protein FG05_30313	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20126	hypothetical protein FG05_30313	F	GO:0003677	DNA binding
20126	hypothetical protein FG05_30313	C	GO:0005634	nucleus
20126	hypothetical protein FG05_30313	F	GO:0019887	protein kinase regulator activity
20151	hypothetical protein FG08862.1	F	GO:0008270	zinc ion binding
20151	hypothetical protein FG08862.1	F	GO:0003676	nucleic acid binding
20168	hypothetical protein FG05_03154	P	GO:0055114	oxidation-reduction process
20168	hypothetical protein FG05_03154	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20168	hypothetical protein FG05_03154	F	GO:0008270	zinc ion binding
20168	hypothetical protein FG05_03154	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20168	hypothetical protein FG05_03154	F	GO:0016491	oxidoreductase activity
20168	hypothetical protein FG05_03154	C	GO:0005634	nucleus
20171	aldolase	P	GO:0006725	cellular aromatic compound metabolic process
20171	aldolase	F	GO:0016830	carbon-carbon lyase activity
20172	cytochrome p450 oxidoreductase	F	GO:0016708	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen,

				NAD(P)H as one donor, and incorporation of two atoms of oxygen into one donor
20172	cytochrome p450 oxidoreductase	P	GO:0019439	aromatic compound catabolic process
20172	cytochrome p450 oxidoreductase	F	GO:0004497	monooxygenase activity
20172	cytochrome p450 oxidoreductase	F	GO:0005506	iron ion binding
20172	cytochrome p450 oxidoreductase	F	GO:0051537	2 iron, 2 sulfur cluster binding
20172	cytochrome p450 oxidoreductase	P	GO:0055114	oxidation-reduction process
20176	pathogenicity protein	F	GO:0005215	transporter activity
20182	hypothetical protein FG05_35086	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20182	hypothetical protein FG05_35086	F	GO:0008270	zinc ion binding
20182	hypothetical protein FG05_35086	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20182	hypothetical protein FG05_35086	F	GO:0003677	DNA binding
20182	hypothetical protein FG05_35086	C	GO:0005634	nucleus
20189	homocitrate synthase	F	GO:0004410	homocitrate synthase activity
20189	homocitrate synthase	C	GO:0016021	integral component of membrane
20189	homocitrate synthase	C	GO:0005634	nucleus
20189	homocitrate synthase	C	GO:0005739	mitochondrion
20189	homocitrate synthase	P	GO:0019878	lysine biosynthetic process via aminoadipic acid
20236	cytochrome p450	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20236	cytochrome p450	F	GO:0008270	zinc ion binding
20236	cytochrome p450	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20236	cytochrome p450	F	GO:0003677	DNA binding
20236	cytochrome p450	C	GO:0005634	nucleus
20239	rta1 domain-containing protein	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20239	rta1 domain-containing protein	F	GO:0008270	zinc ion binding
20239	rta1 domain-containing protein	F	GO:0004553	hydrolase activity, 4hydrolysing O-glycosyl compounds

20239	rta1 domain-containing protein	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20239	rta1 domain-containing protein	F	GO:0016758	transferase activity, transferring hexosyl groups
20239	rta1 domain-containing protein	C	GO:0005634	nucleus
20239	rta1 domain-containing protein	P	GO:0030259	lipid glycosylation
20242	hypothetical protein FG05_04947	F	GO:0046872	metal ion binding
20242	hypothetical protein FG05_04947	F	GO:0004672	protein kinase activity
20242	hypothetical protein FG05_04947	P	GO:0006468	protein phosphorylation
20242	hypothetical protein FG05_04947	F	GO:0005524	ATP binding
20245	hypothetical protein FG05020.1	F	GO:0008270	zinc ion binding
20247	zinc finger protein ozf-like protein	F	GO:0046872	metal ion binding
20251	---NA---	F	GO:0016491	oxidoreductase activity
20251	---NA---	P	GO:0055114	oxidation-reduction process
20255	rpp14 family	F	GO:0004540	ribonuclease activity
20255	rpp14 family	P	GO:0090501	RNA phosphodiester bond hydrolysis
20255	rpp14 family	P	GO:0008033	tRNA processing
20264	phosphatidylinositol 4-kinase	P	GO:0046854	phosphatidylinositol phosphorylation
20264	phosphatidylinositol 4-kinase	F	GO:0004430	1-phosphatidylinositol 4-kinase activity
20264	phosphatidylinositol 4-kinase	P	GO:0006661	phosphatidylinositol biosynthetic process
20264	phosphatidylinositol 4-kinase	P	GO:0048015	phosphatidylinositol-mediated signaling
20264	phosphatidylinositol 4-kinase	F	GO:0005488	binding
20283	hypothetical protein FPSE_02973	P	GO:0009306	protein secretion
20289	fungal transcriptional regulatory protein	F	GO:0000981	sequence-specific DNA binding RNA polymerase II transcription factor activity
20289	fungal transcriptional regulatory protein	P	GO:0006357	regulation of transcription from RNA polymerase II promoter
20289	fungal transcriptional regulatory protein	C	GO:0005634	nucleus
20289	fungal transcriptional regulatory protein	F	GO:0008270	zinc ion binding

20292	hypothetical protein FG06162.1	F	GO:0004190	aspartic-type endopeptidase activity
20292	hypothetical protein FG06162.1	P	GO:0006508	proteolysis
20294	serine threonine protein kinase	F	GO:0004672	protein kinase activity
20294	serine threonine protein kinase	P	GO:0006468	protein phosphorylation
20294	serine threonine protein kinase	F	GO:0005524	ATP binding
20304	meiosis protein mei2	F	GO:0003676	nucleic acid binding
20304	meiosis protein mei2	F	GO:0000166	nucleotide binding
20323	hypothetical protein FG11465.1	F	GO:0020037	heme binding
20323	hypothetical protein FG11465.1	F	GO:0016705	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen
20323	hypothetical protein FG11465.1	F	GO:0004497	monooxygenase activity
20323	hypothetical protein FG11465.1	F	GO:0005506	iron ion binding
20323	hypothetical protein FG11465.1	P	GO:0055114	oxidation-reduction process
20326	hypothetical protein FG05_35376	F	GO:0043565	sequence-specific DNA binding
20326	hypothetical protein FG05_35376	F	GO:0003700	sequence-specific DNA binding transcription factor activity
20326	hypothetical protein FG05_35376	P	GO:0006355	regulation of transcription, DNA-templated
20327	alpha-n-arabinofuranosidase	F	GO:0046556	alpha-L-arabinofuranosidase activity
20327	alpha-n-arabinofuranosidase	C	GO:0005576	extracellular region
20327	alpha-n-arabinofuranosidase	P	GO:0046373	L-arabinose metabolic process
20327	alpha-n-arabinofuranosidase	F	GO:0030248	cellulose binding
20330	related to galactoside o-acetyltransferase	F	GO:0016407	acetyltransferase activity
20330	related to galactoside o-acetyltransferase	P	GO:0008152	metabolic process
20373	alkylated dna repair protein alkb like protein 8	P	GO:0055114	oxidation-reduction process
20373	alkylated dna repair protein alkb like protein 8	F	GO:0016706	oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen, 2-oxoglutarate as one donor, and

				incorporation of one atom each of oxygen into both donors
20377	cutinase	C	GO:0005576	extracellular region
20377	cutinase	P	GO:0008152	metabolic process
20377	cutinase	F	GO:0050525	cutinase activity
20397	hypothetical protein FPSE_09629	P	GO:0032259	methylation
20397	hypothetical protein FPSE_09629	F	GO:0008168	methyltransferase activity
20406	serine threonine protein kinase	P	GO:0006468	protein phosphorylation
20406	serine threonine protein kinase	F	GO:0005524	ATP binding
20406	serine threonine protein kinase	P	GO:0046677	response to antibiotic
20406	serine threonine protein kinase	F	GO:0004674	protein serine/threonine kinase activity
20409	gag protein	F	GO:0003676	nucleic acid binding
20409	gag protein	F	GO:0008270	zinc ion binding
20411	histone h4	C	GO:0000786	nucleosome
20411	histone h4	P	GO:0043581	mycelium development
20411	histone h4	F	GO:0003677	DNA binding
20411	histone h4	P	GO:0006334	nucleosome assembly
20411	histone h4	F	GO:0046982	protein heterodimerization activity
20411	histone h4	P	GO:0006352	DNA-templated transcription, initiation
20411	histone h4	C	GO:0005634	nucleus

Additional file 6. Description of secretome methodology.

RRes secretome prediction found signal domain (RRes:2189), removal of sequences with transmembrane domains (RRes:1776) using Interproscan, exclusion of proteins smaller than 20, and the use of ProtComp (Version 9.0) LocDB and PotLocDB to exclude GPI anchored membrane proteins and other non-extracellular loci proteins (RRes:1602), followed by WoLFPSort [28] (extracellular score > 17, WoLFPSort) (RRes:705)) to identify final destination and big-PI [29] to further remove GPI-anchored proteins (RRes:616).

For the 412 new genes the use of blast NCBI 04/09/14 and Interproscan in blast2go revealed, after the blast2go filtering algorithm, 147 of the new gene protein sequences had blast hits using an e-value of 0.001 before blast2go filtering, and 115 had Interproscan functional annotation. After merging of GO terms from Interproscan, 99 Level 2 biological process and 64 molecular function GO terms were identified, see Table-4. Of the 412, 44 of the new genes have signal domains but 6 have transmembrane of those 38, 12 passed the extracellular filtering using ProtComp (Version 9.0) and WoLFPSort (extracellular score > 17) could be new secretome proteins. The top-hit species blast results are shown in supplementary Table-12 which shows that many of the new genes have similar gene models reported in other genome assemblies/annotations of *F. graminearum* or in lesser cases other fungi.

Complete list of FGRRES_gene ID's of predicted secretome and those uniquely predicted vs the *Brown et al* secretome set. *(MIPS V30 ID) gene identified in *Brown et al* secretome which used MIPS V30 gene models, **(RRes ID) gene model from MIPS version V32. 144 of the RRes predictions were confirmed using proteomics from *Yang et al* and *Paper et al*. [£]30 of the 70 proteins identified in the *Yang et al* study from proteins identified in *F. graminearum* culture supernatants after growth in wheat or barley flour-containing medium were confirmed in the in silico predictions. [§]137 of the 286

from the extracellular proteins in vitro and in planta from the pathogenic fungus *F. graminearum* in the *Paper et al* study were confirmed in the in silica predictions.

RRes (all) (n=616)	Identified in both studies (n= 515)	RRes (unique) (n=101)	Brown <i>et al.</i> , (2012) (unique) (n=59)
00002	00002	00072	00129
00006	00006	00571	00742
00009	00009	01351	00806
00023	00023	01711	01437
00028 [§]	00028	01846	01588
00029	00029	01993	02348
00060 [§]	00060	02190	02685
00061	00061	02204	02976
00062	00062	02293	03406
00072	00100	02328	03495
00100 [§]	00111	02351	03574
00111	00112	02385	03816
00112	00114	02408	03827
00114 [§]	00131	02721	03842
00131	00184	03114	03975
00184 [§]	00230	03157	04372
00230	00260	03662	04397
00260	00294	03713	04521
00294	00411	04345	04597
00411	00569	04758	04699
00569	00642	04781	04824
00571 [§]	00783	05345	04900
00642	00847	05628	05757
00783	00987	05835	06479
00847	00989	06549	06572
00987	01239	06888	06605
00989 [§]	01240	07010	06616
01239	01283	07772	06712
01240	01570	08003	06741(16708)**
01283	01595	08493	07556(16845)**
01351	01598	09522	07678
01570	01603	10604	07871
01595 [§]	01607	11077	07944
01598	01621	11231	08136
01603	01636	11333	08415
01607 [§]	01660	11350	08454(17031)**

01621 [§]	01674	11361	08549
01636	01685	12486	08583(17009)**
01660	01688	12652	09291
01674	01728	13462	09307
01685	01748	15885	09390
01688	01754	15937	09742
01711	01763	15958	10525
01728	01776	16044	10595
01748 [§]	01778	16066	11164
01754	01803	16175	11236
01763	01815	16196	11472
01776	01818	16234	11517(16047)**
01778	01829	16333	11645
01803	01831	16334	11745(15729)**
01815	01872	16408	12337
01829 [§]	01982	16459	12414(16243)**
01831	01988	16547	12439
01846	02015	16584	13583(17212)**
01872	02059	16623	13883
01982	02154	16668	13924(17524)**
01988 [§]	02202	16697	13992(17575)**
01993 [§]	02228	16759	15260(16423)**
02015	02255	16883	15448
02059 [§]	02263	16898	
02190	02269	17022	
02202	02309	17115	
02204 [§]	02314	17419	
02228	02330	17425	
02255 [§]	02332	17434	
02263	02339	17455	
02269	02354	17468	
02293	02360	17469	
02309 [§]	02378	17494	
02314	02422	17545	
02328	02448	17566	
02330	02651	20027	
02332	02658	20104	
02339 [§]	02674	20176	
02351	02686	20208	
02354 [§]	02841	20240	
02360 [§]	02888	20306	
02378	02890	20325	

02385	02893	20327	
02408	02899	20377	
02422	02917	20390	
02448	02918	20400	
02651	02933	20407	
02658 ^{ES}	02962	02909_M	
02674	02977	04165_M	
02686	02987	04374_M	
02721 ^S	02999	04756_M	
02841	03002	07126_M	
02888	03003	07686_M	
02890	03034	08074_M	
02893	03035	09143_M	
02899	03049	10563_M	
02917	03050	10607_M	
02918 ^S	03052	10776_M	
02933	03054	11513_M	
02962	03096	12405_M	
02987	03109	13975_M	
02999	03121	13987_M	
03002 ^S	03122	15136_M	
03003 ^{ES}	03123	15714_M	
03035 ^{ES}	03130	17388_M	
03049 ^S	03131		
03050	03143		
03052 ^S	03156		
03054 ^S	03194		
03096	03209		
03109	03212		
03114 ^S	03217		
03121	03243		
03122	03274		
03123 ^S	03275		
03130	03304		
03131 ^S	03307		
03143 ^S	03315		
03156	03331		
03157	03334		
03194 ^S	03343		
03209	03348		
03212 ^S	03365		
03217	03371		

03243 [§]	03379		
03274	03384		
03275	03387		
03304	03394		
03307	03402		
03315 [§]	03432		
03334 [§]	03436		
03343 [§]	03445		
03348	03457		
03365 ^{ES}	03463		
03371	03467		
03379 [§]	03483		
03384	03507		
03394	03521		
03402	03526		
03432	03529		
03436	03531		
03445	03532		
03457	03544		
03463	03569		
03467 [§]	03570		
03483 [§]	03583		
03507	03584		
03521	03585		
03526	03598		
03529	03599		
03531	03600		
03532	03601		
03569	03609		
03583	03612		
03584 [§]	03614		
03585	03616		
03598 [§]	03624		
03599	03628		
03600 [§]	03629		
03601	03632		
03609 [§]	03640		
03612	03674		
03614	03687		
03616	03695		
03624 [§]	03708		
03628 [§]	03724		

03629	03748		
03632 ^{\$}	03769		
03640	03784		
03662 ^{\$}	03795		
03695 ^{\$}	03813		
03708	03820		
03713	03846		
03724	03858		
03748	03865		
03769	03867		
03784	03883		
03795 ^{\$}	03893		
03813 ^{\$}	03894		
03820	03896		
03846	03901		
03858	03904		
03865 ^{\$}	03905		
03883 [£]	03908		
03893	03911		
03896	03916		
03901	03922		
03904	03944		
03905 ^{\$}	03954		
03908 ^{\$}	03958		
03911	03960		
03916	03969		
03922	03972		
03944	03986		
03954 ^{\$}	04029		
03958	04060		
03960	04074		
03969	04146		
03972	04239		
03986 ^{\$}	04380		
04029	04429		
04060	04434		
04074 ^{£\$}	04439		
04146	04471		
04345	04503		
04380	04504		
04429	04527		
04434	04546		

04439	04563		
04471	04583		
04503	04620		
04504	04624		
04527 ^{£\$}	04646		
04546 ^{\$}	04656		
04563	04678		
04583 ^{\$}	04681		
04620	04685		
04624	04689		
04656	04704		
04678	04732		
04685	04735		
04689	04739		
04704	04740		
04732 [£]	04743		
04735 ^{\$}	04744		
04739	04745		
04740	04768		
04743	04817		
04744	04818		
04745	04841		
04758 ^{£\$}	04848		
04768 ^{\$}	04858		
04781	04864		
04817	04895		
04818	04953		
04841	04980		
04848 ^{\$}	05046		
04858	05052		
04864 ^{\$}	05163		
04895	05245		
04953	05341		
04980	05609		
05046	05719		
05052	05763		
05163	05797		
05245	05841		
05341	05847		
05345	05851		
05609	05906		
05628	05933		

05719	05947		
05763	05970		
05797 ^{£\$}	05983		
05835	06017		
05841	06087		
05847	06110		
05851 ^{\$}	06117		
05906	06278		
05933	06437		
05947	06443		
05970	06451		
05983	06452		
06017 ^{\$}	06463		
06087	06465		
06110	06466		
06117	06469		
06278 [£]	06497		
06443	06506		
06451	06610		
06452 ^{\$}	06612		
06463 ^{\$}	06733		
06465 ^{\$}	06775		
06466 ^{£\$}	06895		
06469	06993		
06497	07207		
06506	07221		
06549 ^{£\$}	07238		
06610	07435		
06612	07551		
06733	07569		
06775	07608		
06888	07625		
06895	07629		
06993 ^{£\$}	07639		
07010	07661		
07207	07684		
07221	07695		
07238	07698		
07435	07699		
07551 ^{\$}	07721		
07569 ^{\$}	07728		
07608 ^{\$}	07755		

07625	07775		
07629	07794		
07639 ^{\$}	07797		
07661	07829		
07695 ^{\$}	07838		
07698	07899		
07699	07921		
07721	07988		
07728	07993		
07755	07996		
07772 ^{\$}	08002		
07775 ^{£\$}	08011		
07794	08021		
07797	08026		
07829	08041		
07838 ^{£\$}	08048		
07899	08085		
07921 ^{\$}	08090		
07988 ^{\$}	08094		
07993	08115		
08002	08116		
08003	08122		
08011 ^{\$}	08150		
08026	08175		
08048 ^{\$}	08180		
08085	08196		
08094	08210		
08115	08238		
08116	08265		
08122 ^{\$}	08344		
08150	08387		
08175	08464		
08180	08824		
08196	08825		
08210	08907		
08238 ^{\$}	08958		
08265 ^{£\$}	08978		
08344	08987		
08387	09032		
08464	09066		
08493	09071		
08824	09085		

08825 [§]	09093		
08907	09099		
08958	09109		
08978	09118		
08987	09127		
09066	09132		
09071	09134		
09093	09142		
09099	09181		
09109	09345		
09118	09353		
09127	09358		
09132	09382		
09134	09445		
09142	09475		
09181	09586		
09345	09646		
09353	09821		
09358	10206		
09382	10212		
09445	10316		
09475 [§]	10357		
09522 [§]	10435		
09586	10495		
09646	10500		
09821	10543		
10206	10551		
10212 [§]	10554		
10316 [§]	10561		
10357	10587		
10435	10592		
10495	10594		
10500	10609		
10543 [§]	10611		
10551	10670		
10554	10675		
10561	10676		
10587	10677		
10592	10712		
10604	10713		
10609 [§]	10729		
10611 [§]	10972		

10670 ^{\$}	10982		
10675 ^{\$}	10986		
10676	10998		
10677 ^{\$}	10999		
10712 ^{\$}	11006		
10713	11008		
10729	11011		
10972	11032		
10982 ^{\$}	11033		
10986	11036		
10998	11037		
10999 ^{\$}	11046		
11006	11047		
11008 [£]	11048		
11011 ^{\$}	11052		
11032	11066		
11033 ^{£\$}	11094		
11036 ^{\$}	11095		
11037 ^{£\$}	11097		
11046	11101		
11047	11106		
11048 ^{\$}	11112		
11052	11136		
11066	11143		
11077	11156		
11094	11163		
11095	11166		
11097 ^{£\$}	11169		
11106	11184		
11112 ^{\$}	11190		
11136	11204		
11143	11205		
11156	11206		
11163 ^{\$}	11208		
11166	11225		
11169 ^{£\$}	11227		
11184 ^{£\$}	11228		
11190 ^{\$}	11229		
11204 [£]	11232		
11205 ^{\$}	11238		
11206	11249		
11208 ^{\$}	11276		

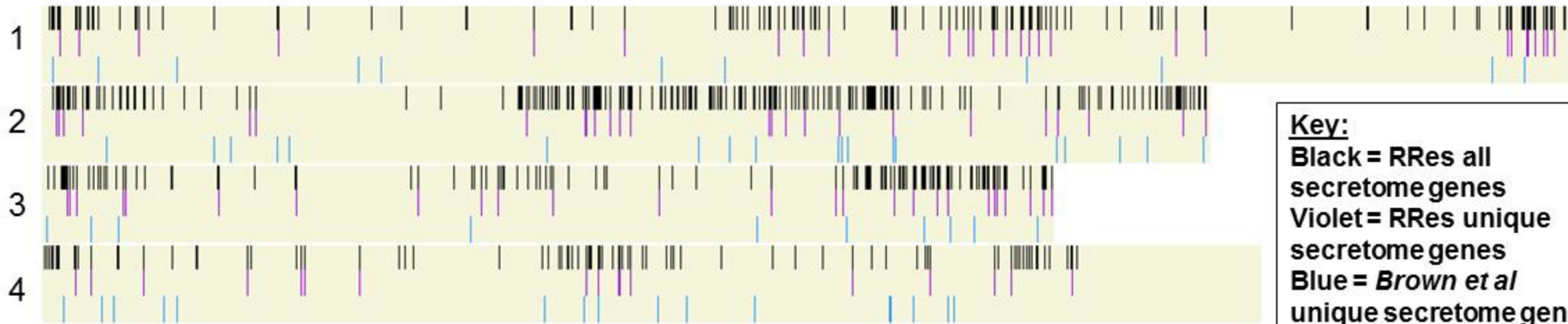
11225	11280		
11227	11296		
11228 ^{\$}	11304		
11229	11315		
11231	11318		
11232	11326		
11238	11348		
11249 ^{\$}	11360		
11276	11366		
11280 ^{£\$}	11379		
11296	11386		
11304 ^{\$}	11428		
11315	11487		
11318	11488		
11326 ^{\$}	11493		
11348 ^{\$}	11496		
11350	11516		
11360	11528		
11361	11548		
11366 ^{\$}	11559		
11379	11602		
11386 ^{£\$}	11647		
11428 ^{\$}	11648		
11487 ^{\$}	11675		
11488	11715		
11493	12067		
11496	12081		
11516	12119		
11528	12142		
11548 ^{\$}	12206		
11559	12210		
11602	12214		
11647	12251		
11675	12300		
11715	12369		
12067	12390		
12081	12434		
12119	12504		
12142	12514		
12206	12548		
12210	12551		
12214	12554		

12251	12591		
12300	12622		
12369 [£]	12644		
12390	12835		
12434	12918		
12486	13189		
12504	13245		
12514	13412		
12548	13436		
12551	13443		
12554	13450		
12591	13505		
12622	13515		
12644	13628		
12652	13692		
12835	13876		
12918	13958		
13189	14020		
13412	15003		
13443	15048		
13450	15123		
13462	15142		
13505	15251		
13515	15437		
13628	15469		
13692	15661		
13876	15679(00031)*		
13958	16013(02551)*		
15003	16371(02898)*		
15048	16353(03012)*		
15123	16328(03095)*		
15142	16324(03129)*		
15251	16542(12807)*		
15437	17151(09046)*		
15469	17535(11490)*		
15661	17084(08213)*		
15679	17063(08289)*		
15885	17119(08007)*		
15937	16930(07934)*		
15958	16877(07691)*		
15982	16148(04097)*		
16013	16104(04501)*		
16044	16074(04684)*		

16066	16635(06332)*		
16074	16657(06438)*		
16104	17402(10615)*		
16148	17405(10622)*		
16163	17424(10656)*		
16175	17462(11170)*		
16196	15982(12127)*		
16234	16163(12326)*		
16324	17459(13840)*		
16328	12545_16364_M(12544)*		
16333	16312(12492)*		
16334	16111(12256)*		
16353	16018(12160)*		
16371	15917(12047)*		
16408	16051(11532)*		
16459	16431(04849)*		
16542			
16547			
16584			
16623			
16635			
16657			
16668			
16697			
16759			
16877			
16883			
16898			
16930			
17022			
17063			
17084			
17115			
17119			
17151			
17402			
17405			
17419			
17424			
17425			
17434			
17455			
17459			
17462			

17468			
17469			
17494			
17535			
17545			
17566			
20027			
20104			
20176			
20208			
20240			
20306			
20325			
20327			
20377			
20390			
20400			
20407			
01818_M ^{£\$}			
02154_M			
02909_M			
02977_M ^{\$}			
03034_M			
03331_M			
03387_M			
03544_M			
03570_M			
03674_M ^{\$}			
03687_M ^{£\$}			
03867_M [£]			
03894_M			
04165_M			
04239_M ^{\$}			
04374_M			
04646_M			
04681_M			
04756_M			
06437_M			
07126_M			
07684_M			
07686_M			
07996_M			
08021_M			
08041_M			

08074_M			
08090_M			
09032_M			
09085_M ^s			
09143_M			
10563_M			
10594_M			
10607_M			
10776_M			
11101_M			
11333_M			
11513_M			
11648_M			
12405_M			
12545_16364_M			
13245_M			
13435_6_M			
13975_M			
13987_M			
14020_M			
15136_M			
15714_M			
15917_M			
16018_M			
16051_M			
16111_M			
16312_M			
16431_M			
17388_M			



Additional file 8. Gene ID's with GO term 0001071 which are identified as having nucleic acid binding transcription factor activity from either altered gene predictions or genes that were not found in the MIPS v3.2 annotation. *Gene ID to be changed to a modified gene annotation in RRes v4.1.

Modified RRes	New RRes	New RRes blast hits
02787_M	20236 (*12631_M)	cytochrome p450
03862_M	20126	hypothetical protein FG05_30313
03873_M	20239	rta1 domain-containing protein
04170_M	20168	hypothetical protein FG05_03154
04203_M	20326	hypothetical protein FG05_35376
07575_M	20048	hypothetical protein FG05_30608
09075_M	20289	fungal transcriptional regulatory protein
09465_M	20182	hypothetical protein FG05_35086
10014_M		
10684_M		
10891_M		
10977_M		
10980_M		
11462_M		
11524_M		
12018_M		
14003_M		
15730_M		
16506_M		
16822_M		
17196_M		
17345_M		
17658_M		
12427_B_M		

Additional file 9. Distal centromere and telomere gene ID's and annotations.

Minimum	MIPS Annotation	RRes Annotation
Centromere 1-5'		
5S_rRNA		
16043	hypothetical protein	
02778	probable histone H3 methyltransferase DIM-5	
02779	conserved hypothetical protein	
02780	probable coatamer complex beta chain	
11636	probable mago nashi protein	
11637	probable VID24 - required for vacuolar import and degradation of Fbp1p	
11632	conserved hypothetical protein	
11633	conserved hypothetical protein	
11634	related to ankyrins	
11594	related to tRNA ligase	
11595	conserved hypothetical protein	
11596	related to ribosomal protein YmL11 precursor, mitochondrial	
11597	probable 26S proteasome regulatory particle chain RPT5	
11602	probable brefeldin A resistance protein	
11603	related to KM-PA-2 protein	
16044	related to cholinesterase precursor	
centromere 1-3'		
13662	probable RPO31 - DNA-directed RNA polymerase III, 160 KD subunit	
13663	conserved hypothetical protein	
09861	conserved hypothetical protein	
09862	probable fimbrin	
09863	related to acyl-CoA cholesterol acyltransferase	
09864	related to mRNA splicing factor	
09865_M	NA	related to mitochondrial carrier protein
09866	probable ribosomal protein L18, cytosolic	
17283	related to MRM2 - Mitochondrial rRNA methyltransferase	
09868	conserved hypothetical protein	
09869	related to OST3 - oligosaccharyltransferase gamma subunit	
09870	probable CPC2 protein	
09871	conserved hypothetical protein	
09872	conserved hypothetical protein	
09873	probable glycine hydroxymethyltransferase	
17593	probable translation activator GCN1	

centromere 2 - 5'		
08944	probable glucosamine-phosphate N-acetyltransferase	
08945	conserved hypothetical protein	
08946	probable alpha-N-arabinofuranosidase / alpha-L-arabinofuranosidase	
08947	conserved hypothetical protein	
16939	related to response regulator Mcs4	
08949	conserved hypothetical protein	
08950	conserved hypothetical protein	
20155	not found	
16938_M	NA	hypothetical protein
16937	hypothetical protein	
11609_M	NA	conserved hypothetical protein
20156	not found	
11610_M	NA	conserved hypothetical protein
11590	probable TAD2 - tRNA-specific adenosine deaminase 2	
16936	hypothetical protein	
00592_M	NA	conserved hypothetical protein
13260	hypothetical protein	
centromere 2 - 3'		
16403	probable sterol glucosyltransferase	
16402	hypothetical protein	
12579	hypothetical protein	
11617	#N/A	
11618	conserved hypothetical protein	
11619	probable NADH-ubiquinone oxidoreductase 21 kDa subunit	
11622	conserved hypothetical protein	
11623	conserved hypothetical protein	
12578_M	NA	probable class E vacuolar-protein sorting and endocytosis factor
02781	probable SKI2 - antiviral protein and putative helicase	
02782	probable pyruvate dehydrogenase (lipoamide) beta chain precursor (PDB1)	
02783	probable DELTA(24)-STEROL C-METHYLTRANSFERASE (ERG6)	
012577_M	NA	related to cell cycle control protein cwf5
02784_M	NA	conserved hypothetical protein
02785	probable zuotin	
16401	probable ATP-binding cassette multidrug transport protein ATRC	
centromere		

e 3 - 5'		
16651_M	related to CECR1 protein	
06423	conserved hypothetical protein	
06424	related to SRP40 - suppressor of mutant AC40 of RNA polymerase I and III	
06425	conserved hypothetical protein	
12916_M	NA	related to probable gtp-binding protein engb
16652	hypothetical protein	
16653	related to ubiquitin-protein ligase	
06427	related to YBR267w	
06428	conserved hypothetical protein	
centromere 3 - 3'		
13936	hypothetical protein	
11598	related to L-fucose permease	
17530	related to neurofibromin	
10716	related to dna exoribonuclease Dhp1p	
10717	related to transport protein USO1	
10718	related to protein arginine N-methyltransferase 3	
17529	related to permeases - unknown function	
13934	related to RAV1 - similarity to Drosophila DmX gene	
17739	conserved hypothetical protein	
10721	conserved hypothetical protein	
10722	probable isoleucine--tRNA ligase	
10723	probable hymA gene	
10724	related to SWI/SNF complex 60 kDa subunit	
centromere 4- 5'		
07954	related to mouse proteinase activated receptor 2	
07955	related to GPI-anchor transamidase GAA1	
07956	related to 26S proteasome regulatory particle chain RPN9	
13257	probable NADH2 dehydrogenase (ubiquinone) 40K chain	
13258	probable NADH2 dehydrogenase (ubiquinone) 40K chain	
07958	related to ubiquitin-specific protease 22	
07959	conserved hypothetical protein	
07960	related to YTP1	
07961	conserved hypothetical protein	
07962	probable Rad54 homolog MUS-25	
16933	hypothetical protein	
07963	related to ornithine transporter 2	
07964	related to member of RSC complex	

16934	hypothetical protein	
07966	probable transmembrane protein UsgS	
07967	conserved hypothetical protein	
16935	hypothetical protein	
13259	hypothetical protein	
07969	conserved hypothetical protein	
07970_M	NA	NONE
centromere 4- 3'		
11628	conserved hypothetical protein	
11627	probable HTA2 - histone H2A.2	
11626	probable HTB1 - histone H2B	
13661	conserved hypothetical protein	
17282_M	NA	related to ATG4 - essential for autophagocytosis
09857	related to dnaJ-like proteins	
09856	conserved hypothetical protein	
09855	related to O-succinylhomoserine (thiol)-lyase	
17281	hypothetical protein	
09853	conserved hypothetical protein	
09852	conserved hypothetical protein	
09851	probable SAP1 - member of the AAA-protein family	
09850	related to COG6 - conserved oligomeric golgi complex	
13660	probable APG9 - integral membrane protein required for Cvt and autophagy transport	
13659	hypothetical protein	

Telomeres -first 2 genes			
	ID		
chr 1 - 5'	16072_M	hypothetical protein	homology in chr2 ?
	11606	related to hexose transporter protein	
chr 1 - 3'	17444	related to flavin-containing monooxygenase	
	13822_M	related to flavin-containing monooxygenase	
chr 2 - 5'	14031_M	conserved hypothetical protein	no homology in other chr
	15663	hypothetical protein	

chr 2 - 3'	11586	conserved hypothetical protein	
	11587_M	probable sterol glucosyltransferase	no homology in other chr
chr 3 - 5'	20223	NA	
	12581_2_M	hypothetical protein	NA but MIPS given;
chr 3 - 3'			
	15657_M	hypothetical protein	
	20327	NA	
	20328	NA	Has homology to chromosome 2 -5' end !
chr 4 - 5'	14027_16070_11614_M	Ankyrin repeat domain protein; amily: Ank_2 (PF12796)	
	11613_M	conserved hypothetical protein	
	16071	hypothetical protein	
chr 4 - 3'	tRNA		
(def ense resp onse relat ed)	08956_M	NB-ARC doamin signalling motif; related to kinesin light chain	
	13478	conserved hypothetical protein	
	08955_M	related to WD40-repeat protein (notchless protein)	
	8954	conserved hypothetical protein	
	13477	conserved hypothetical protein	
	13476	conserved hypothetical protein	
	17133	related to RSA4 - WD-repeat protein required for maturation and efficient intra-nuclear transport or pre-60S ribosomal subunits	NA but MIPS given
(def ense resp onse relat ed)	17709_M	NB-ARC domain signalling motif; related to calcium-independent phospholipase A2	
	20409	NA	
	20412	NA	
	rRNA cluster	28S(?)rRNA	

neocentromere			
	ID		
chr 3 - 5'			
	17445	Aconitase/3-isopropylmalate dehydratase	iron
	13824	P-loop containing nucleoside triphosphate hydrolase, NACHT	
	17446	Multi-copper oxidase	copper
	11236	phosphoesterase	
	17447	NACHT domain	
	11234_M	TRANSCRIPTION TERMINATION FACTOR 2	
	11232	Amine oxidase	
	11231	MOLYBDOPTERIN COFACTOR SULFURASE	
chr 3 - 3'			
	11620	ARRESTIN-RELATED TRAFFICKING ADAPTER 3-RELATED	
	17589	Marvel domain, MAL and related proteins for vesicle trafficking and membrane link	
	17597_112 39_M	Potassium channel	
	11240	ATP-BINDING CASSETTE TRANSPORTER	
	11241	EXOPOLYPHOSPHATASE	
	17587	Aerobactin siderophore biosynthesis	iron

Additional file 10. The alignment positions of the AT rich regions from the *F. verticillioides* chromosomes to the sequences immediately flanking the centromeres in RRes v4.0 *F. graminearum*.

FG Chromosome centromere	Amino centromere flank			Carboxyl centromere flank		
	Chromosome	Position (Mbp)	Orientation (R/F)*	Chromosome	Position (Mbp)	Orientation (R/F)*
1	5	2.65	F	5	0.00	F
2	6	3.90	F	N/A	N/A	N/A
3	2	2.06	F	DS486012 (supercontig)	0.00	R
4	4	3.30	F	4	2.30	R
4 (neocentromere)	N/A	N/A	N/A	N/A	N/A	N/A

*R/F = reverse complement or forward orientation of reference sequence

**N/A = no homology

Additional file 11. RRes v4.0 gene models that have changed since MIPS v3.2 that are present in PHI-base. ^A MIPS gene ID version 30 unless ^B MIPS gene call version 3.2; ^C gene homolog name provided in brackets.

PHI-base ID	MIPS gene ID ^A	Mutant phenotype in planta	Gene name or homolog ^C
PHI:1182	FGSG_13318	Reduced virulence	Sc Mec1
PHI:1206	FGSG_04053	Reduced virulence	Sp Prp4
PHI:1251	FGSG_01506 ^B	Unaffected pathogenicity	
PHI:1270	FGSG_03146 ^B	Unaffected pathogenicity	
PHI:1276	FGSG_09150 ^B	Unaffected pathogenicity	
PHI:1278	FGSG_11614 ^B	Unaffected pathogenicity	
PHI:1281	FGSG_02153 ^B	Unaffected pathogenicity	
PHI:1286	FGSG_07812 ^B	Unaffected pathogenicity	
PHI:1338	FGSG_13313	Unaffected pathogenicity	GzbZIP020
PHI:1339	FGSG_13759	Unaffected pathogenicity	GzbZIP021
PHI:1359	FGSG_02788	Unaffected pathogenicity	GzC2H019
PHI:1377	FGSG_06168	Unaffected pathogenicity	GzC2H040
PHI:1385	FGSG_07075	Unaffected pathogenicity	GzC2H048
PHI:1386	FGSG_07187	Unaffected pathogenicity	GzC2H049
PHI:1392	FGSG_07751	Unaffected pathogenicity	GzC2H055
PHI:1397	FGSG_07952	Unaffected pathogenicity	GzC2H060
PHI:1400	FGSG_08246	Unaffected pathogenicity	GzC2H063
PHI:1431	FGSG_12970	Unaffected pathogenicity	Pac1
PHI:1432	FGSG_12973	Unaffected pathogenicity	GzC2H100
PHI:1435	FGSG_13492	Unaffected pathogenicity	GzC2H104
PHI:1437	FGSG_13896	Unaffected pathogenicity	GzC2H106
PHI:1466	FGSG_07087	Unaffected pathogenicity	GzCCAAT006
PHI:1483	FGSG_04557	Unaffected pathogenicity	GzHMG015
PHI:1486	FGSG_05604	Unaffected pathogenicity	GzHMG018
PHI:1498	FGSG_10762	Lethal	GzHMG030

PHI:1502	FGSG_12323	Unaffected pathogenicity	GzHMG034
PHI:1527	FGSG_13431	Lethal	GzHOMEL040
PHI:1627	FGSG_08551	Unaffected pathogenicity	GzTF2S002
PHI:1645	FGSG_07433	Unaffected pathogenicity	GzWing017
PHI:1660	FGSG_11696	Unaffected pathogenicity	GzCCCH004
PHI:1669	FGSG_10277	Unaffected pathogenicity	GzCCHC009
PHI:1675	FGSG_06542	Unaffected pathogenicity	GzDHHC003
PHI:1684	FGSG_04258	Unaffected pathogenicity	GzRad002
PHI:1695	FGSG_10977	Unaffected pathogenicity	GzZC010
PHI:1696	FGSG_11051	Unaffected pathogenicity	GzZC011
PHI:1697	FGSG_11462	Unaffected pathogenicity	GzZC012
PHI:1698	FGSG_10891	Unaffected pathogenicity	GzZC013
PHI:1701	FGSG_04496	Unaffected pathogenicity	GzZC016
PHI:1702	FGSG_07575	Unaffected pathogenicity	GzZC017
PHI:1707	FGSG_04170	Unaffected pathogenicity	GzZC022
PHI:1713	FGSG_02825	Unaffected pathogenicity	GzZC028
PHI:1714	FGSG_06448	Lethal	GzZC029
PHI:1715	FGSG_06380	Unaffected pathogenicity	GzZC030
PHI:1718	FGSG_13652	Unaffected pathogenicity	GzZC033
PHI:1721	FGSG_13344	Unaffected pathogenicity	GzZC036
PHI:1748	FGSG_12528	Unaffected pathogenicity	GzZC063
PHI:1751	FGSG_09726	Unaffected pathogenicity	GzZC066
PHI:1805	FGSG_08028	Reduced virulence	GzZC120
PHI:1823	FGSG_09075	Unaffected pathogenicity	GzZC138
PHI:1826	FGSG_10980	Unaffected pathogenicity	GzZC141
PHI:1833	FGSG_04203	Unaffected pathogenicity	GzZC148
PHI:1838	FGSG_02787	Unaffected pathogenicity	GzZC153
PHI:1840	FGSG_06460	Unaffected pathogenicity	GzZC155
PHI:1846	FGSG_12729	Unaffected pathogenicity	GzZC161
PHI:1851	FGSG_09064	Lethal	GzZC166
PHI:1863	FGSG_04671	Unaffected pathogenicity	GzZC178

PHI:1864	FGSG_14003	Unaffected pathogenicity	GzZC179
PHI:1865	FGSG_13455	Unaffected pathogenicity	GzZC180
PHI:1886	FGSG_07693	Unaffected pathogenicity	GzZC201
PHI:1894	FGSG_05716	Unaffected pathogenicity	GzZC209
PHI:1905	FGSG_03294	Unaffected pathogenicity	GzZC220
PHI:1908	FGSG_03226	Unaffected pathogenicity	GzZC223
PHI:1911	FGSG_03207	Unaffected pathogenicity	GzZC226
PHI:1949	FGSG_03485	Lethal	GzZC264
PHI:1957	FGSG_04888	Unaffected pathogenicity	GzZC272
PHI:1963	FGSG_06503	Unaffected pathogenicity	GzZC278
PHI:1995	FGSG_00196	Unaffected pathogenicity	GzZC310
PHI:2386	FGSG_12857	Loss of pathogenicity	ACL1
PHI:2395	FGSG_00950	Reduced virulence	GzSYN1

Additional file 12. RRes *F. graminearum* recombinant and non-recombinant region comparison. No observable pattern was identified between the high recombinant and non-highly recombinant regions when chromosomal positions were also taken into account.

Region	Chrom	Position	length	AT (%)	Gene no.	Gene no. per 100 kb	Secretome no.	Modified Gene no.	New gene no.
Recombinant region 1	1 (5' long arm)	32,785 575,729	542,944	53.4	233	42.91	20	26	7
Recombinant region 2	3 (5' long arm)	12,014 401,069	389,055	53.6	169	43.44	18	22	8
Recombinant region 3	1 (central long arm)	5,639,082 6,032,689	393,607	52.4	161	40.90	13	20	2
Recombinant region 4	2 (3' end long arm)	8,387,552 8,806,410	418,858	53.7	171	40.83	18	33	3
Recombinant region 5	4 (central long arm)	4,064,366 4,608,192	543,826	53.2	249	45.79	19	37	8
Non-recombinant region 1	1 (adjacent to 5' end long arm)	866,719 1,471,900	605,181	50.9	221	36.52	2	13	5
Non-recombinant region 2	3 (adjacent to 5' end long arm)	651,956 1,687,347	1,035,391	51.4	419	40.47	7	28	13
Non-recombinant region 3	4 (3' end short arm)	7,219,363 7,842,145	622,782	52.4	263	42.23	18	46	9
Non-recombinant region 4	2 (central short arm)	1,757,141 2,235,287	478,146	51.2	170	35.55	0	29	4

Additional file 13. RRes Repeat masking details and gene ID's identified with transposon domains.

In total using RepeatModeller, 12 repeat motif containing sequences were identified including a long terminal direct repeat (LTR) of the gypsy group, TcMar-Ant1, TcMar-Pogo, TcMar-Marin and TcMar-Fot1 transposon identified as two large repeat sequences, 15 other unknown repeat sequences, one SINE and one rRNA located in the chromosomal 4 repetitive region (previously discussed).

Transposon Gene ID (FGRRES_)	Transposon class
08954	TcMar-Ant1
06490	TcMar-Ant1, TcMar-ISRm1
13478	TcMar-Ant1, LTR-Gypsy
05781	TcMar-Tc1
15950	TcMar-Fot1
17263	TcMar-Fot1
16975	TcMar-Fot1
13431_M	TcMar-Fot1
07243	TcMar-Pogo
05264	TcMar-Pogo
08846	TcMar-Marin
20409	LTR/Gypsy

Additional file 14. Transposon and repeat content of centromeres and subtelomeric AT rich regions.

Chromosome/region	Simple repeat (total length bp) no.	Low complexity repeat (total length bp) no.	Transposon (total length bp) no.	Length of region (bp)	Transposon classes
1 Amino subtelomere	(69) 1	0	(1,725) 1	6,000	MarCry-1_FO
1 Centromere	(854) 17	(586) 9	(31,828) 34	56,581	MarCry-1_FO, LTR-Gypsy
1 Carboxyl subtelomere	(1,744) 38	(630) 8	(141)	15,700	MarCry-1_FO
2 Amino subtelomere	(60) 1	0	(3,738) 2	6,000	MarCry-1_FO
2 Centromere	(1,848) 33	(442) 8	(35,049) 43	65,181	MarCry-1_FO, LTR-Gypsy
2 Carboxyl subtelomere	(123) 2	0	(1,159) 3	3,600	MarCry-1_FO, LTR-Gypsy
3 Amino subtelomere	(384) 7	(42) 1	(746) 2	10,800	MarCry-1_FO, LTR-Gypsy
3 Centromere	(278) 7	0	(35,251) 25	56,387	MarCry-1_FO, LTR-Gypsy
3 Carboxyl subtelomere	(1,003) 18	(250) 5	(6,080) 12	21,300	MarCry-1_FO, LTR-Gypsy
3 neocentromere	(513) 10	(679) 2	(7,901) 10	14,951	MarCry-1_FO, LTR-Gypsy
4 Amino subtelomere	0	0	(372) 1	661	LTR-Gypsy
4 centromere	(746) 16	(228) 5	(33,842) (31)	60,933	MarCry-1_FO, LTR-Gypsy

Additional file 15. Unique ger A table of *Fusarium graminearum* species specific genes identified in th

#chrom	start	end	strand	name
Chromosome_1	59208	59741	-	00001
Chromosome_1	540287	540640	+	00164
Chromosome_1	3058264	3058707	-	00927
Chromosome_1	3757352	3757940	+	01148
Chromosome_1	4305969	4307320	-	01303
Chromosome_1	5406584	5407009	-	01628
Chromosome_1	6864667	6865301	-	02086
Chromosome_1	7284465	7285153	+	02243
Chromosome_1	8097774	8098171	-	02517
Chromosome_1	8168437	8169083	-	02537
Chromosome_2	3670699	3671902	-	02889
Chromosome_2	3849234	3849548	+	02955
Chromosome_2	4481609	4482061	+	03196
Chromosome_2	4748425	4748921	+	03297
Chromosome_2	5409654	5409785	-	03539
Chromosome_2	6074199	6075092	-	03793
Chromosome_2	8155516	8155931	-	04472
Chromosome_2	8216660	8217527	+	04493
Chromosome_2	8221051	8221469	+	04495
Chromosome_2	8524929	8525412	+	04600
Chromosome_2	8563401	8565405	+	04617
Chromosome_3	27693	28493	-	04695
Chromosome_3	236115	237192	+	04766
Chromosome_3	296989	297434	-	04792
Chromosome_3	536163	536622	-	04878
Chromosome_3	655012	655303	-	04924
Chromosome_3	736707	737040	+	04955
Chromosome_3	746751	747806	-	04959
Chromosome_3	755547	756041	-	04962
Chromosome_3	3503222	3503628	+	05832
Chromosome_3	3643120	3643482	-	05881
Chromosome_3	5136706	5137400	-	06360
Chromosome_4	209260	210493	-	06487
Chromosome_4	1018746	1019048	-	06734
Chromosome_4	1175886	1176864	+	06773
Chromosome_4	1373271	1373868	+	06829
Chromosome_4	3672169	3672624	+	07490
Chromosome_4	3994472	3994871	-	07605
Chromosome_2	236943	237593	-	08036
Chromosome_2	549628	550100	+	08152
Chromosome_2	585491	585989	+	08167
Chromosome_2	846930	849260	+	08261
Chromosome_2	1178388	1179015	-	08370
Chromosome_2	2056531	2057079	+	08625
Chromosome_2	2352902	2354194	-	08705
Chromosome_2	2910379	2911797	-	08854
Chromosome_4	7164516	7166736	+	09248

Chromosome_4	7070067	7070681 +	09278
Chromosome_4	6443986	6444757 +	09458
Chromosome_1	9818722	9819434 +	10093
Chromosome_1	10759515	10760090 +	10393
Chromosome_1	10775376	10776112 +	10400
Chromosome_1	10792917	10793466 +	10406
Chromosome_1	11016906	11017947 -	10485
Chromosome_1	11023629	11025052 -	10487
Chromosome_1	11434365	11434841 +	10603
Chromosome_1	11463714	11464539 +	10614
Chromosome_1	11495454	11496363 -	10625
Chromosome_1	11680115	11680805 -	10697
Chromosome_1	11717950	11718956 -	10709
Chromosome_1	11719810	11721464 +	10710
Chromosome_3	6199255	6199770 +	10953
Chromosome_3	7261812	7262497 +	11323
Chromosome_3	7389112	7391835 +	11371
Chromosome_3	7546697	7547559 -	11426
Chromosome_3	7622856	7623569 +	11455
Chromosome_2	23588	24550 +	11593
Chromosome_1	106402	106725 -	11650
Chromosome_1	257818	258636 -	11666
Chromosome_1	379575	380260 -	11678
Chromosome_1	539442	539684 +	11693
Chromosome_1	680680	682127 +	11712
Chromosome_1	785494	785908 -	11732
Chromosome_1	808293	808714 -	11735
Chromosome_1	876258	876988 -	11740
Chromosome_1	1093787	1094212 +	11751
Chromosome_1	1193752	1194490 +	11758
Chromosome_1	1613391	1613711 -	11774
Chromosome_1	1656296	1656652 +	11776
Chromosome_1	1825103	1825969 -	11785
Chromosome_1	1873878	1874318 -	11788
Chromosome_1	2311758	2312174 +	11813
Chromosome_1	2356890	2357380 +	11817
Chromosome_1	3105424	3106336 -	11863
Chromosome_1	3235302	3235969 +	11871
Chromosome_1	3315074	3316223 +	11881
Chromosome_1	3523861	3524382 -	11888
Chromosome_1	3922835	3923644 +	11911
Chromosome_1	4287082	4288172 +	11928
Chromosome_1	4712939	4713292 -	11951
Chromosome_1	5129464	5130069 -	11965
Chromosome_1	5261696	5262081 +	11971
Chromosome_1	5724685	5725053 +	12001
Chromosome_1	5814456	5815944 -	12008
Chromosome_1	6613528	6613977 -	12054
Chromosome_1	6665905	6666269 +	12060
Chromosome_1	6986388	6986762 -	12072

Chromosome_1	7113433	7114029 -	12083
Chromosome_1	7288889	7289883 +	12102
Chromosome_1	8815448	8816210 +	12196
Chromosome_2	8478169	8478702 -	12227
Chromosome_2	8473163	8473800 -	12228
Chromosome_2	8185550	8186030 -	12248
Chromosome_2	8108844	8109407 -	12253
Chromosome_2	7898685	7898996 +	12262
Chromosome_2	7837106	7837405 -	12265
Chromosome_2	7563676	7564089 -	12277
Chromosome_2	7280485	7280820 +	12293
Chromosome_2	6950303	6950951 +	12304
Chromosome_2	6814517	6814904 -	12318
Chromosome_2	6637697	6638029 +	12332
Chromosome_2	5834763	5835074 -	12394
Chromosome_2	5819331	5819716 -	12395
Chromosome_2	5108548	5109148 +	12437
Chromosome_2	4951165	4951335 -	12450
Chromosome_2	4836523	4836976 +	12457
Chromosome_2	4795147	4795690 +	12462
Chromosome_2	4770780	4771158 -	12466
Chromosome_2	4760108	4760615 -	12467
Chromosome_2	4756029	4756668 +	12468
Chromosome_2	4685318	4685818 -	12473
Chromosome_2	4673517	4673742 -	12476
Chromosome_2	4337259	4337758 +	12500
Chromosome_2	3582679	3583045 +	12564
Chromosome_2	3467999	3468656 +	12575
Chromosome_3	632512	633646 -	12652
Chromosome_3	1028817	1029167 +	12673
Chromosome_3	1344801	1345536 +	12695
Chromosome_3	1755376	1755910 +	12707
Chromosome_3	1827754	1828436 +	12709
Chromosome_3	1984313	1984912 +	12718
Chromosome_3	1985804	1986155 +	12719
Chromosome_3	2004903	2005326 +	12721
Chromosome_3	2111274	2111618 -	12727
Chromosome_3	2133320	2134078 -	12730
Chromosome_3	2266289	2266657 +	12739
Chromosome_3	2344195	2344757 -	12740
Chromosome_3	2484726	2485322 -	12749
Chromosome_3	2579891	2580427 -	12752
Chromosome_3	3045916	3046123 +	12780
Chromosome_3	3211576	3211904 -	12797
Chromosome_3	3287828	3288130 +	12803
Chromosome_3	3517235	3517684 -	12822
Chromosome_3	3524755	3525135 -	12824
Chromosome_3	3893504	3893920 -	12847
Chromosome_3	4158216	4158533 -	12861
Chromosome_3	4402823	4403488 +	12872

Chromosome_4	175881	176201 +	12924
Chromosome_4	200513	200944 +	12930
Chromosome_4	555254	555426 +	12944
Chromosome_4	727122	727616 -	12957
Chromosome_4	731947	732240 -	12958
Chromosome_4	777572	777928 -	12959
Chromosome_4	859974	860306 +	12964
Chromosome_4	1089873	1090250 +	12982
Chromosome_4	1178478	1179160 -	12988
Chromosome_4	1262518	1262919 -	12990
Chromosome_4	1300816	1301590 +	12992
Chromosome_4	1668553	1669283 -	13016
Chromosome_4	1754722	1755373 +	13020
Chromosome_4	1790186	1790644 -	13023
Chromosome_4	2138022	2138710 -	13049
Chromosome_4	2222271	2223172 -	13055
Chromosome_4	2420662	2420907 +	13062
Chromosome_4	2632785	2633172 -	13083
Chromosome_4	2727565	2728247 -	13090
Chromosome_4	2830534	2831236 -	13101
Chromosome_4	2845000	2845383 -	13102
Chromosome_4	3095505	3095917 +	13116
Chromosome_4	3099250	3099576 -	13117
Chromosome_4	3150222	3150602 +	13121
Chromosome_4	3205072	3205473 -	13126
Chromosome_4	3392330	3393269 -	13143
Chromosome_4	3412357	3412686 +	13144
Chromosome_4	3969042	3969831 +	13168
Chromosome_4	4117198	4117503 +	13178
Chromosome_4	4724339	4724656 +	13235
Chromosome_4	4745562	4745939 +	13237
Chromosome_4	4884291	4884945 +	13249
Chromosome_4	4896415	4897152 +	13250
Chromosome_4	4902430	4902992 -	13251
Chromosome_2	3105112	3106122 +	13270
Chromosome_2	3066996	3067508 +	13271
Chromosome_2	3052046	3052538 -	13272
Chromosome_2	2839352	2839729 +	13285
Chromosome_2	2294155	2294818 -	13317
Chromosome_2	2040200	2040606 -	13330
Chromosome_2	2035725	2036051 +	13331
Chromosome_2	1632319	1632786 -	13351
Chromosome_2	1506546	1507014 +	13359
Chromosome_2	1418129	1418668 -	13362
Chromosome_2	1151229	1151588 -	13376
Chromosome_2	1098898	1099215 -	13387
Chromosome_2	870398	871106 +	13405
Chromosome_2	864594	865195 +	13406
Chromosome_2	844151	846538 -	13407
Chromosome_2	496512	496853 +	13434

Chromosome_2	462777	463597 -	13437
Chromosome_2	277477	277806 +	13447
Chromosome_2	142840	143349 +	13460
Chromosome_2	96793	97065 +	13464
Chromosome_4	7741035	7741694 +	13506
Chromosome_4	7614106	7614555 -	13512
Chromosome_4	7496373	7497273 -	13520
Chromosome_4	7316692	7317429 +	13529
Chromosome_4	7203460	7203777 +	13533
Chromosome_4	7173371	7173962 +	13537
Chromosome_4	7013684	7014424 -	13548
Chromosome_4	6581952	6582394 -	13577
Chromosome_4	6332315	6332860 +	13598
Chromosome_4	6328559	6329062 +	13599
Chromosome_4	6189047	6189505 -	13602
Chromosome_4	6109045	6109523 +	13608
Chromosome_4	5408572	5409000 +	13639
Chromosome_4	5381687	5382302 +	13642
Chromosome_4	5171150	5171554 +	13656
Chromosome_1	9427572	9428447 +	13677
Chromosome_1	9446906	9447360 +	13679
Chromosome_1	9610258	9610893 -	13690
Chromosome_1	9717128	9718137 +	13696
Chromosome_1	9896766	9897806 +	13706
Chromosome_1	10259840	10260383 +	13721
Chromosome_1	10314527	10314957 -	13724
Chromosome_1	10355123	10355724 -	13727
Chromosome_1	10801775	10802483 -	13753
Chromosome_1	11010751	11011216 +	13768
Chromosome_1	11013076	11013445 -	13769
Chromosome_1	11019657	11019971 -	13771
Chromosome_1	11165511	11165831 +	13779
Chromosome_1	11420039	11420395 -	13794
Chromosome_1	11728645	11729228 -	13818
Chromosome_3	6917738	6918366 +	13833
Chromosome_3	6482451	6482630 +	13863
Chromosome_3	6412398	6412640 -	13869
Chromosome_3	6407733	6408065 +	13872
Chromosome_3	6202567	6203623 -	13882
Chromosome_3	6168788	6169275 +	13885
Chromosome_3	6089034	6089914 -	13893
Chromosome_3	6016924	6017429 +	13895
Chromosome_3	5864911	5865344 -	13907
Chromosome_3	5817349	5818106 -	13909
Chromosome_3	5682365	5683249 +	13919
Chromosome_3	5594081	5594633 -	13921
Chromosome_3	7387843	7388619 -	13977
Chromosome_3	7112853	7113158 +	13997
Chromosome_2	8920574	8921637 -	14013
Chromosome_2	8889209	8889634 +	14014

supercontig_3_12	4606	5349 -	14025
supercontig_3_12	1227	2321 -	14026
Chromosome_1	849900	850115 -	15005
Chromosome_2	716996	717292 +	15034
Chromosome_4	6812745	6813184 -	15037
Chromosome_1	298549	298839 +	15048
Chromosome_1	534188	534452 +	15053
Chromosome_1	6679149	6679445 +	15059
Chromosome_1	1263453	1263749 +	15067
Chromosome_1	1636973	1637266 -	15070
Chromosome_1	1705944	1706237 +	15071
Chromosome_1	1868400	1868555 +	15073
Chromosome_1	1876371	1876553 -	15074
Chromosome_1	2281954	2282445 -	15080
Chromosome_1	2545783	2546043 +	15095
Chromosome_1	2986704	2987036 -	15098
Chromosome_1	3098514	3098818 +	15104
Chromosome_4	568023	568214 -	15106
Chromosome_1	3140701	3140893 -	15109
Chromosome_1	4038362	4038791 +	15118
Chromosome_1	4178605	4178856 -	15120
Chromosome_1	4258972	4259250 +	15121
Chromosome_3	49462	49755 -	15122
Chromosome_4	5631716	5631895 -	15124
Chromosome_1	4631589	4631931 -	15125
Chromosome_1	4785278	4785553 -	15127
Chromosome_1	4906502	4906860 -	15128
Chromosome_1	5773629	5774339 -	15133
Chromosome_1	7110494	7110883 -	15140
Chromosome_1	7875956	7876191 -	15143
Chromosome_1	8307183	8307341 +	15149
Chromosome_2	3876671	3876887 +	15164
Chromosome_2	3986026	3986302 +	15168
Chromosome_2	4639344	4639601 +	15177
Chromosome_2	4749611	4750069 +	15179
Chromosome_2	5050616	5050858 +	15183
Chromosome_2	5055838	5056428 +	15184
Chromosome_2	5204763	5205639 +	15186
Chromosome_2	5294173	5294389 -	15191
Chromosome_2	5455766	5455926 +	15192
Chromosome_2	5560302	5560597 -	15196
Chromosome_2	5688288	5688702 +	15198
Chromosome_2	5909645	5909998 +	15202
Chromosome_2	5977260	5977523 -	15204
Chromosome_2	6276837	6277010 +	15211
Chromosome_2	6663720	6664208 -	15214
Chromosome_2	6396892	6397098 +	15216
Chromosome_2	6457714	6458140 +	15217
Chromosome_2	6875472	6875798 -	15222
Chromosome_2	6884407	6884587 +	15223

Chromosome_2	6889427	6889606 +	15224
Chromosome_2	6978138	6978624 +	15227
Chromosome_2	7449328	7449788 -	15232
Chromosome_2	7691309	7691865 +	15236
Chromosome_2	7767161	7767370 -	15239
Chromosome_2	8012733	8012977 -	15244
Chromosome_2	8017619	8017984 -	15245
Chromosome_2	8146218	8146642 -	15248
Chromosome_2	8216669	8216851 -	15249
Chromosome_2	8230614	8230959 +	15250
Chromosome_2	8527286	8527489 -	15251
Chromosome_2	8731141	8731508 +	15257
Chromosome_2	8579199	8579713 +	15258
Chromosome_3	502119	502357 -	15261
Chromosome_3	568803	569057 -	15262
Chromosome_3	748349	748570 +	15268
Chromosome_3	753758	754018 -	15270
Chromosome_3	869745	870304 -	15272
Chromosome_3	871468	871760 -	15273
Chromosome_3	872342	872515 +	15274
Chromosome_3	1083449	1083977 -	15278
Chromosome_3	1123423	1123578 -	15279
Chromosome_3	1452503	1452926 +	15290
Chromosome_3	1953668	1953938 -	15293
Chromosome_3	2645069	2645592 +	15306
Chromosome_3	3017024	3017235 -	15310
Chromosome_3	3170389	3170606 -	15314
Chromosome_3	3555055	3555238 -	15319
Chromosome_3	5023191	5023412 +	15345
Chromosome_4	344813	345156 +	15354
Chromosome_4	129329	129784 +	15357
Chromosome_4	163239	163595 +	15361
Chromosome_4	208809	209103 +	15363
Chromosome_4	261293	261564 -	15367
Chromosome_4	638911	639095 +	15374
Chromosome_4	925648	925935 +	15376
Chromosome_4	935848	936148 -	15377
Chromosome_4	1019860	1020154 +	15381
Chromosome_4	1049741	1050012 -	15382
Chromosome_4	1720239	1720617 -	15398
Chromosome_4	1758552	1758886 +	15400
Chromosome_4	1759131	1759460 -	15401
Chromosome_4	2216521	2216868 -	15409
Chromosome_4	2764990	2765422 -	15417
Chromosome_4	3194412	3194624 -	15421
Chromosome_4	3430252	3430542 -	15427
Chromosome_4	3805232	3805432 -	15434
Chromosome_4	3933106	3933397 +	15436
Chromosome_4	4382997	4383173 +	15438
Chromosome_4	4444284	4444515 +	15439

Chromosome_4	4447824	4448069 +	15440
Chromosome_4	4452971	4453314 -	15441
Chromosome_4	4244202	4244523 -	15443
Chromosome_4	4725758	4726325 +	15452
Chromosome_4	4779235	4779405 -	15453
Chromosome_4	4577359	4577734 -	15456
Chromosome_2	98299	98593 -	15458
Chromosome_2	148178	148360 +	15459
Chromosome_2	44998	45269 -	15462
Chromosome_2	456506	456658 +	15464
Chromosome_2	885363	885671 -	15466
Chromosome_2	717532	717711 -	15470
Chromosome_2	996334	996527 -	15475
Chromosome_2	1099456	1100083 +	15477
Chromosome_2	1154035	1154262 +	15478
Chromosome_2	1180541	1180756 +	15479
Chromosome_2	1508509	1508676 -	15488
Chromosome_2	1557303	1557586 -	15489
Chromosome_2	1818001	1818174 -	15494
Chromosome_2	1884011	1884248 -	15495
Chromosome_2	2057537	2057797 -	15497
Chromosome_2	2115244	2115441 +	15499
Chromosome_2	2416909	2417133 +	15504
Chromosome_2	2574807	2575016 -	15510
Chromosome_2	2855106	2855273 -	15516
Chromosome_2	2782652	2782804 +	15517
Chromosome_2	2904827	2905072 +	15518
Chromosome_2	2906216	2906649 +	15519
Chromosome_2	2946602	2947090 -	15520
Chromosome_2	2878417	2878701 +	15521
Chromosome_2	2984599	2984932 +	15522
Chromosome_2	2985780	2986035 -	15523
Chromosome_2	3072461	3072649 -	15527
Chromosome_4	7888924	7889122 +	15534
Chromosome_4	7969380	7969592 -	15535
Chromosome_4	7572163	7572659 -	15538
Chromosome_4	6609602	6609943 -	15549
Chromosome_4	5963343	5963775 +	15557
Chromosome_4	5350157	5350398 -	15559
Chromosome_4	5213722	5214214 +	15561
Chromosome_1	9404027	9404419 +	15568
Chromosome_1	9519967	9520164 +	15570
Chromosome_1	9531546	9531797 +	15571
Chromosome_1	9532394	9532650 +	15572
Chromosome_1	9870832	9871020 -	15576
Chromosome_1	10137056	10137780 +	15583
Chromosome_1	10231729	10232166 +	15585
Chromosome_1	10789055	10789280 -	15600
Chromosome_1	10806866	10807096 +	15602
Chromosome_1	10910815	10911091 +	15604

Chromosome_1	11051450	11051768 +	15606
Chromosome_1	11127463	11127867 -	15611
Chromosome_1	11130845	11131108 -	15612
Chromosome_1	11485077	11485603 -	15616
Chromosome_3	5597880	5598309 -	15623
Chromosome_3	5638527	5638703 +	15627
Chromosome_3	6394605	6394826 -	15631
Chromosome_3	6716029	6716408 -	15633
Chromosome_3	6792996	6793445 +	15637
Chromosome_3	6903923	6904694 +	15639
Chromosome_3	7370041	7370286 -	15648
Chromosome_3	7431136	7431336 +	15651
Chromosome_3	7545745	7546026 +	15652
Chromosome_3	7711250	7711517 -	15655
Chromosome_2	8976040	8976198 +	15659
Chromosome_2	8932256	8932478 -	15660
Chromosome_1	21463	22193 +	15662
Chromosome_1	526448	526690 +	15699
Chromosome_1	704893	706203 +	15709
Chromosome_1	4222804	4224070 -	15840
Chromosome_1	7151561	7151887 -	15949
Chromosome_1	7157128	7158717 -	15950
Chromosome_2	6903789	6904206 +	16151
Chromosome_2	5979315	5979733 -	16210
Chromosome_2	3560654	3560921 +	16382
Chromosome_3	2146423	2146755 +	16480
Chromosome_4	281233	281499 +	16671
Chromosome_4	5023023	5023574 +	16933
Chromosome_1	11643151	11644314 +	17433
Chromosome_3	7046841	7048625 +	17587
Chromosome_2	6902169	6902618 -	17620
Chromosome_3	1143327	1143722 -	17647
Chromosome_1	138963	139465 +	20002
Chromosome_1	516693	517255 +	20007
Chromosome_1	944042	944738 +	20008
Chromosome_1	1091180	1091376 +	20009
Chromosome_1	1307846	1308441 -	20010
Chromosome_1	1434958	1435251 -	20012
Chromosome_1	1524747	1525269 -	20013
Chromosome_1	1627218	1627457 -	20014
Chromosome_1	1638936	1639211 +	20015
Chromosome_1	2136768	2137001 -	20017
Chromosome_1	2294123	2294335 +	20018
Chromosome_1	2294383	2294604 -	20019
Chromosome_1	2818235	2818623 -	20020
Chromosome_1	2914469	2914696 +	20021
Chromosome_1	3334099	3334432 -	20022
Chromosome_1	3432538	3432732 +	20023
Chromosome_1	3559431	3559622 -	20024
Chromosome_1	3610128	3610534 +	20025

Chromosome_1	3793854	3794115 +	20027
Chromosome_1	4017300	4017539 +	20029
Chromosome_1	4049510	4049694 +	20032
Chromosome_1	4113763	4114409 +	20033
Chromosome_1	4160625	4161052 +	20036
Chromosome_1	4195947	4196183 +	20037
Chromosome_1	4313080	4314790 -	20038
Chromosome_1	4380004	4380283 -	20039
Chromosome_1	4479903	4480198 -	20040
Chromosome_1	4545689	4546093 -	20041
Chromosome_1	4798376	4799082 -	20043
Chromosome_1	5101349	5101561 -	20044
Chromosome_1	5187067	5187301 -	20045
Chromosome_1	5230545	5230727 -	20046
Chromosome_1	5249341	5249828 -	20047
Chromosome_1	5255329	5255648 +	20049
Chromosome_1	5284329	5284538 +	20050
Chromosome_1	5400965	5401596 +	20051
Chromosome_1	5402914	5403629 -	20052
Chromosome_1	5469084	5469263 -	20054
Chromosome_1	5559782	5560121 -	20055
Chromosome_1	5683691	5683957 +	20057
Chromosome_1	6067965	6068144 +	20058
Chromosome_1	6090203	6090694 -	20059
Chromosome_1	6100715	6101474 -	20060
Chromosome_1	6299067	6299291 -	20062
Chromosome_1	6735963	6736943 +	20065
Chromosome_1	6783736	6784065 +	20066
Chromosome_1	7039832	7040080 -	20068
Chromosome_1	7045882	7046124 +	20069
Chromosome_1	7208043	7208198 +	20071
Chromosome_1	7274138	7274374 +	20072
Chromosome_1	8461042	8461293 +	20080
Chromosome_1	8653321	8653524 -	20081
Chromosome_1	8668272	8668927 +	20082
Chromosome_1	8682665	8682931 +	20083
Chromosome_1	9119894	9120366 -	20085
Chromosome_1	9287306	9287497 +	20087
Chromosome_1	9796259	9796694 +	20089
Chromosome_1	9835829	9835995 +	20090
Chromosome_1	10038291	10038515 -	20092
Chromosome_1	10057127	10057309 +	20093
Chromosome_1	10133104	10133735 +	20094
Chromosome_1	10424750	10425004 +	20096
Chromosome_1	10518393	10518531 +	20097
Chromosome_1	10714964	10715226 -	20099
Chromosome_1	10806231	10806597 -	20100
Chromosome_1	11485520	11485949 +	20103
Chromosome_1	11737013	11737303 -	20106
Chromosome_2	57477	57713 -	20109

Chromosome_2	65607	66745 +	20110
Chromosome_2	210806	211024 +	20111
Chromosome_2	324692	325003 +	20115
Chromosome_2	625746	626162 +	20118
Chromosome_2	967922	968158 -	20119
Chromosome_2	1016979	1017236 -	20121
Chromosome_2	1052041	1053100 -	20122
Chromosome_2	1149615	1149857 -	20123
Chromosome_2	1197391	1197582 +	20125
Chromosome_2	1270244	1270456 +	20127
Chromosome_2	1322434	1322763 -	20129
Chromosome_2	1343898	1344396 -	20130
Chromosome_2	1481550	1481804 -	20131
Chromosome_2	1691140	1691463 -	20134
Chromosome_2	1934838	1935023 -	20135
Chromosome_2	1982729	1984284 -	20136
Chromosome_2	2104180	2104528 -	20137
Chromosome_2	2126944	2127150 +	20138
Chromosome_2	2354259	2354438 +	20140
Chromosome_2	2377470	2377691 +	20141
Chromosome_2	2666532	2666792 +	20142
Chromosome_2	2728032	2728351 +	20144
Chromosome_2	2763065	2763265 -	20145
Chromosome_2	2840707	2841051 +	20147
Chromosome_2	2844452	2844961 -	20148
Chromosome_2	2900842	2901425 +	20149
Chromosome_2	2905929	2906135 -	20150
Chromosome_2	3114773	3115029 -	20154
Chromosome_2	3257823	3258098 -	20156
Chromosome_2	3389810	3390103 -	20157
Chromosome_2	3407182	3407376 +	20158
Chromosome_2	3564390	3564737 +	20159
Chromosome_2	3586325	3586528 -	20160
Chromosome_2	3601355	3601897 +	20161
Chromosome_2	3948880	3949205 -	20164
Chromosome_2	4191475	4191660 -	20165
Chromosome_2	4290279	4290506 -	20166
Chromosome_2	4760117	4760362 +	20169
Chromosome_2	5478887	5479087 +	20174
Chromosome_2	5493236	5493572 +	20175
Chromosome_2	5683613	5683807 -	20177
Chromosome_2	5846816	5847013 -	20181
Chromosome_2	6403659	6403882 +	20185
Chromosome_2	6408497	6408924 -	20186
Chromosome_2	6603576	6603767 +	20187
Chromosome_2	6612917	6613333 -	20189
Chromosome_2	6668121	6668404 -	20190
Chromosome_2	6867161	6867427 +	20194
Chromosome_2	6889237	6889440 +	20195
Chromosome_2	6974566	6974784 -	20196

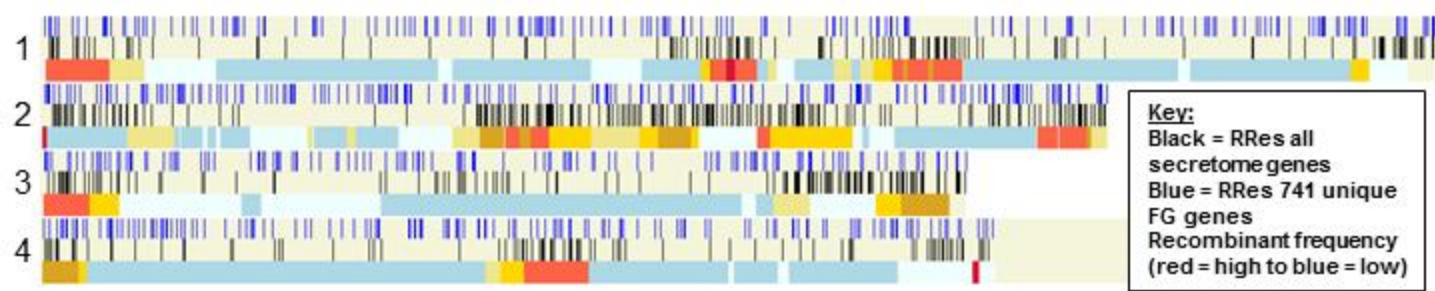
Chromosome_2	7208392	7211433 -	20197
Chromosome_2	7218362	7218634 +	20198
Chromosome_2	7366827	7367573 +	20200
Chromosome_2	7436510	7436704 +	20201
Chromosome_2	7593004	7594973 +	20202
Chromosome_2	7669495	7669755 -	20203
Chromosome_2	7747142	7747360 +	20204
Chromosome_2	7947985	7948359 -	20206
Chromosome_2	7960591	7960821 +	20207
Chromosome_2	8101240	8102433 +	20209
Chromosome_2	8102374	8102679 -	20210
Chromosome_2	8208958	8209311 -	20212
Chromosome_2	8217972	8218085 +	20213
Chromosome_2	8261281	8261484 +	20214
Chromosome_2	8298352	8298762 +	20215
Chromosome_2	8652655	8652981 +	20217
Chromosome_2	8910332	8910894 -	20220
Chromosome_2	8922231	8922593 -	20221
Chromosome_2	8978246	8978758 -	20222
Chromosome_3	30235	30441 -	20224
Chromosome_3	52810	53292 +	20225
Chromosome_3	54124	54315 -	20226
Chromosome_3	84325	84429 -	20227
Chromosome_3	239373	239723 -	20230
Chromosome_3	350081	350362 +	20232
Chromosome_3	426932	427492 -	20235
Chromosome_3	460534	460981 -	20237
Chromosome_3	480251	480439 +	20238
Chromosome_3	711384	712536 +	20242
Chromosome_3	882350	882541 +	20244
Chromosome_3	1113534	1113758 +	20250
Chromosome_3	1407247	1407637 -	20252
Chromosome_3	1862592	1867177 +	20256
Chromosome_3	1874110	1878283 -	20257
Chromosome_3	1991553	1992076 +	20259
Chromosome_3	2138162	2138377 +	20262
Chromosome_3	2496926	2497198 -	20265
Chromosome_3	2578260	2578961 +	20266
Chromosome_3	2713326	2713710 +	20267
Chromosome_3	2768860	2769163 -	20268
Chromosome_3	2986597	2986952 -	20269
Chromosome_3	3068505	3068741 -	20270
Chromosome_3	3129573	3129766 +	20271
Chromosome_3	3521707	3521925 -	20275
Chromosome_3	3629133	3629741 -	20277
Chromosome_3	3647890	3648138 +	20278
Chromosome_3	4234047	4234280 +	20284
Chromosome_3	4266939	4267340 +	20285
Chromosome_3	4267971	4268164 -	20286
Chromosome_3	4288396	4288720 -	20287

Chromosome_3	4449169	4452302 -	20290
Chromosome_3	4454409	4454609 +	20291
Chromosome_3	4510774	4511094 -	20292
Chromosome_3	4640778	4641106 -	20295
Chromosome_3	4780416	4782555 +	20296
Chromosome_3	5709740	5709955 +	20298
Chromosome_3	5928972	5929190 -	20300
Chromosome_3	5972078	5972326 -	20301
Chromosome_3	5992546	5992848 +	20302
Chromosome_3	6086844	6087077 -	20303
Chromosome_3	6107776	6108146 -	20305
Chromosome_3	6390818	6391162 +	20307
Chromosome_3	6464627	6464902 -	20308
Chromosome_3	6474416	6474804 -	20309
Chromosome_3	6585499	6585964 +	20311
Chromosome_3	6667873	6668208 +	20312
Chromosome_3	6742412	6742754 -	20313
Chromosome_3	7049827	7050063 -	20315
Chromosome_3	7236930	7237121 -	20318
Chromosome_3	7370747	7371216 +	20320
Chromosome_3	7560965	7562096 -	20321
Chromosome_3	7706113	7706539 +	20324
Chromosome_3	7789097	7789770 -	20328
Chromosome_4	15034	15517 +	20329
Chromosome_4	62366	62646 +	20331
Chromosome_4	378130	378369 +	20333
Chromosome_4	378929	379113 -	20334
Chromosome_4	486345	489652 +	20335
Chromosome_4	580283	580540 +	20337
Chromosome_4	591995	592183 -	20338
Chromosome_4	593911	594629 +	20339
Chromosome_4	675898	676125 +	20342
Chromosome_4	811529	812028 -	20343
Chromosome_4	958080	958304 -	20344
Chromosome_4	1007432	1008679 +	20345
Chromosome_4	1136888	1136998 -	20348
Chromosome_4	1203622	1203933 +	20349
Chromosome_4	1774398	1774685 -	20350
Chromosome_4	1870178	1871319 +	20351
Chromosome_4	2220256	2220450 -	20352
Chromosome_4	2268551	2268822 +	20353
Chromosome_4	2530216	2530413 +	20355
Chromosome_4	2817603	2818111 +	20359
Chromosome_4	3135003	3135375 -	20361
Chromosome_4	3186148	3186479 -	20362
Chromosome_4	3390224	3390839 -	20365
Chromosome_4	3412720	3413321 +	20366
Chromosome_4	3486746	3486964 +	20367
Chromosome_4	3501015	3501294 -	20368
Chromosome_4	3519231	3520495 +	20369

Chromosome_4	3556163	3556342 -	20370
Chromosome_4	3683556	3683735 -	20371
Chromosome_4	3741513	3741740 +	20372
Chromosome_4	4172925	4173502 -	20374
Chromosome_4	4316929	4317215 -	20375
Chromosome_4	4600970	4601164 +	20381
Chromosome_4	5413865	5414611 -	20382
Chromosome_4	5597936	5598579 +	20383
Chromosome_4	5928357	5928954 +	20385
Chromosome_4	6166141	6166356 +	20387
Chromosome_4	6368530	6368871 -	20388
Chromosome_4	6452497	6452706 +	20389
Chromosome_4	6848504	6848743 +	20391
Chromosome_4	6868131	6868409 +	20392
Chromosome_4	7094486	7094719 +	20393
Chromosome_4	7171774	7172121 -	20395
Chromosome_4	7199591	7199841 +	20396
Chromosome_4	7203889	7204071 +	20397
Chromosome_4	7293457	7293645 +	20398
Chromosome_4	7336939	7337166 -	20401
Chromosome_4	7383815	7384006 +	20402
Chromosome_4	7937554	7939330 -	20407
supercontig_3_12	1126	1230 +	20410
Chromosome_4	8010523	8010651 +	20412
Chromosome_1	3241762	3242739 +	00980_M
Chromosome_1	5052817	5053418 +	01528_M
Chromosome_2	5316027	5317377 +	03503_M
Chromosome_2	8229161	8230630 -	04498_M
Chromosome_2	8338216	8339080 +	04532_M
Chromosome_2	8339572	8340914 -	04533_M
Chromosome_2	8588273	8589087 +	04623_M
Chromosome_2	8663069	8663836 +	04649_M
Chromosome_3	28922	30832 +	04696_M
Chromosome_3	222558	224450 -	04760_M
Chromosome_4	4360256	4361099 -	07741_M
Chromosome_4	4866630	4866987 +	07919_M
Chromosome_2	317493	319301 +	08073_M
Chromosome_2	849471	851825 -	08262_M
Chromosome_2	1675223	1676392 -	08512_M
Chromosome_4	7887525	7888614 -	08990_M
Chromosome_4	7726953	7729163 +	09044_M
Chromosome_4	7633555	7634624 -	09079_M
Chromosome_1	10678940	10679431 +	10370_M
Chromosome_1	10811390	10812723 +	10408_M
Chromosome_1	10930288	10930836 -	10457_M
Chromosome_1	11660431	11662423 +	10689_M
Chromosome_1	11681905	11682599 -	10698_M
Chromosome_1	1558559	1558780 +	11772_M
Chromosome_1	2136364	2137195 +	11801_M
Chromosome_1	2753545	2754006 -	11838_M

Chromosome_1	3811809	3812269 -	11907_M
Chromosome_1	4201628	4203149 -	11923_M
Chromosome_1	7312275	7312895 -	12103_M
Chromosome_1	8015027	8015899 -	12148_M
Chromosome_1	8164337	8164799 -	12157_M
Chromosome_1	8647078	8648280 -	12177_M
Chromosome_3	460519	461104 +	12639_M
Chromosome_3	462174	462644 -	12640_M
Chromosome_3	748952	749285 +	12654_M
Chromosome_3	1826014	1826484 -	12708_M
Chromosome_3	2895194	2895409 -	12770_M
Chromosome_3	4659273	4659653 +	12891_M
Chromosome_4	198083	198364 -	12929_M
Chromosome_4	289745	290056 +	12933_M
Chromosome_4	1043135	1043466 +	12981_M
Chromosome_4	4332072	4333571 +	13207_8_M
Chromosome_2	2088888	2089527 -	13327_M
Chromosome_2	1144952	1145199 -	13378_M
Chromosome_4	6892182	6892594 -	13559_M
Chromosome_4	5379498	5380513 -	13643_M
Chromosome_1	10413964	10414787 -	13732_M
Chromosome_1	10797858	10798451 +	13751_M
Chromosome_3	6368857	6369254 +	13877_M
Chromosome_3	6012103	6014061 +	13896_M
Chromosome_3	5843416	5843744 -	13908_M
Chromosome_1	581027	581359 -	15054_M
Chromosome_1	2546951	2547148 -	15096_M
Chromosome_1	3765472	3765603 -	15116_M
Chromosome_2	3772699	3772892 -	15160_M
Chromosome_2	3728285	3728413 +	15163_M
Chromosome_3	1375509	1375903 +	15283_M
Chromosome_3	3175259	3175401 -	15315_M
Chromosome_3	4415567	4415872 -	15332_M
Chromosome_3	4835095	4835330 +	15340_M
Chromosome_4	1540647	1540808 +	15394_M
Chromosome_4	3732682	3732851 -	15432_M
Chromosome_2	2495715	2496167 -	15508_M
Chromosome_1	10986986	10987219 -	15605_M
Chromosome_1	11072650	11073017 -	15607_M
Chromosome_3	5643785	5643922 +	15624_M
Chromosome_1	70921	71208 -	15677_M
Chromosome_4	1945523	1945798 +	16203_M
Chromosome_2	548397	548667 -	15465_M
Chromosome_3	7711250	7711517 -	15655

1e RRes v4.0 gene annotation



The figure displays genomic tracks for two strains, 04289 and 0421_M, across a range of positions from 1 to 900. The tracks are organized into four main sections, each showing a reference sequence (04289) and a comparison sequence (0421_M).

- Section 1 (Positions 1-150):** Shows the CDS (Coding DNA Sequence) for both strains. The reference sequence (04289) is shown in blue, and the comparison sequence (0421_M) is shown in red. The CDS is indicated by a blue arrow pointing right.
- Section 2 (Positions 160-310):** Shows the CDS for both strains. The reference sequence (04289) is shown in blue, and the comparison sequence (0421_M) is shown in red. The CDS is indicated by a blue arrow pointing right.
- Section 3 (Positions 320-460):** Shows the CDS for both strains. The reference sequence (04289) is shown in blue, and the comparison sequence (0421_M) is shown in red. The CDS is indicated by a blue arrow pointing right.
- Section 4 (Positions 470-620):** Shows the CDS for both strains. The reference sequence (04289) is shown in blue, and the comparison sequence (0421_M) is shown in red. The CDS is indicated by a blue arrow pointing right.
- Section 5 (Positions 630-770):** Shows the five_prime_UTR (5' Untranslated Region) for both strains. The reference sequence (04289) is shown in blue, and the comparison sequence (0421_M) is shown in red. The five_prime_UTR is indicated by a blue arrow pointing right.
- Section 6 (Positions 780-900):** Shows the CDS for both strains. The reference sequence (04289) is shown in blue, and the comparison sequence (0421_M) is shown in red. The CDS is indicated by a blue arrow pointing right.

The sequences are color-coded by amino acid: A (blue), C (green), G (yellow), T (red), and N (purple). The CDS and five_prime_UTR regions are highlighted with blue arrows. The comparison sequence (0421_M) is shown in red, and the reference sequence (04289) is shown in blue.

Additional file 18. RRes *F. graminearum* detailed statistics of chromosome arms.

Chromosome	Genes	FG unique	Secretome	GC (%)	C E N T R O M E R E	Genes	FG unique	Secretome	GC (%)	Total GC content (%)*
1	3,354	143	111	48.5		1,039	66	34	48.2	48.3
2	1,250	99	42	48.1		2,395	130	192	47.8	47.6
3	2,104	102	61	48.6		986	58	81	47.6	48.0
4	1,937	119	61	48.4		1,095	48	34	49.7 48.6**	48.8
Total	8,645	463	275	48.4		5,515	302	341	48.3 48.1**	48.2 (Avg.)

*Excluding N bases

** Excluding large repetitive sequence at the carboxyl end of chromosome 4

This table excludes the 4 unique genes found on supercontig 3.12

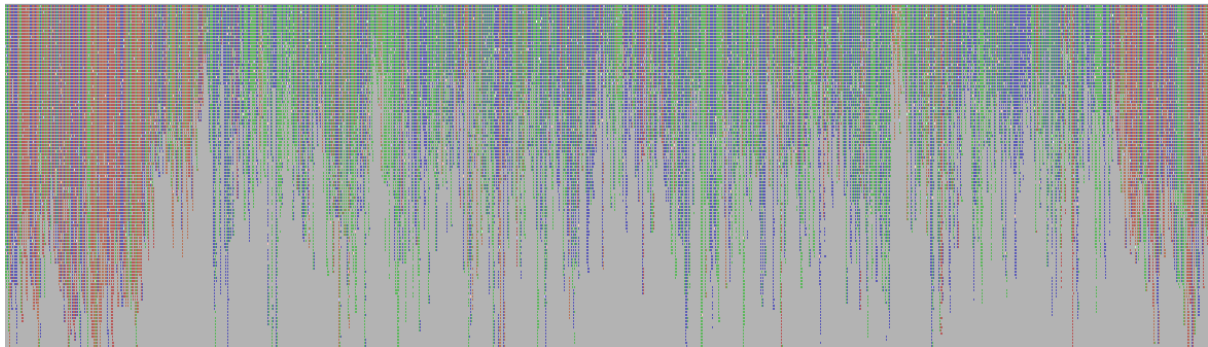
Additional file 19. The composition of the centromere sequences in chromosomes 1-4 and read coverage.

Summary frequency of bases, percentage composition and percentage GC content for each of the centromeres of chromosomes 1-4.

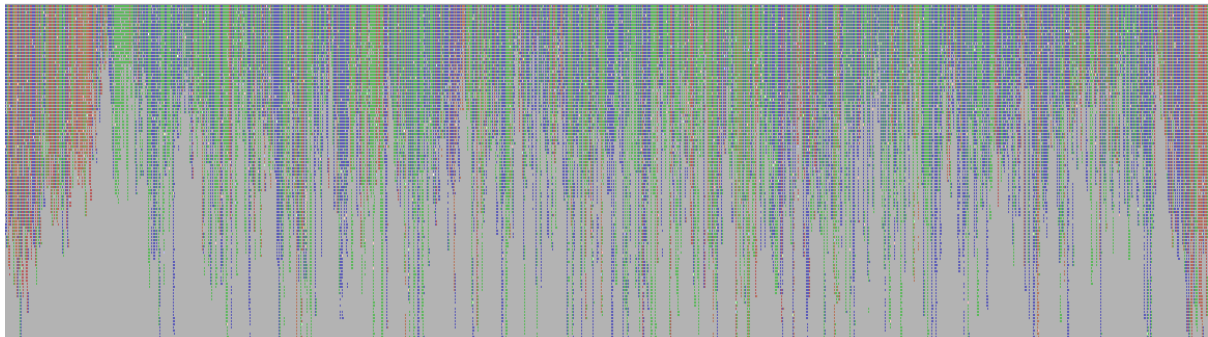
	1		2		3		4	
Composition	Freq	%	Freq	%	Freq	%	Freq	%
A:	24,365	43.10%	29,288	44.90%	23,774	42.20%	27,334	44.90%
C:	3,661	6.50%	4,594	7.00%	3,523	6.20%	3,643	6.00%
G:	3,588	6.30%	4,277	6.60%	3,722	6.60%	3,754	6.20%
T:	24,955	44.10%	27,022	41.50%	25,368	45.00%	26,202	43.00%
N:	12	0.00%	0	0.00%	0	0.00%	0	0.00%
GC:	7,249	12.80%	8,871	13.60%	7,245	12.80%	7,397	12.10%

Image of mapped reads to centromere sequences (75 kbp window). Green = A, Blue = T, Yellow = C, Red = G.

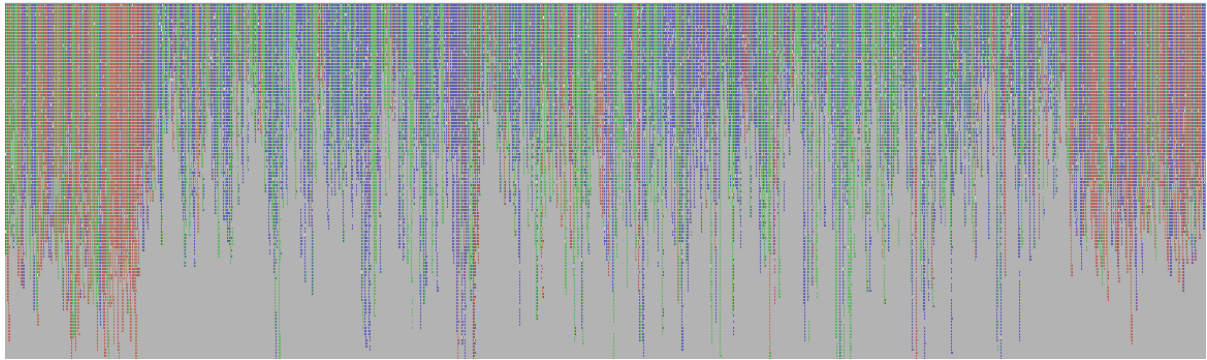
Centromere1



Centromere2



Centromere3



Centromere4

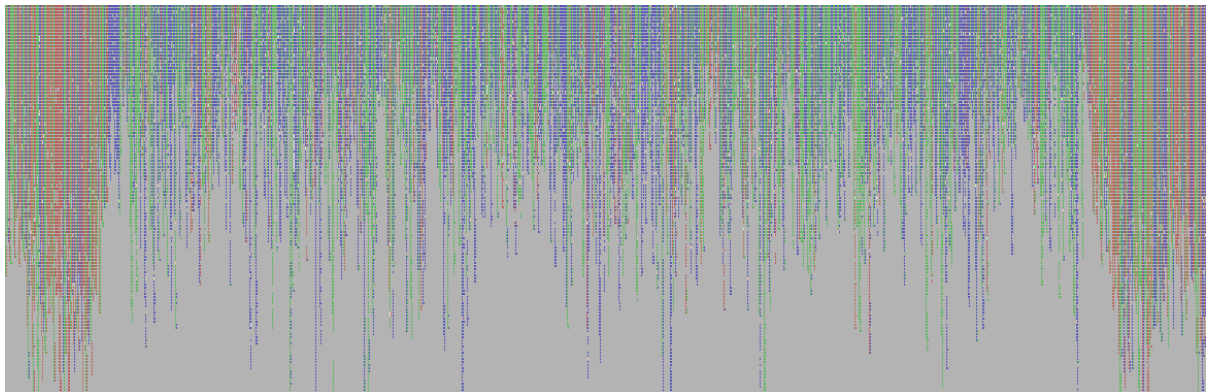
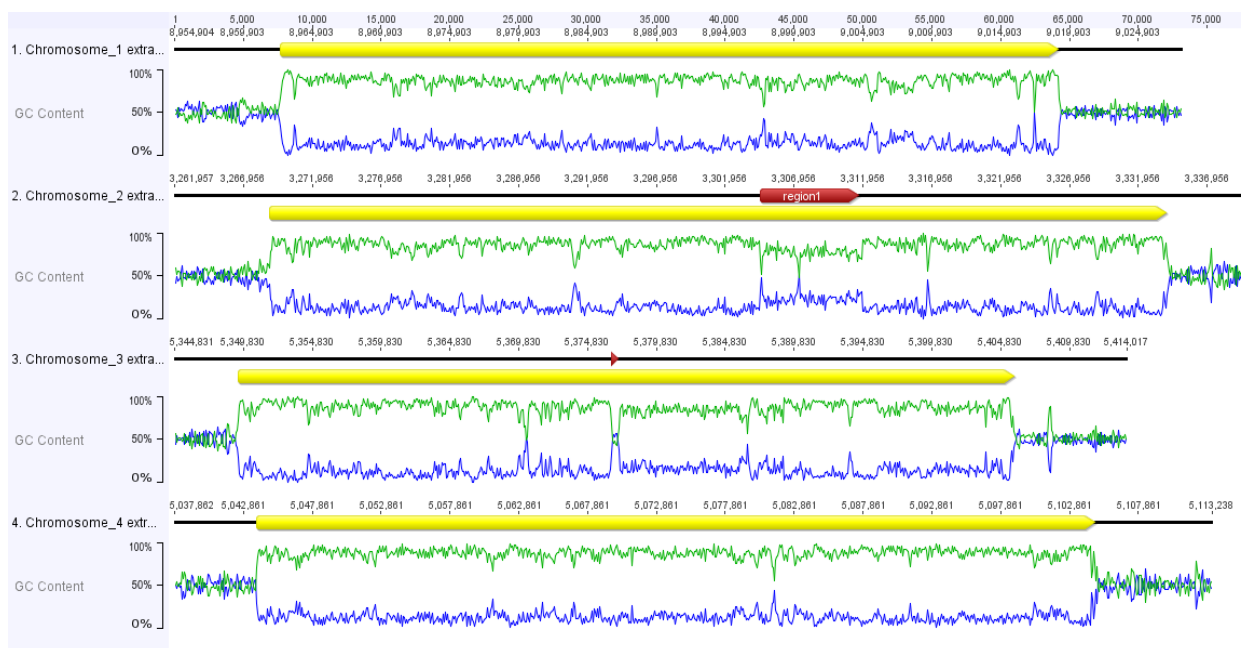


Image of Chromosome regions GC/AT graphs. Yellow represents the proposed centromere sequences and the region annotated with a red arrow represent regions of lower AT content and higher GC. Region 1 of chromosome 2 and region 3 of chromosome 3 represents a region of respectively approximately 7,194 bp and 450 bp and 22.8% and 53.8% GC content.



Additional file 20. Instructions to join the mailing list for the *F. graminearum* RRes v4.1 annotation. Subscribers can submit new or corrected gene models and receive updated annotation versions via custom tracks in Ensembl Fungi.

Joining the mailing list:

Step 1: Copy the text into your browser

"<https://www.lists.rothamsted.ac.uk/mailman/listinfo/fgrresv41>"

Step2: In the "Subscribing to Fgrresv41" section fill in your name, email address and password.

Step3: A confirmation email will be sent to you. Click the confirmation link. Please check spam folder and promotional folder if using gmail.

Step4: To post a message use this email address "fgrresv41@lists.rothamsted.ac.uk". Postings may appear in the forum tab of a gmail account.

Submission of gene models new or curated:

Step 1: send a gff3 file to the mailing group or within the email (if only 1-10 gene models) of your gene models to be curated by the admin. Include in your message any supporting evidence to justify the change. Please see example gff3 below which shows two exons of the gene 11979 on chromosome 1 which is the strand, and the locus_tag which is the name of the gene.

1	.	CDS	5340879	5342780	.	-	.	locus_tag=11979_M;ID=RK1
1	.	CDS	5339808	5340830	.	-	.	locus_tag=11979_M;ID=RK1

The administrator will then transfer the gene model or curate an existing entry to the mailing site set and include in an updated custom track URL for Ensembl Fungi for all users to see. After approx. 6 months periodically an update will be submitted to ENA and subsequently Ensembl Fungi.