Reverse transcriptase-related enzymes are associated with horizontal chromosome transfer in an asexual pathogen

Xiaoqiu Huang, Anindya Das, Binod B Sahu, Subodh K Srivastava, Leonor F Leandro, Kerry O'Donnell, Madan K Bhattacharyya

Supernumerary chromosomes have been shown to transfer horizontally from one isolate to another. However, the mechanism by which horizontal chromosome transfer (HCT) occurs is unknown. In this study, we compared the genomes of 11 isolates comprising six *Fusarium* species that cause soybean sudden death syndrome (SDS) or bean root rot (BRR), and detected numerous instances of HCT in supernumerary chromosomes. We also identified a statistically significant number (21 standard deviations above the mean) of single nucleotide polymorphisms (SNPs) in the supernumerary chromosomes between isolates of the asexual pathogen *F. virguliforme*. Supernumerary chromosomes carried reverse transcriptase-related genes (*RVT*); the presence of long *RVT* open reading frames (ORFs) in the supernumerary chromosome was correlated with the presence of two or more chromosome copies with a significant number of SNPs between them. Our results suggest that supernumerary chromosomes transfer horizontally via an RNA intermediate. Understanding the mechanism by which HCT occurs will have a profound impact on understanding evolution and applying biotechnology as well as accepting HCT as a natural source of genetic variation.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

Reverse transcriptase-related enzymes are associated with horizontal chromosome transfer in an asexual pathogen

Xiaoqiu Huang[1,2,3], Anindya Das[1], Binod B. Sahu[4], Subodh K. Srivastava[5], Leonor F. Leandro[6], Kerry O'Donnell[7], Madan K. Bhattacharyya[4]

[1]Department of Computer Science and [2]Plant Sciences Institute, Iowa State University, Ames, Iowa 50011, USA

[4]Department of Agronomy, Iowa State University, Ames, Iowa 50011, USA

[5]Crop, Soil and Environmental Sciences, University of Arkansas, Fayetteville, Arkansas 72701, USA

[6]Department of Plant Pathology, Iowa State University, Ames, Iowa 50011, USA

[7]National Center for Agricultural Utilization Research, US Department of Agriculture, Agricultural Research Service, Peoria, Illinois 61604, USA

Corresponding Author:

[3]Xiaoqiu Huang

Department of Computer Science, Iowa State University, Ames, Iowa 50011, USA

E-mail: xqhuang@iastate.edu, Phone: (515) 294-2432, Fax: (515) 294-0258

Key Words: Asexual reproduction, Horizontal chromosome transfer, Reverse transcriptase-related enzymes, Duplication-induced mutation

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

Abstract

Supernumerary chromosomes have been shown to transfer horizontally from one isolate to another. However, the mechanism by which horizontal chromosome transfer (HCT) occurs is unknown. In this study, we compared the genomes of 11 isolates comprising six *Fusarium* species that cause soybean sudden death syndrome (SDS) or bean root rot (BRR), and detected numerous instances of HCT in supernumerary chromosomes. We also identified a statistically significant number (21 standard deviations above the mean) of single nucleotide polymorphisms (SNPs) in the supernumerary chromosomes between isolates of the asexual pathogen *F. virguliforme*. Supernumerary chromosomes carried reverse transcriptase-related genes (*RVT*); the presence of long *RVT* open reading frames (ORFs) in the supernumerary chromosome was correlated with the presence of two or more chromosome copies with a significant number of SNPs between them. Our results suggest that supernumerary chromosomes transfer horizontally via an RNA intermediate. Understanding the mechanism by which HCT occurs will have a profound impact on understanding evolution and applying biotechnology as well as accepting HCT as a natural source of genetic variation.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

Introduction

Conventional wisdom holds that asexual organisms lack a mechanism for generating genetic variation. However, evidence of sexual reproduction has not been detected in many important plant pathogens (Masel et al., 1996; Michielse & Rep, 2009). Asexual fungal pathogens are known to have variable electrophoretic karyotypes (Kistler & Miao, 1992). It was shown through pulsed-field gel eletrophoresis that extra nonessential chromosomes (called supernumerary chromosomes) are present only in some individuals of a species (Masel, Irwin & Manners, 1993). It was demonstrated under laboratory conditions that a 2-Mb supernumerary chromosome was transferred between two vegetatively incompatible isolates of an asexual fungus (He et al., 1998). A whole-genome comparative study also suggests that supernumerary chromosomes were horizontally acquired (Ma et al., 2010). In addition, the genomes of some asexual fungal pathogens contain lineage-specific (LS) regions that are highly variable among isolates (Klosterman et al., 2011). An analysis of several nonhomologous recombination forms and polymorphic sequence types of each form in LS regions from different isolates suggests that LS sequences were horizontally acquired (Huang, 2014). These results indicate that horizontal transfer generates genetic variation in asexual fungal pathogens, yet the mechanism and extent of transfer are unknown. It is also unclear whether transfer creates not only presence/absence polymorphisms but also single nucleotide polymorphisms (SNPs).

Fusarium virguliforme is an economically important fungal pathogen that causes sudden death syndrome (SDS) in soybean in North and South America (Aoki et al., 2003). Although different *F. virguliforme* isolates show variation in aggressiveness on soybean plants, studies with various molecular markers detected an extremely low level of genetic variation within *F. virguliforme* isolates from North and South America (O'Donnell et al., 2010; Mbofung et al., 2012). Moreover, mating experiments with 17 US isolates of *F. virguliforme* indicated that they all belonged to a single mating type (Covert et al., 2007). A genome assembly of a *F. virguliforme* isolate was produced recently (Srivastava et al., 2014), and the mating type locus in *F. virguliforme* and its six close relatives were characterized. A PCR assay based on both mating type sequences revealed that all 129 isolates of *F. virguliforme* in North and South America possessed the *MAT1-1* mating type (Hughes et al., 2014). These data suggest that the reproduction mode of *F. virguliforme* on soybean is asexual. It is unknown how genetic variation in this asexual pathogen is generated for the disease to have reached all major soybean-growing areas in the USA since its first detection in the early 1970s.

F. virguliforme is related to the sexual fungal pathogen *Nectria haematococca* MPVI, which is known to contain supernumerary chromosomes (Miao, Covert & VanEtten, 1991). These supernumerary chromosomes contain genes involved in resistance to plant antimicrobial compounds and in host-specific pathogenicity (Covert, 1998). Sequences of the *N. haematococca* MPVI supernumerary chromosomes (Coleman et al., 2009) can be used to determine if their homologs are present in *F. virguliforme*.

F. virguliforme is closely related to five morphologically distinct *Fusarium* species that cause SDS or bean root rot (BRR): *F. azukicola*, *F. brasiliense*, *F. cuneirostrum*, *F. phaseoli* and *F. tucumaniae* (Aoki et al., 2012). *F. tucumaniae* is the only known sexually reproducing fungus among these species (Covert et al., 2007). In this study, we selected ten isolates of these closely related species — three (*F. virguliforme*), three (*F. tucumaniae*), and one (each of the other four species) — for next-generation genome sequencing and analysis in comparison with the *F. virguliforme* genome assembly (Srivastava et al., 2014) as a reference (Table 1). Note that including the reference isolate leads to a total of eleven isolates, four of which are *F. virguliforme* isolates.

We compared the genomes of 11 isolates comprising six *Fusarium* species that cause SDS or BRR, and detected numerous instances of HCT (horizontal chromosome transfer) in supernumerary chromosomes. The genome of the asexual pathogen *F. virguliforme* was composed of a large core genome and a small supernumerary portion; there was little variation in the core between isolates, but there are a statistically significant number (21 standard deviations above the mean) of SNPs in the supernumerary chromosomes between isolates. Supernumerary chromosomes carry reverse transcriptase-related genes (*RVT*); they were highly variable in length, and the presence of long *RVT* open reading frames (ORFs) in the supernumerary chromosome was correlated with the presence of two or more chromosome copies with a significant number of SNPs between them. Our results suggest

PeerJ PrePrints

that supernumerary chromosomes transfer horizontally via an RNA intermediate; their high SNP and length variation rates were attributed to the high error rates of the RVT enzymes. We report an extensive body of evidence to suggest that the *F. virguliforme* genome evolved by two mechanisms: duplication-induced mutation for the core and replication via an RNA intermediate for the supernumerary.

Materials & Methods

Sequence data

We selected 10 isolates of six *Fusarium* species and produced Illumina paired-end reads of 102 bp for each of them. We previously produced a genome assembly (NCBI BioProject Accession: PRJNA63281) of isolate *F. virguliforme* Mont-1 (Srivastava et al., 2014), which was used as a reference genome assembly in this study. The origin, year of collection, and name abbreviation of each of these 11 isolates are in Table 1.

Read mapping and SNP detection

A SNP between the reference isolate and another isolate (query) has two or more alleles called REF and ALT. The REF allele refers to the allele in the reference and ALT alleles refer to alternate non-reference alleles. A SNP is of type 2 if both the REF allele and the ALT allele are present in the query isolate, and of type 1 if only the ALT allele is present in the query isolate.

The sets of Illumina paired-end reads for each query isolate were mapped onto the reference genome assembly with Bowtie2 (Langmead & Salzberg, 2012). The output from Bowtie2 in SAM format was redirected to Samtools (Li et al., 2009) with the view command to produce output in BAM format, which was sorted with the sort command. The sorted output in BAM format was piled up on the reference with the mpileup command. For command options and parameter values, see Huang (2014). The sorted BAM output files for all the isolates along with the reference genome assembly were uploaded into Inte-

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

grative Genomics Viewer (Robinson et al., 2011) for viewing SNPs and presence/absence polymorphisms in each isolate.

Assembly of short reads

An assembly of paired-end reads for each isolate was performed with an Illumina version of PCAP (Huang et al., 2003) with the following data and options: a pair of mate files in fastq format; a minimum insert length of 100 bp and a maximum insert length of 700 bp; an average insert length of 400 bp with a standard deviation of 100. The minimum length of overlaps with no base mismatch match was set to 84 bp, and that of overlaps with up to three base mismatches was set to 90 bp. No overlap with more than three base mismatches was accepted. Each data set was of size up to 49 Gb, and each assembly could be produced in a day on a processor with 100 Gb of main memory. One feature of the program is that it is conservative in joining reads into contigs by avoiding reads in the overlap between two potential contigs that can not be merged into one.

Assembly mapping

Each assembly of Illumina reads was mapped to the reference genome assembly by using BWA-MEM (Li, 2013) with the default options. The output from BWA-MEM in SAM format was redirected to Samtools (Li et al., 2009) with the view command and -bS options to produce output in BAM format, which was sorted with the sort command. An output file of SNPs and indels in VCF format was produced in the same way as in the read mapping. The assembly mapping was useful in finding long indels between contigs in the reference assembly and query assembly, respectively. The coordinates of an indel between two contigs were found by computing an alignment of the contigs with GAP3 (Huang & Chao, 2003).

Gene identification and functional annotation

Ab initio gene identification in a *Fusarium* genomic sequence was performed using Augustus (Stanke & Waack, 2003) with training data from *F. graminearum*. A non-redundant

PeerJ PrePrints

protein sequence database at National Center for Biotechnology Information was searched using Blastx (Gish & States, 1993) with a genomic coding region as a query to find a set of protein database sequences that are most similar to the coding region. The gene structure from Augustus was refined by using AAT (Huang et al., 1997) on the set of protein database sequences. Functional annotation of genes was performed using the HMMER web server (Finn, Clements & Eddy, 2011).

Phylogenetic analysis

A maximum-likelihood tree of the 11 SDS/BRR *Fusarium* isolates was inferred from genome-wide SNP data. The data were produced by mapping reads from each of 10 of the 11 isolates onto a genome assembly of the reference *Fv* Mont-1. A covered SNP position is a position of the reference that was sufficiently covered by reads from each isolate and had an alternative allele (a SNP) in the read coverage of this position from one of the 10 isolates. A total of 297,076 covered SNP positions were aligned in the 11 isolates. The multiple sequence alignment was analyzed to infer the tree with 200 bootstrap samples.

Results

Rapid evolution in a small portion of the F. virguliforme genome

We mapped short reads from each of the ten isolates onto a 50.5-Mb genome assembly of isolate *Fv* Mont-1 (Srivastava et al., 2014) as a reference. The length of the reference covered by reads from the isolate and the distribution of SNPs between the reference and the isolate are given in Table 2. Table 2 reveals significant variation in evolutionary rate among the isolates. First, the four *F. virguliforme* isolates possessed a low genome-wide SNP rate of less than 1 in 10,000 bp, which is consistent with an asexual mode of reproduction. Isolate *Fv* 34551 collected in South America in 2002 was closer to *Fv* Mont-1 collected in the USA in 1991 than the other two *F. virguliforme* isolates collected in the USA. Second, the genome-wide SNP rate of about 1 in 200 bp between the reference and each non-*F. virguliforme* isolate was at least 58 times higher than that between the reference and each *F. virguliforme* isolate, indicating a significantly higher level of polymorphism and suggesting a much longer divergence time between *F. virguliforme* and the other species. Third, the genome-wide SNP rate of 1 in 200 bp between the reference and each non-*F. virguliforme* isolate was not high enough to explain why at least 10 Mb of the reference genome was covered by reads from every *F. virguliforme* isolate, but not by reads from any non-*F. virguliforme* isolate.

To shed light on the last observation, we selected all of the contigs that were at least 1 kb in the reference assembly and calculated the total number of contig bases covered by reads from *Fc* 31157 as well as that not covered by reads from this isolate. The size of the covered portion was 39.5 Mb; that of the uncovered portion was 10.9 Mb. The uncovered portion was A+T rich (68%), whereas the covered portion was A+T poor (45%). The duplicated content of the uncovered portion was 70%, with 48% made up of sequences with copy numbers above 20. In sharp contrast, the duplicated content of the covered portion was 3.8%, with 0.56% made up of 20-plus-copy sequences. These results indicate that since the divergence between *F. virguliforme* and *F. cuneirostrum*, the uncovered portion of the *F. virguliforme* genome evolved much faster in association with duplication and C-to-T/G-to-A mutation. For example, a maximum likelihood tree of 13 duplicated sequences 3,772 bp in length from the genome assembly of *Fv* Mont-1 showed that the more recently duplicated sequences have a higher A+T content (Fig. 1).

Although the genome-wide SNP rate between the reference *F. virguliforme* isolate and each of the other three *F. virguliforme* isolates was at most 0.00007, we found high levels of variation among the four *F. virguliforme* isolates in a small portion ($<= 2\%$) of the genome; the maximum SNP rate between the reference *F. virguliforme* isolate and any other *F. virguliforme* isolate was at least 21 standard deviation units above the mean SNP rate. In addition, the maximum SNP rate for isolate *Fv* LL0009 was even larger than that for each of the non-*F. virguliforme* isolates, three of which belonged to the sexual species *F. tucumaniae*. This suggests that different evolutionary forces may have shaped their genomes.

The maximum SNP rate for each of the top six isolates in Table 2 was all contained

PeerJ PrePrints

in one of the two contigs (the second contig of 52,027 bp and the fourth contig of 68,285 bp) in scaffold 28 of the reference assembly. Scaffold 28 contained 12 contigs with a total length of 217,558 bp that were ordered and oriented by using 454 read pairs with two insert sizes of 3 kb and 20 kb (Srivastava et al., 2014). The two contigs, referred to as mc28.2 and mc28.4 (m for Mont-1 and c for contig), were separated by the third contig (referred to as mc28.3) of 36,918 bp. Scaffold 28 was linked by 14 read pairs (with an insert size of 20,000 bp) downstream to scaffold 66 with three contigs, the largest one of which was contig mc66.3 at 27,852 bp. Many SNPs were also found in mc66.3 in each of the top six isolates in Table 2. Thus, scaffold 28 was expanded to include scaffold 66.

Many of the SNPs in contigs mc28.2 and mc28.4 between the reference and each of the top six isolates were of type 2. In fact, the maximum type 2 SNP rate between the reference and each of the three *F. virguliforme* isolates was at least 0.00117. The high type 2 SNP rates indicated that two or more sequence types were present in each *F. virguliforme* isolate. In addition, high type 2 SNP rates in the small portion of the genome were found in the isolates of *F. cuneirostrum*, *F. phaseoli* and *F. brasiliense*, whereas low type 2 SNP rates in every region of the genome were observed in the isolates of *F. tucumaniae* and *F. azukicola*.

We inferred evolutionary relationships among the 11 isolates by constructing a phylogenetic tree (Fig. 2) based on concatenation of 297,076 SNPs. The tree showed three clearly separate clusters: a first one formed by the four *F. virguliforme* isolates; a second one by *Fb* 31757, *Fc* 31157, and *Fp* 31156; a third one by the three *F. tucumaniae* isolates. The four *F. virguliforme* isolates formed a close cluster with extremely low levels of genome-wide variation among them. On the other hand, high levels of genome-wide variation were observed within the sexually reproducing species *F. tucumaniae*.

The rapidly evolving portion is homologous to a known supernumerary chromosome

Scaffold 28 of the *Fv* Mont-1 genome assembly was compared with the genome assembly of *Nectria haematococca* MPVI, the most closely related species whose genome sequence

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

was determined previously (Coleman et al., 2009). Two unique significant matches (with at least 90% identity over 10,000 bp) were found in chromosome 14 of *N. haematococca* MPVI, a known supernumerary chromosome; one match was in mc28.4 and the other in mc28.10. The sequence of chromosome 14 was compared with the rest of the *Fv* Mont-1 assembly to find additional strong matches. No match meeting the above requirement was found; we found only one additional match (with 95% identity over 5,000 bp) in contig mc71.1. Like mc28.4, mc71.1 was rich in SNPs for some *F. virguliforme* isolates (see below). The unique significant matches between scaffold 28 of *Fv* Mont-1 and chromosome 14 of *N. haematococca* MPVI suggest the possibility that scaffold 28 was supernumerary.

Scaffold 28 was also highly variable among the *F. virguliforme* isolates, with several presence/absence polymorphisms. For example, contig mc28.3 was fully covered by reads from *Fv* 34551 with no SNPs, mostly covered by reads from *Fv* LL0009 with many SNPs, but barely covered by reads from *Fv* Clinton-1B. In addition, mc28.3 was highly variable among *Fc* 31157, *Fp* 31156, and *Fb* 31757. Similarly, contigs mc28.8 of 5 kb, mc28.11 of 8 kb, and mc28.12 of 8 kb were highly variable among the *F. virguliforme* isolates. Thus, scaffold 28 was supernumerary by definition.

The rapid evolution is linked to horizontal transfer

Contigs mc28.2 and mc28.4 were compared with a genome assembly of each isolate to find corresponding contigs in the assembly with unique significant matches (with $\geq 94\%$ identity over ≥ 5 kb). Corresponding contigs were found in each of the top six isolates in Table 2. In addition, mc28.2 and mc28.4 were sufficiently covered by reads from each of these isolates. However, mc28.2 and mc28.4 were barely covered by reads from any of the bottom four isolates in Table 2. In addition, little variation in mc28.2 was detected between the reference isolate and *Fv* 34551. For *Fv* 34551, the major differences in read coverage depth and type 2 SNP number between mc28.2 and mc28.4 indicate the presence of a long segment and a short segment in *Fv* 34551 that were highly polymorphic over mc28.4.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

By contrast, we detected significant variation in mc28.2 and mc28.4 between the reference isolate and *Fv* Clinton-1B by finding unique significant matches in a comparison of these contigs with the *Fv* Clinton-1B genome assembly. Some of the matches suggest a chromosomal rearrangement between the reference isolate and *Fv* Clinton-1B, and the presence of two genomic segments in the reference isolate that were highly similar over some of their lengths but were quite different over the rest (Fig. 3). The sequence integrity of cc26.1 over the breakpoint (marked by a green arrow in Fig. 3) was confirmed by a match of 96% identity between a region of cc26.1 from 28,492 to 52,548 bp and a region of a contig of 27,382 bp from a genome assembly of *Fv* LL0009; the percent identity of the match around the breakpoint was nearly 99%. In addition, by mapping short reads from each isolate onto the *Fv* Clinton-1B genome assembly, we found that cc26.1 was deeply covered over the breakpoint by reads from the five isolates: *Fv* Clinton-1B (at a depth of 414), *Fv* LL0009 (319), *Fc* 31157 (722), *Fp* 31156 (494), and *Fb* 31757 (231). However, cc26.1 was not covered at the breakpoint by any reads from *Fv* 34551, although cc26.1 was deeply covered before and after the breakpoint by these reads. Therefore, the rearrangement type in cc26.1 of *Fv* Clinton-1B was not present in *Fv* 34551; the rearrangement type in mc28.2 and mc28.4 of *Fv* Mont-1 was present only in *Fv* 34551 based on the deep read coverage of mc28.2 around the breakpoint (at a depth of 240) and of mc28.4 around the breakpoint (231). Furthermore, a type 2 SNP G/A (G, REF allele; A, ALT allele) was found near the breakpoint in cc26.1 in *Fv* Clinton-1B (G at a coverage depth of 253; A at 153), *Fc* 31157 (567/179), and *Fp* 31156 (278/224), a sign that two polymorphic segments were present in each of these three isolates.

A total of eight contig sequence alignments showing SNPs and small indels between the reference isolate and *Fv* Clinton-1B are shown in Fig. 4. Each alignment contained two or more instances of polymorphism, all of which were close enough to be linked by 102-bp reads. We checked for the presence/absence of these polymorphic sequences in each of the top six isolates in Table 2. This was done by mapping short reads from each of the six isolates onto the genome assembly of the reference isolate and again onto that of *Fv* Clinton-1B. We found additional types of polymorphic sequences by examining the read coverage of each contig sequence. Thus, some alignments in Figure 4 contained three polymorphic sequences. For each isolate and for each sequence in each alignment, Table 3 shows the number of reads from the isolate that matched and linked all alleles in the sequence.

Table 3 contains unexpected data. Five sequence types in the *Fv* Clinton-1B genome assembly (A3.Tc, A4.Tb, A5.Tc, A7.Tc, and A8.Tb) were covered by reads from *Fv* Clinton-1B, *Fc* 31157, and *Fp* 31156; they were not covered by any read from *Fv* LL0009 and *Fv* 34551 although these two isolates showed little variation from *Fv* Clinton-1B over most of the genome. Similarly, three sequence types in *Fv* Clinton-1B (A1.Tb, A2.Tb and A3.Tb) were covered by reads from *Fv* Clinton-1B and *Fv* LL0009, and from one or more of *Fc* 31157, *Fp* 31156, and *Fb* 31757; they were not covered by any read from *Fv* 34551. In addition, a sequence type in the reference genome assembly (A8.Ta) was covered by reads from *Fv* LL0009, *Fv* 34551, *Fc* 31157, and *Fp* 31156; it was not covered by any read from *Fv* Clinton-1B. A chromosomal rearrangement type in the *Fv* Clinton-1B genome assembly was covered by reads from five of the six isolates but not by any read from *Fv* 34551 (see above). Moreover, Table 3 shows that every isolate except *Fb* 31757 contained two or more polymorphic sequence types, i.e., two or more copies of an element. Analysis of *Fb* 31757 revealed that it contained two alleles at each SNP position in its deep read coverage (≥ 500) of two large regions of mc28.4, a sign that the isolate contained two copies of an element. These observations suggest that copies of the element in scaffold 28 were transferred horizontally.

After discovering the short common sequence types in cc26.1 and cc440.1 between *Fv* Clinton-1B and *Fc* 31157, we checked to see if the two isolates were closer in the whole contigs than the other isolates. Contig cc26.1 was completely covered at a high depth by reads from *Fc* 31157, but only partially at a high depth by reads from each of the other four isolates. Thus, *Fv* Clinton-1B was most similar to *Fc* 31157 in this contig, which is another species, and less similar to *Fv* LL0009 and *Fv* 34551 of its own species. We also made a similar observation regarding contig cc440.1. These observations also suggest that the element (i.e., a chromosome or part of it) in contigs cc26.1 and cc440.1 of *Fv* Clinton-1B was horizontally acquired from another species. The presence of two or more DNA segments homologous to scaffold 28 and with numerous small and large variations

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

in each of the top six SDS/BRR isolates suggests that horizontal transfer was a frequent process in this clade of closely related species.

Additional supernumerary elements

We discovered another reference contig (mc74.1 of 75 kb) in which a high SNP rate between the reference isolate and *Fv* LL0009 was observed; it was 4.8 standard deviation units above the mean SNP rate. Isolate *Fv* 34551 was most similar to the reference isolate in contig mc74.1, as indicated by a low SNP rate between them. Contig mc74.1 was the first of a three-contig scaffold of 80 kb. We found a total of 119 type 2 SNPs in the *Fv* LL0009 read coverage of mc74.1, suggesting that the isolate contained two or more polymorphic copies of the element in mc74.1. Contig mc74.1 (over its separate regions) had unique significant matches (with 99% identity over 10 kb) to three contigs (lc47.1 of 16 kb, lc25.1 of 33 kb, and lc220.1 of 18 kb) in the *Fv* LL0009 genome assembly. Contig lc220.1 was a nearly perfect match over its whole length (except its short ends) to a region of mc74.1. However, contig lc25.1 was only a local match to mc74.1; a region of lc25.1 from positions 9,409 to 27,568 bp was 99% identical to a region of mc74.1 from positions 41,232 to 23,076 bp (in reverse orientation). Moreover, only this region of lc25.1 was covered in high depth by reads from *Fv* Mont-1, *Fv* 34551, *Fc* 31157, and *Fp* 31156.

On the other hand, contig lc25.1 from positions 4,845 to 12,262 bp was 99% identical to contig cc714.1 (from positions 7,419 to 1 bp) of *Fv* Clinton-1B; contig lc25.1 from positions 4,830 to 13,682 bp was 99% identical to contig bc299.1 from positions 8,853 to 1 bp of *Fb* 31757. The two strong matches confirmed the integrity of the region of contig lc25.1 from positions 4,830 to 9,408 bp. In addition, a region of lc25.1 from positions 606 to 4,153 bp was 99% identical to contig bc2776.1 (from positions 2 to 3,537 bp) of *Fb* 31757. This region of lc25.1 was not covered by any read from the other isolates. Contig cc714.1 was another contig in which not all of the six SDS/BRR isolates were the same in their read coverage of this contig. Taken together, these observations suggest that copies of this element were transferred horizontally.

We screened the reference assembly for additional contigs with a high SNP rate or

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

contigs in which some of the isolates were different in their read coverage of these contigs. A total of 18 scaffolds with such contigs were found (Table 4). These scaffolds were candidate supernumerary elements.

Genes in supernumerary elements

We annotated genes in two supernumerary elements by combining ab initio gene structure prediction with protein database matching. A list of proteins along with their functions in each element are shown in Figure 5. We found two types of proteins. One type of proteins was involved in drug metabolism, for example, cytochrome P450 and epoxide hydrolase. The other type was related to cell cycle (e.g., cyclin), cell calcium control (e.g., calcium exchanger), cell wall (e.g. endochitinase), DNA replication (e.g., reverse transcriptase-related enzyme) and repair (e.g., double-strand-break-repair protein). The second type of proteins provide hints regarding the mechanism of horizontal transfer based on the assumption that selection acts on those genes in the element to make its horizontal transfer successful.

We examined variation in some of the genes among the isolates. Contig mc74.1 harbors a gene encoding a cytochrome P450 (CYP) enzyme of 643 residues. This enzyme, a member of family CYP53 (e-value = 1.0e-152), is capable of detoxifying plant defensive compounds, including benzoic acid derivatives (Durairaj et al., 2015). The gene was present in the top seven isolates including *Ft* 31096, but not in the other three isolates including *Ft* 31781 and *Ft* 34546. No SNPs were found in this gene in each of *Fv* Clinton-1B, *Fv* 34551, *Fc* 31157, *Fp* 31156, whose reads covered the reference locus at depths between 70 and 380; 2 SNPs were found in *Fb* 31757. In contrast, we found 11 type 2 SNPs in *Fv* LL0009. Of the 11 SNPs, 8 were nonsynonymous, 1 synonymous, 1 in an intron, and 1 in a $5'$ untranslated region (UTR). In addition, in *Ft* 31096, 12 SNPs were found, of which 8 were nonsynonymous.

The supernumerary *CYP53* gene was 43% identical at the amino acid level to another region (contig mc2.51) in the core genome, where the two genes share the same 4-exon gene structure with two short exons followed by two long exons. The core *CYP53* gene

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

was present in all of the isolates with no SNPs among the *F. virguliforme* isolates and a total of 12 SNPs between the *F. virguliforme* isolates and the non-*F. virguliforme* isolates. Of the 12 SNPs, 3 were present in all the non-*F. virguliforme* isolates, 2 were in all the *F. tucumaniae* isolates, 3 were in *Fc* 31157 and *Fb* 31757, 3 were only in *Fa* 54364, 1 was only in *Fp* 31156. The core CYP53 enzyme was 90% identical to a CYP53 enzyme of 635 residues from *N. haematococca* MPVI; which was also the best match (at 43% identity) for the supernumerary CYP53 enzyme when searched against all of the *N. haematococca* MPVI proteins. These results suggest that the supernumerary *CYP53* gene came from the core genome and was under positive selection.

Similar results were obtained for the following supernumerary genes: a 4-exon gene encoding 517-residue P450 enzyme in contig mc28.4. a single-exon gene coding for a 248-residue G1/S-specific cyclin in contig mc66.3, and a 2-exon gene encoding a 514-residue protein with a heterokaryon incompatibility (HET) domain in contig mc74.1. Details are omitted.

Reverse transcriptase-related enzymes in supernumerary elements

The supernumerary element in scaffold 74 carried both *RVT1* and *RVT2* genes (Fig. 5), which were conserved among the top six SDS/BRR isolates based on read coverage of the reference element. The *RVT1* gene contained 4 predicted introns; the *RVT2* gene had one. The *RVT1* gene was predicted to encode a protein of 1,619 residues with an endonuclease/exonuclease (e-value = 3.2e-18) domain, a reverse transcriptase (3.1e-27), and an RNase H (8.9e-18). The endonuclease/exonuclease domain was in exon 4 encoding 430 residues, and the other two domains were mostly in exon 5 encoding 692 residues, with the two exons separated by an intron of 58 bp. The *RVT2* gene was predicted to encode a protein of 957 residues with an integrase core domain (4.6e-18) and a reverse transcriptase domain (2.9e-88) but without an endonuclease/exonuclease or RNase domain. The two domains were in exon 2 encoding 710 residues. Scaffold 74 had a G+C content of 52%. We searched the rest of the reference genome for strong matches to either RVT protein and found 7 additional *RVT1* regions and 3 additional *RVT2* ones. For each region,

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

we checked whether its scaffold was variable among the isolates, and if so, we checked whether the presence (or absence) of long *RVT* ORFs in the region was correlated with the presence (or absence) of type 2 SNPs in the read coverage of this region by some isolates.

The results of these searches revealed that the region with the largest-scoring match to the RVT1 protein was part of contig mc71.1 from positions 11,784 to 15,975 bp, which was 76% identical to part of the protein from residues 249 to 1619. The alignment identified two introns and long ORFs with no in-frame stop codons. The 19-kb contig was fully covered by reads from each *F. virguliforme* isolate with 23 type 2 SNPs in *Fv* Clinton-1B and 21 type 2 SNPs in *Fv* LL0009. The contig was partially covered by *Fp* 31156, but barely covered by any of the other isolates. This contig was part of scaffold 71 of 85 kb with a G+C content of 51%. In the 31-kb contig mc71.2, 8 type 2 SNPs were found in *Fv* Clinton-1B, and 5 in *Fv* LL0009; in the 17-kb contig mc71.4, 9 or more were found in each of the three *F. virguliforme* isolates. Contigs mc71.2 and mc71.4 were fully covered by the three *F. virguliforme* isolates; mc71.4 was partially by *Fc* 31157 and *Fp* 31156. This region was an instance of long *RVT1* ORFs in a supernumerary element with a significant number of SNPs between copies. Such instances were detected in the *RVT1* regions of scaffolds 28 (contigs mc28.11 and mc28.12), 54 (contig mc54.2), and 88 (contig mc88.6).

A region with a high-scoring match to the RVT1 protein was found in contig mc117.2 of 18 kb. Part of mc117.2 from positions 4,722 to 1,792 bp was 62% identical to part of the protein from residues 610 to 1619 with 9 in-frame stop codons scattered over the whole region. The DNA-protein alignment predicted an intron of 57 bp between residues 927 and 928 in the *RVT1* gene in mc117.2; an intron of 58 bp was also present between the residues in the *RVT1* gene in mc74.1. Contig mc117.2 was part of scaffold 117 of 24 kb with a G+C content of 42%; the G+C content of mc117.2 was 50%. This contig was fully covered by each of the three *F. virguliforme* isolates, mostly covered by *Ft* 31781 and *Ft* 34546, but was not covered by any of the other isolates including *Ft* 31096. Contig mc117.1 of 6 kb was covered only by each of the three *F. virguliforme* isolates. Few SNPs were detected among the three *F. virguliforme* isolates in this scaffold.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

Variation in mc117.2 among the three *F. tucumaniae* isolates indicated that scaffold 117 was supernumerary. The absence of long *RVT1* ORFs in this element was consistent with the absence of SNPs in it.

A region of mc51.6 from positions 26,034 to 23,606 bp was 71% identical to part of the *RVT1* protein from residues 952 to 1,619 with 1 intron and 37 in-frame stop codons. Contig mc51.6 was part of a scaffold of 137 kb with a G+C content of 45%; the G+C content of mc51.6 was 42%. A large part (excluding the *RVT1* gene) of mc51.6 was covered by reads from all the ten isolates; a large part of this scaffold (contig mc51.4 of 88 kb with a G+C content of 50%) was mostly covered by reads from each of the ten isolates. The rest of the scaffold with a low G+C content was covered by only the three *F. virguliforme* isolates. Few SNPs were detected among the *F. virguliforme* isolates in this scaffold. However, in the parts of this scaffold covered by the ten isolates, many SNPs were identified in each non-*F. virguliforme* isolate. These observations indicate that this scaffold was part of the core genome. The large number of in-frame stop codons and the low G+C content revealed that the *RVT1* gene was subjected to G-to-A and C-to-T mutation.

The last of the 7 regions was a 1,986-bp ORF in contig mc15.5 of 408 kb with a G+C content of 53%, which was part of scaffold 15 of 900 kb. The ORF was 78% identical to part of the RVT1 protein from residues 950 to 1,619. The ORF was fully covered by reads from each *F. virguliforme* isolate with a maximum depth of 232, and by reads from isolate *Fa* 54364 with a maximum depth of 3,009, although it was barely covered by the other isolates. The rest of mc15.5, however, was densely covered by reads from each of the ten isolates. Few SNPs were detected in this 408-kb contig among the *F. virguliforme* isolates. These observations indicate that this ORF was part of the core genome.

We examined the 3 regions with a strong match to the RVT2 protein. One of them was 91% identical to the entire protein with an in-frame stop codon shown on the DNA-protein alignment. The region was in contig mc41.8 with a G+C content of 50%, part of scaffold 41 of 206 kb. We noted a large variation in the read coverage of this contig between *Fv* Clinton-1B and the other two *F. virguliforme* isolates; we found 117 SNPs in this contig of 21 kb between *Fv* Clinton-1B and the reference isolate, and 36 (29 type

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

2) SNPs between *Fv* LL0009 and the reference isolate. This region was an instance of long *RVT2* ORFs in a supernumerary element with a significant number of SNPs between copies. Such an instance was also detected in the *RVT2* region of scaffold 50 (contig mc50.3).

The third one was 43% identical to part of the protein from residues 113 to 957 with 21 in-frame stop codons. This region was in contig mc57.6 with a G+C content of 44%, part of scaffold 57 of 114 kb. We observed two instances of presence/absence variation among the three *F. virguliforme* isolates in their read coverage of this contig: (1) Part of the contig from positions 2,301 to 2,376 bp was covered at a minimum depth of 85 by reads from *Fv* Clinton-1B, but was not covered by reads from *Fv* LL0009 or *Fv* 34551. (2) Part of the contig from positions 5,920 to 6,373 was covered at depths between 44 and 230 by reads from *Fv* 34551, but was sparsely covered by reads from *Fv* Clinton-1B or *Fv* LL0009. In addition, we found presence/absence variation among the non-*F. virguliforme* isolates in their read coverage of this scaffold: more than half of mc57.1 was covered by all isolates except *Fa* 54364; mc57.2 was covered by *Ft* 31096, but not by the other two isolates of the same species; mc57.2 was mostly covered by *Fc* 31157, *Fp* 31156 and *Fb* 31757, but mc57.6 was sparsely covered by these three isolates. Few SNPs were detected among the *F. virguliforme* isolates in this scaffold. This region was an instance of short *RVT2* ORFs in a supernumerary element without a significant number of SNPs.

Discussion

Although the four *F. virguliforme* isolates show virtually no variation (at a rate less than 1 in 10,000 bp) in most of the genome, they are highly variable in a small portion of the genome with variation including SNPs and small indels as well as large segment presence/absence polymorphisms. This portion consists of supernumerary chromosomes by definition and by unique strong matches to known supernumerary chromosomes in the related species, *Nectria haematococca* MPVI (Coleman et al., 2009). Some of the supernumerary chromosomes are present in two or more copies with a significant number of SNPs between them. *F. virguliforme* possesses genome-wide variation (at a rate greater

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

than 1 in 230 bp) from three other species: *F. brasiliense*, *F. cuneirostrum*, and *F. phaseoli*. Remarkably, supernumerary chromosome sequence types and rearrangement patterns in some of the *F. virguliforme* isolates are present in an isolate of another species, but not in the other *F. virguliforme* isolates. These observations suggest that some supernumerary chromosomes were acquired by horizontal transfer between these species.

Supernumerary chromosomes carry one or two *RVT* genes, which polymerize DNA via an RNA template. Core chromosomes, which are transmitted vertically from parent to offspring, possess few SNPs between *F. virguliforme* isolates, indicating that the DNA polymerases are very accurate. In contrast, supernumerary chromosomes from the same source have a significant number of SNPs between and within *F. virguliforme* isolates. These observations suggest that supernumerary chromosomes are synthesized by the RNA-dependent DNA polymerase in the RVT enzyme during their horizontal transfer, with their SNPs resulting from the high error rate of the polymerase. This is supported by evidence that supernumerary chromosomes with an *RVT* gene that lacks long ORFs show few SNPs between *F. virguliforme* isolates; these supernumerary chromosomes have lost the ability to generate SNPs through horizontal transfer. Conversely, the presence of long *RVT* ORFs in a supernumerary chromosome is associated with the presence of two or more copies of the chromosome with a significant number of SNPs between them. In addition, supernumerary chromosomes tend to contain more more truncations than core chromosomes, another known error type of the polymerase.

It is reasonable to assume that elements carrying RTs may be able to transfer horizontally from one species to another. Retrotransposons carrying RTs are able to transfer horizontally from one species to another (He et al., 1996). Fungal *RVT* genes were found in the genome of microscopic invertebrate animals *Bdelloid rotifers* (Gladyshev & Arkhipova, 2011). Another source of evidence for the involvement of RTs in horizontal transfer is that retrotransposons are associated with a horizontal LS element transfer between the asexual pathogens *Verticillium dahliae* and *V. albo-atrum* (Huang, 2014). Because the RT in the retrotransposon transcribes RNA into cDNA, the RT-related enzyme in the supernumerary chromosome is expected to use an RNA template too. However, because the product of this reverse transcription is a supernumerary chromosome with introns and

PeerJ PrePrints

intergenic regions, the RNA template should be continuously synthesized from the DNA of the supernumerary chromosome. Instead of using RNA polymerase II, it is synthesized by a single-subunit DNA-dependent RNA polymerase, similar to RNA synthesized from mitochondrial DNA. Note that eukaryotic mitochondria use such an RNA polymerase that is structurally and mechanistically related to that of many viruses; a 1378-residue RNA polymerase of this kind was found in the core genome of *F. virguliforme*. Our analysis suggests that supernumerary chromosomes evolve quickly by replicating via an RNA intermediate with single-subunit RNA polymerases and RTs, whereas core chromosomes evolve slowly by replicating once from one generation to next with more accurate DNA polymerases. HCT is linked to replication via an RNA intermediate.

Previous sequence analysis indicates that supernumerary chromosomes possess a different evolutionary history from the core genome (Covert, 1998). However, we found that supernumerary chromosomes of *F. virguliforme* carry genes (e.g., P450 enzymes and a cyclin protein) that are related to those in the core genome. Thus, parts of the *F. virguliforme* supernumerary genome appear to have been derived from the core genome. Because some unique significant sequence matches were discovered between the supernumerary chromosomes of *F. virguliforme* and *N. haematococca* MPVI, we posit that these supernumerary chromosomes have persisted in this lineage for an extended period of evolutionary time. The unique contribution of supernumerary chromosomes to genetic diversity and adaptability in a community of fungal species may be summarized as follows: they acquire genes from the community via HCT, generate mutations in these genes quickly via an RNA template, and donate genes with beneficial mutations to the community.

Mechanisms exist that generate genetic variation both in sexual reproduction and in asexual reproduction. Homologous recombination is used in sexual reproduction to generate genome-wide genetic variation. In the asexual pathogen *F. virguliforme*, duplication-induced mutation is used to generate variation in the core genome between *F. virguliforme* and its relatives; HCT and replication via an RNA intermediate are used to generate variation including SNPs and presence/absence polymorphisms in the supernumerary genome between *F. virguliforme* isolates. The ability to generate variation in asexual reproduction makes it a viable alternative to sexual reproduction. This viability helps explain how asex-

PeerJ PrePrints

ual reproduction in eukaryotes could first emerge, survive instead of becoming a dead end in evolution, and lead to the development of sexual reproduction in eukaryotes.

The species *F. tucumaniae* is an example of a sexual pathogen that appears to have recently jumped to soybean as a host. The high SNP rate in the core chromosomes among the three *F. tucumaniae* isolates is consistent with the fact that the reproductive mode of *F. tucumaniae* is sexual (Covert et al., 2007). This rate is as high as that between the isolates of two different BRR species, *F. cuneirostrum* and *F. phaseoli*. The absence of *F. virguliforme* supernumerary chromosomes in *F. tucumaniae* suggests that they are not essential. The extremely low SNP rate in the essential chromosomes among the four *F. virguliforme* isolates indicates that the reproductive mode of *F. virguliforme* on soybean is asexual (Covert et al., 2007). Our analysis helps explain how the asexual pathogen *F. virguliforme* is more aggressive on soybean than the sexual pathogen *F. tucumaniae* (Scandiani et al., 2004).

The discovery of mechanisms for generating genetic variation in the asexual pathogen *F. virguliforme* raises questions about our understanding of the forces in molecular evolution. Mutations are thought to be stochastic and often occur randomly across the genome. However, in *F. virguliforme*, mutations mostly occur in the supernumerary genome. Presumably some of these mutations are beneficial as they help the species produce more variants or compete with the plant host in a toxin arms race. Genetic drift is thought to be the chief cause of molecular evolution (Kimura, 1983). However, without these novel mechanisms to generate genetic variation, genetic drift with primitive random mutation would be too slow to produce any beneficial variation for selection on act on in this asexual pathogen. These novel mutational mechanisms have significantly increased the chances of beneficial mutations. In addition, the mutation rates in eukaryotes may be significantly higher than previously thought because the supernumerary chromosomes of the eukaryotic species *F. virguliforme* may replicate via an RNA intermediate.

Identification of the RT-related enzymes in the supernumerary chromosome that is transferred within and between fungal species is expected to have a major impact on biotechnology by introducing a new tool in transgenic applications involving eukaryotes. This tool can be used not only to move genes from one eukaryotic species to another, but

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

also to set the genes in an automatic mode to quickly produce more beneficial mutations on their own. At the same time, the risk associated with this tool needs to be understood.

Conclusions

Supernumerary chromosomes evolved much more rapidly than core chromosomes in *F. virguliforme*. Supernumerary chromosomes were acquired by horizontal transfer between *F. virguliforme* and some of its closely related species. Supernumerary chromosomes may replicate and transfer horizontally like retrotransposons.

Additional Information and Declarations

Competing Interests

The authors declare there are no competing interests.

Author Contributions

Xiaoqiu Huang conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper.

Kerry O'Donnell and Anindya Das contributed reagents/materials/analysis tools, prepared figures and/or tables, reviewed drafts of the paper.

Madan K. Bhattacharyya, Binod B. Sahu, Leonor F. Leandro and Subodh K. Srivastava contributed reagents/materials/analysis tools, reviewed drafts of the paper.

Data Availability

The sequence data from this study have been submitted to the NCBI Sequence Read Archive (SRA) under BioProject PRJNA289542.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

Funding

This work was supported in part by a grant from USDA National Institute of Food and Agriculture.

Acknowledgements

Xiaoqiu Huang would like to dedicate this paper to Webb Miller and Ross Overbeek for helping him start his career. The mention of firm names or trade products does not imply that they are endorsed or recommended by the US Department of Agriculture over other firms or similar products not mentioned. The USDA is an equal opportunity provider and employer.

References

[1] Aoki T, O'Donnell K, Homma Y, Lattanzi AR. 2003. Sudden-death syndrome of soybean is caused by two morphologically and phylogenetically distinct species within the *Fusarium solani* species complex – *F. virguliforme* in North America and *F. tucumaniae* in South America. *Mycologia* **95**:660–684.

[2] Aoki T, Tanaka F, Suga H, Hyakumachi M, Scandiani MM, O'Donnell K. 2012. *Fusarium azukicola* sp. nov., an exotic azuki bean root-rot pathogen in Hokkaido, Japan. *Mycologia* **104**:1068–1084.

[3] Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, Schmutz J, Taga M, White GJ, Zhou S, Schwartz DC, Freitag M, Ma L-J, Danchin EGJ, Henrissat B, Coutinho PM, Nelson DR, Straney D, Napoli CA, Barker BM, Gribskov M, Rep M, Kroken S, Molnr I, Rensing C, Kennell JC, Zamora J, Farman ML, Selker EU, Salamov A, Shapiro H, Pangilinan J, Lindquist E, Lamers C, Grigoriev IV, Geiser DM, Covert SF, Temporini E, VanEtten HD. 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. *PLoS Genetics* **5**:e1000618. doi: 10.1371/journal.pgen.1000618

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

[4] Covert SF. 1998. Supernumerary chromosomes in filamentous fungi. *Current Genetics* **33**:311–319.

[5] Covert SF, Aoki T, O'Donnell K, Starkey D, Holliday A, Geiser DM, Cheung F, Town C, Strom A, Juba J, Scandiani M, Yang XB. 2007. Sexual reproduction in the soybean sudden death syndrome pathogen *Fusarium tucumaniae*. *Fungal Genetics and Biology* **44**:799–807.

[6] Durairaj P, Jung E, Park HH, Kim B-G, Yun H. 2015. Comparative functional characterization of a novel benzoate hydroxylase cytochrome P450 of *Fusarium oxysporum*. *Enzyme and Microbial Technology* **70**:58–65.

[7] Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research* **39**:W29–W37.

[8] Gladyshev EA, Arkhipova IR. 2011. A widespread class of reverse transcriptase-related cellular genes. *Proceedings of the National Academy of Sciences of the United States of America* **108**:20311–20316.

[9] Gish W, States DJ. 1993. Identification of protein coding regions by database similarity search. *Nature Genetics* **3**:266–272.

[10] He C, Nourse JP, Kelemu S, Irwin JAG, Manners JM. 1996. *Cg* T1: a non-LTR retrotransposon with restricted distribution in the fungal phytopathogen *Colletotrichum gloeosporioides*. *Molecular and General Genetics* **252**:320-331.

[11] He C, Rusu AG, Poplawski AM, Irwin JA, Manners JM. 1998. Transfer of a supernumerary chromosome between vegetatively incompatible biotypes of the fungus *Colletotrichum gloeosporioides*. *Genetics* **150**:1459–1466.

[12] Huang X. 2014. Horizontal transfer generates genetic variation in an asexual pathogen. *PeerJ* **2**:e650. https: //dx.doi.org/10.7717/peerj.650

[13] Huang X, Adams MD, Zhou H, Kerlavage AR. 1997. A tool for analyzing and annotating genomic sequences. *Genomics* **46**:37–45.

PeerJ PrePrints

[14] Huang X, Chao K-M. 2003. A Generalized global alignment algorithm. *Bioinformatics* **19**:228–233.

[15] Huang X, Wang J, Aluru S, Yang SP, Hillier L. 2003. PCAP: a whole-genome assembly program. *Genome Research* **13**:2164–2170.

[16] Hughes TJ, O'Donnell K, Sink S, Rooney AP, Scandiani MM, Luque A, Bhattacharyya MK, Huang X. 2014. Genetic architecture and evolution of the mating type locus in fusaria that cause soybean sudden death syndrome and bean root rot. *Mycologia* **106**:686–697.

[17] Kimura M. 1983. The neutral theory of molecular evolution. Cambridge: Cambridge University Press.

[18] Kistler HC, Miao VPW. 1992. New modes of genetic change in filamentous fungi. *Annual Review of Phytopathology* **30**:131–152.

[19] Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BPHJ, Zehua Chen, Henrissat B, Lee Y-H, Park J, Garcia-Pedrajas MD, Barbara DJ, Anchieta A, de Jonge R, Santhanam P, Maruthachalam K, Atallah Z, Amyotte SG, Paz Z, Inderbitzin P, Hayes RJ, Heiman DI, Young S, Zeng Q, Engels R, Galagan J., Cuomo CA, Dobinson KF, Ma L-J. 2011. Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathogens* **7**:e1002137. doi: 10.1371/journal.ppat.1002137

[20] Langmead B, Salzberg S. 2012. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**:357–359.

[21] Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: 1303.3997v2 [q-bio.GN].

[22] Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map (SAM) format and SAMtools. *Bioinformatics* **25**:2078-2079.

PeerJ PrePrints

PeerJ PrePrints

[23] Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim W-B, Woloshuk C, Xie X, Xu J-R, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RAE, Chapman S, Coulson R, Coutinho PM, Danchin EGJ, Diener A, Gale LR, Gardiner DM, Goff S, Hammond-Kosack KE, Hilburn K, Hua-Van A, Jonkers W, Kazan K, Kodira CD, Koehrsen M, Kumar L, Lee Y-H, Li L, Manners JM, Miranda-Saavedra D, Mukherjee M, Park G, Park J, Park S-Y, Proctor RH, Regev A, Ruiz-Roldan MC, Sain D, Sakthikumar S, Sykes S, Schwartz DC, Turgeon BG, Wapinski I, Yoder O, Young S, Zeng Q, Zhou S, Galagan J, Cuomo CA, Kistler HC, Rep M. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* **464**:367–373.

[24] Masel AM, He C, Poplawski AM, Irwin JAG, Manners JM. 1996. Molecular evidence for chromosome transfer between biotypes of *Colletotrichum gloeosporioides*. *Molecular Plant-Microbe Interactions* **9**:339–348.

[25] Masel AM, Irwin JAG, Manners JM. 1993. DNA addition or deletion is associated with a major karyotype polymorphism in the fungal phytopathogen *Colletotrichum gloeosporioides*. *Molecular and General Genetics* **237**:73–80.

[26] Mbofung GYC, Harrington TC, Steimel JT, Navi SS, Yang XB, Leandro LF. 2012. Genetic structure and variation in aggressiveness in *Fusarium virguliforme* in the Midwest United States. *Canadian Journal of Plant Pathology* **34**:83–97.

[27] Miao VP, Covert SF, VanEtten HD. 1991. A fungal gene for antibiotic resistance on a dispensable (B) chromosome. *Science* **254**:1773–1776.

[28] Michielse CB, Rep M. 2009. Pathogen profile update: *Fusarium oxysporum*. *Molecular Plant Pathology* **10**:311–324.

[29] O'Donnell K, Sink SL, Scandiani MM, Luque A, Colletto A, Biasoli M, Lenzi L, Salas G, González V, Ploper LD, Formento N, Pioli RN, Aoki T, Yang XB, Sarver BAJ. 2010. Soybean sudden death syndrome species diversity within North and South America revealed by multilocus genotyping. *Phytopathology* **100**:58–71.

[30] Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP. 2011. Integrative genomics viewer. *Nature Biotechnology* **29**:24–26.

[31] Scandiani M, Ruberti D, O'Donnell K, Aoki T, Pioli R, Giorda L, Luque A, Biasoli M. 2004. Recent outbreak of soybean sudden death syndrome caused by *Fusarium virguliforme* and *F. tucumaniae* in Argentina. *Disease Notes* **88**:1044.

[32] Srivastava SK, Huang X, Brar HK, Fakhoury AM, Bluhm BH, Bhattacharyya MK. 2014. The genome sequence of the fungal pathogen *Fusarium virguliforme* that causes sudden death syndrome in soybean. *PLoS ONE* **9**:e81832. doi: 10.1371/journal.pone.0081832

[33] Stanke M, Waack S. 2003. Gene prediction with a hidden-markov model and a new intron submodel. *Bioinformatics* **19**:ii215-ii225.

PeerJ PrePrints

Table 1: Isolates used in this study.

Isolate[a]	Origin	Year	Abbreviation
F. virguliforme Mont-1	USA, Illinois	1991	*Fv* Mont-1
F. virguliforme Clinton-1B	USA, Iowa	1993	*Fv* Clinton-1B
F. virguliforme LL0009	USA, Iowa	2006	*Fv* LL0009
F. virguliforme NRRL 34551	Argentina, Buenos Aires	2002	*Fv* 34551
F. cuneirostrum NRRL 31157	USA, Michigan	1992	*Fc* 31157
F. phaseoli NRRL 31156	USA, Michigan	Unknown	*Fp* 31156
F. brasiliense NRRL 31757	Brazil, Distrito Federal	1992	*Fb* 31757
F. tucumaniae NRRL 31096	Argentina, Tucumán	2001	*Ft* 31096
F. tucumaniae NRRL 31781	Argentina, Tucumán	Unknown	*Ft* 31781
F. tucumaniae NRRL 34546	Argentina, Buenos Aires	2000	*Ft* 34546
F. azukicola NRRL 54364	Japan, Hokkaido	1997	*Fa* 54364

[a] NRRL= Agricultural Research Service Culture Collection, National Center for Agricultural Utilization Research, USDA-ARS, Peoria, IL. No NRRL number is known for some isolates.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

Table 2: Length of coverage and distribution of SNPs when reads were mapped onto reference *Fv* Mont-1.

Isolate	Length of coverage (Mb)	Number of SNPs	Mean SNP rate/ standard deviation[a]	Max SNP rate[b]
Fv 34551	49.9	4,955	0.00003/0.00007	0.00177 (23.7)
Fv Clinton-1B	49.6	8,269	0.00006/0.00044	0.00960 (21.5)
Fv LL0009	49.2	8,541	0.00007/0.00052	0.01129 (21.7)
Fc 31157	39.5	176,065	0.00446/0.00123	0.01126 (5.5)
Fp 31156	40.0	178,511	0.00447/0.00125	0.01097 (5.2)
Fb 31757	39.3	172,100	0.00435/0.00117	0.00903 (4.0)
Ft 31096	39.3	181,420	0.00462/0.00128	0.00943 (3.8)
Ft 31781	39.2	172,823	0.00441/0.00114	0.00829 (3.4)
Ft 34546	38.9	157,076	0.00412/0.00102	0.00726 (3.1)
Fa 54364	37.9	188,209	0.00506/0.00119	0.00957 (3.8)

[a] The mapped reference was partitioned into at least 700 disjoint windows each with 35-kb sufficiently covered base positions. The mean and standard deviation were calculated for the SNP rates of these windows.

[b] The number in the parentheses is the maximum SNP rate measured in units of standard deviation above the mean.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

Table 3: The number of reads from the isolate that link all alleles in the sequence.

Sequence[a]	Number of reads from the isolate that cover the sequence[b]					
	Fv Clinton-1B	*Fv* LL0009	*Fv* 34551	*Fc* 31157	*Fp* 31156	*Fb* 31757
A1.Ta	16	16	16	0	0	0
A1.Tb	32	18	0	48	46	8
A2.Ta	88	149	147	0	0	9
A2.Tb	85	114	0	115	0	0
A3.Ta	52	41	61	0	0	0
A3.Tb	78	52	0	152	0	0
A3.Tc	33	0	0	34	46	0
A4.Ta	162	134	77	121	0	97
A4.Tb	54	0	0	127	242	0
A5.Ta	39	27	57	0	0	18
A5.Tb	0	8	0	39	0	0
A5.Tc	85	0	0	69	65	0
A6.Ta	0	0	46	0	0	0
A6.Tb	72	0	0	0	209	0
A6.Tc	116	121	35	554	244	74
A7.Ta	0	0	98	0	0	0
A7.Tb	0	31	0	0	54	0
A7.Tc	35	0	0	42	50	39
A8.Ta	0	35	34	40	42	0
A8.Tb	70	0	0	51	59	42

[a] Each sequence is denoted by its alignment number and type letter (Fig. 4): e.g., Types a and b in Alignment 1 are denoted by A1.Ta and A1.Tb, respectively.

[b] A read covers a sequence in a set of polymorphic sequences if the read and the sequence have the same allele at every occurrence of polymorphism.

Table 4: Scaffolds with supernumerary elements.

Scaffold	Length (kb)	Contig with type 2 SNPs or coverage variation (CV)[a]
26	379	mc26.1 (CV: *Fv* LL0009, *Fv* 34551)
28	218	mc28.2 (SNPs: *Fv* Clinton-1B)
33	330	mc33.8 (CV: *Fv* Clinton-1B, *Fv* LL0009)
41	207	mc41.8 (SNPs: *Fv* LL0009)
46	158	mc46.2 (CV: *Fv* LL0009, *Fv* 34551)
50	140	mc50.4 (SNPs: *Fv* Clinton-1B)
58	96	mc58.2 (CV: *Fv* Clinton-1B, *Fv* LL0009)
71	85	mc71.2 (SNPs: *Fv* Clinton-1B)
74	80	mc74.1 (SNPs: *Fv* LL0009)
79	73	mc79.6 (SNPs: *Fv* LL0009)
80	69	mc80.1 (CV: *Fv* LL0009, *Fv* 34551)
88	51	mc88.6 (SNPs: *Fp* 31156)
90	61	mc90.7 (CV: *Fv* 34551, *Fc* 31157)
91	44	mc91.1 (SNPs: *Fp* 31156)
98	45	mc98.3 (CV: *Fv* Clinton-1B, *Fv* LL0009)
100	37	mc100.2 (CV: *Fv* Clinton-1B, *Fv* LL0009)
117	24	mc117.2 (CV: *Ft* 31096, *Ft* 31781)
158	12	mc158.2 (CV: *Fc* 31157, *Fp* 31156)

[a] Shown in the parentheses are the names of two isolates in which read coverage variation was detected in the contig or the name of an isolate in which type 2 SNPs were detected in the contig.

PeerJ PrePrints

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

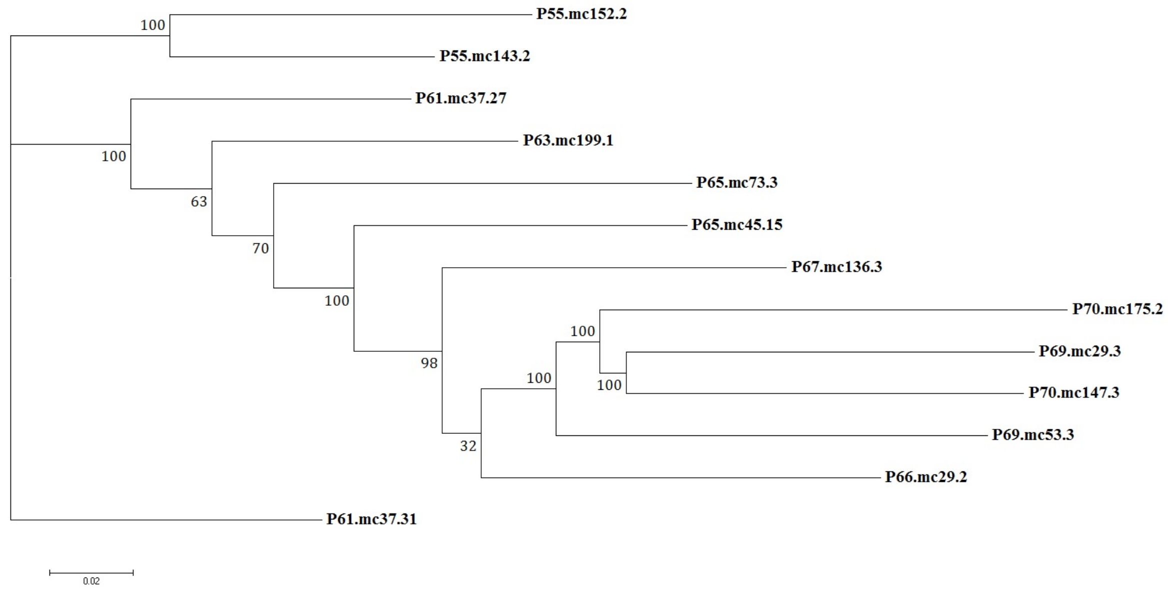

Figure 1: Maximum likelihood tree of 13 duplicated sequences in the *Fv* Mont-1 assembly. Each sequence was named based on its A+T content followed by its contig name. For example, sequence P61.mc37.31 indicates an A+T content of 61% and mc37.31 as its source contig. Support values from 100 bootstrap replicates are provided at internodes.

PeerJ PrePrints

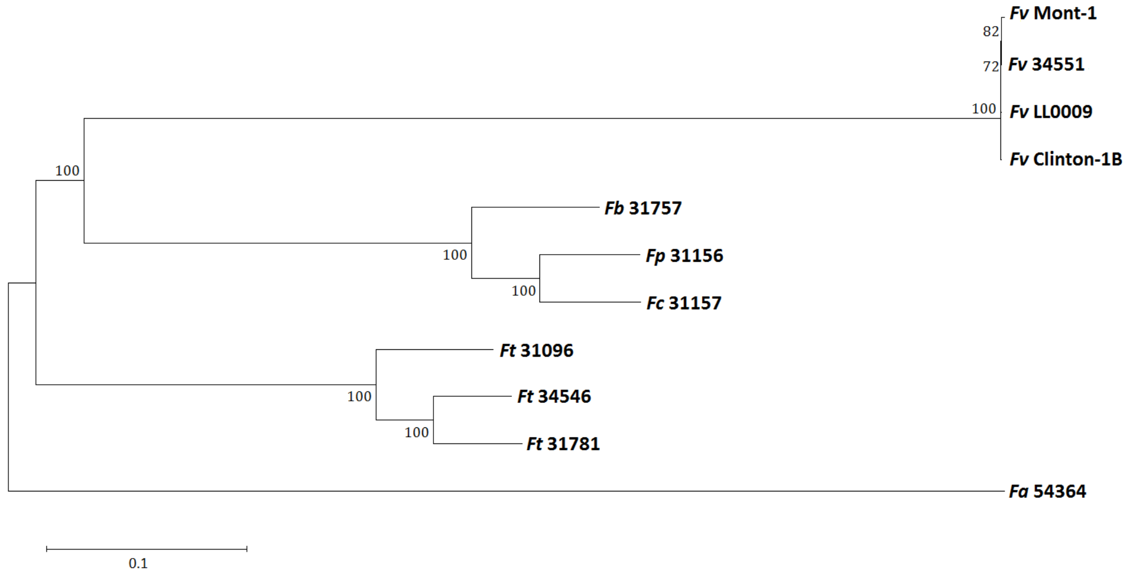

Figure 2: Maximum-likelihood midpoint rooted tree of 11 SDS/BRR *Fusarium* isolates, inferred from genome-wide SNP data with 200 bootstrap samples.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015

PeerJ PrePrints

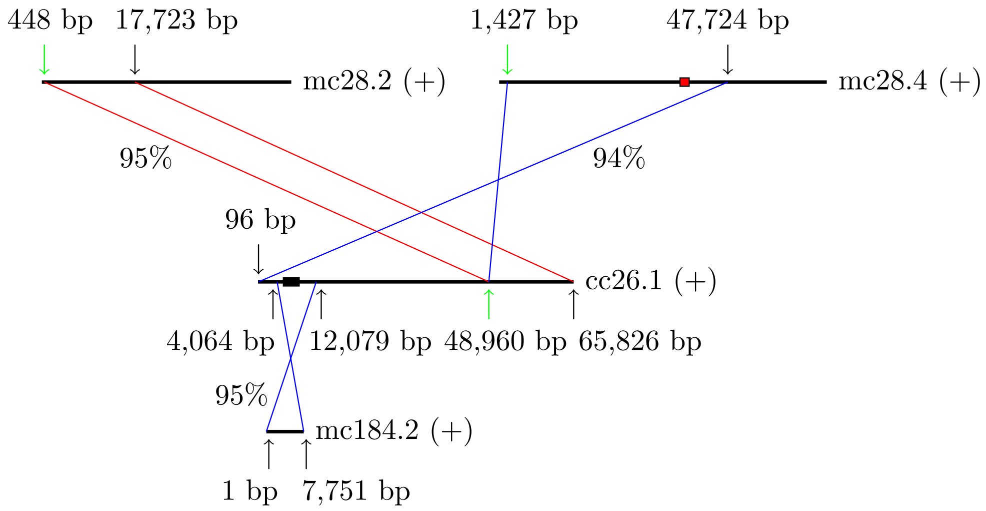

Figure 3: Chromosomal rearrangement between *Fv* Mont-1 and *Fv* Clinton-1B. Each horizontal line represents a contig with its name and orientation (+ denotes forward) given on the right. A unique significant match between contig regions in opposite orientations is indicated by a pair of cross lines; one in the same orientation by a pair of parallel lines. In each case, the percent identity of the match is shown next to the lines. The beginning and end of each contig region in the match are marked with vertical arrows along with their positions in bp. A red box in contig mc28.4 and a black box in contig cc26.1 represent different islands surrounded by the match; the black box is part of the match with contig mc184.2.

PeerJ PrePrints

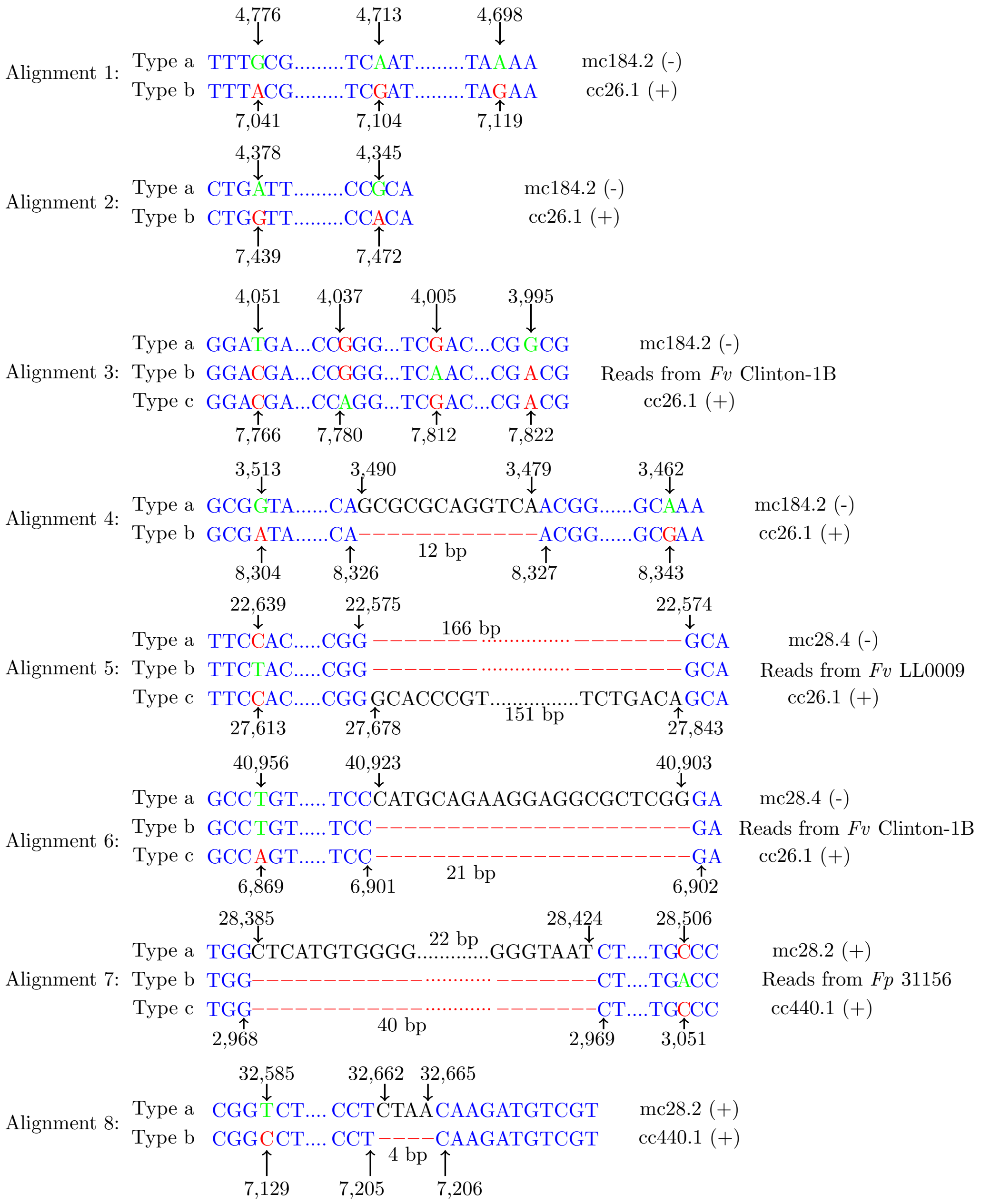
Alignment 1:
4,776 4,713 4,698
Type a TTTGCG.........TCAAT.........TAAAA mc184.2 (-)
Type b TTTACG.........TCGAT.........TAGAA cc26.1 (+)
7,041 7,104 7,119
Alignment 2:
4,378 4,345
Type a CTGATT.........CCGCA mc184.2 (-)
Type b CTGGTT.........CCACA cc26.1 (+)
7,439 7,472
Alignment 3:
4,051 4,037 4,005 3,995
Type a GGATGA...CCGGG...TCGAC...CG GCG mc184.2 (-)
Type b GGACGA...CCGGG...TCAAC...CG ACG Reads from Fv Clinton-1B
Type c GGACGA...CCAGG...TCGAC...CG ACG cc26.1 (+)
7,766 7,780 7,812 7,822
Alignment 4:
3,513 3,490 3,479 3,462
Type a GCGGTA......CAGCGCGCAGGTCAACGG......GCAAA mc184.2 (-)
Type b GCGATA......CA–––––––––––––ACGG......GCGAA cc26.1 (+)
12 bp
8,304 8,326 8,327 8,343
Alignment 5:
22,639 22,575 22,574
Type a TTCCAC.....CGG 166 bp GCA mc28.4 (-)
Type b TTCTAC.....CGG GCA Reads from Fv LL0009
Type c TTCCAC.....CGG GCACCCGT...............TCTGACAGCA cc26.1 (+)
151 bp
27,613 27,678 27,843
Alignment 6:
40,956 40,923 40,903
Type a GCCTGT.....TCCCATGCAGAAGGAGGCGCTCGGGA mc28.4 (-)
Type b GCCTGT.....TCC–––––––––––––––––––––GA Reads from Fv Clinton-1B
Type c GCCAGT.....TCC–––––––––––––––––––––GA cc26.1 (+)
21 bp
6,869 6,901 6,902
Alignment 7:
28,385 28,424 28,506
Type a TGGCTCATGTGGGG.............GGGTAAT CT....TGCCC mc28.2 (+)
22 bp
Type b TGG–––––––––––––––– CT....TGACC Reads from Fp 31156
Type c TGG–––––––––––––––– CT....TGCCC cc440.1 (+)
40 bp
2,968 2,969 3,051
Alignment 8:
32,585 32,662 32,665
Type a CGGTCT.... CCTCTAACAAGATGTCGT mc28.2 (+)
Type b CGGCCT.... CCT––––CAAGATGTCGT cc440.1 (+)
4 bp
7,129 7,205 7,206

Figure 4: Eight sequence alignments with SNPs and small indels (4 to 166 bp). Each alignment is composed of two or three sequence types (denoted by Types a, b and c): a reference contig, a contig in the *Fv* Clinton-1B assembly, and sometimes short reads from one of the ten isolates, which were mapped to one of the two contigs. The name of each contig along with its orientation (+ denotes forward and - denotes reverse), or the name of the isolate if present, is shown to the right of its sequence type. Every allele in the contig is marked with an arrow and a number in bp showing its position. Notation: mc184.2, *Fv* Mont-1 contig 184.2; cc26.1, *Fv* Clinton-1B contig 26.1.

PeerJ PrePrints

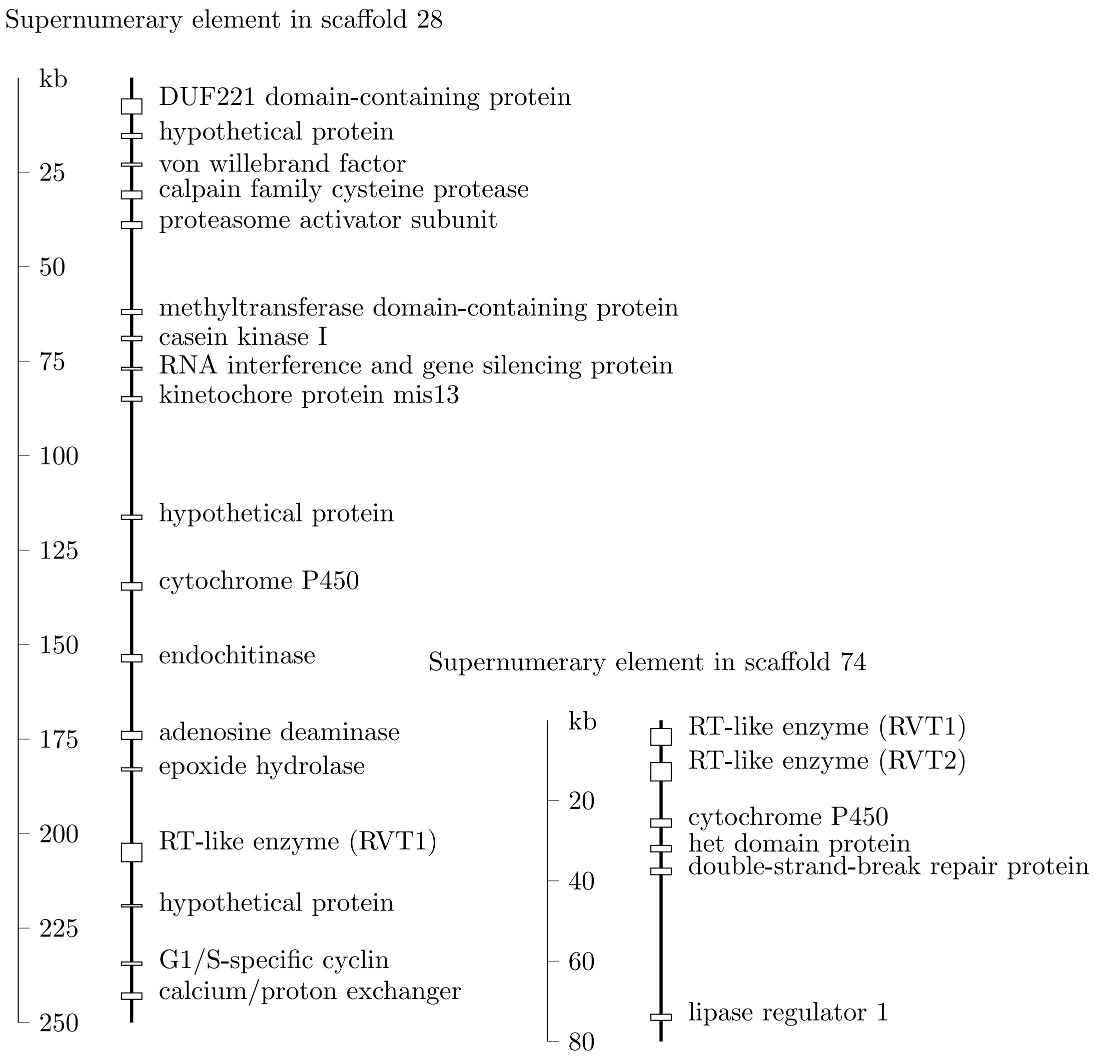

Figure 5: Proteins encoded by two supernumerary elements. The related proteins between the elements are P450 enzymes and reverse transcriptase-like (RT-like) enzymes. The larger element encodes a G1/S-specific cyclin protein.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1324v1 | CC-BY 4.0 Open Access | rec: 25 Aug 2015, publ: 25 Aug 2015