

RESEARCH

Open Access



# Insights into genomic evolution from the chromosomal and mitochondrial genomes of *Ustilagoidea virens*

Kang Zhang<sup>1,2\*</sup>, Zaixu Zhao<sup>1</sup>, Ziding Zhang<sup>3</sup>, Yuejiao Li<sup>1</sup>, Shaojie Li<sup>4</sup>, Nan Yao<sup>5</sup>, Tom Hsiang<sup>6</sup> and Wenxian Sun<sup>1,7\*</sup> 

## Abstract

*Ustilagoidea virens*, the causal agent of rice false smut, is an economically important filamentous fungal pathogen. A high-quality reference genome of *U. virens* promotes understanding of molecular mechanisms underlying its virulence and pathogenicity. Here, we report the first chromosome-level assembly of *U. virens* genome consisting of seven chromosomes ranging from 2.4 to 7.5 Mb. The assembly has dramatic improvements over previous assemblies, including considerably longer contigs, higher proportion of repetitive elements and more functionally annotated genes. Phylogenetic analyses revealed an extremely low intraspecific sequence divergence in *U. virens*. By contrast, intraspecific genome comparisons uncovered dynamic genomic alterations including massive structural variations and widespread lineage-specific regions (LSRs) among *U. virens* strains, which were mainly generated by recent burst of retrotransposons. Genomic plasticity created by structural variations and LSRs might drive rapid evolution of *U. virens*. High-quality mitochondrial genomes of eight *U. virens* strains exhibit size variations from 94 to 102 kb. Consistently, *U. virens* contains conserved lengths of exons and highly dynamic mobile introns, which contribute to intraspecific size variations due to gain/loss of homing endonuclease genes. This study highlights unique characteristics in nuclear and mitochondrial genomic divergence and provides new insights into genomic and mitochondrial evolution of *U. virens*.

**Keywords:** *Ustilagoidea virens*, Comparative genomics, Structural variations, Genomic evolution

## Background

Rice false smut (RFS), caused by the Sordariomycete *Ustilagoidea virens*, is one of the most devastating rice grain diseases (Tanaka et al. 2008; Sun et al. 2020). It was first reported in India in the 1870s, and was once considered as a minor disease (Cooke 1878). However, the disease has expanded to the majority of rice-production regions worldwide in the past few decades and has led to huge economic losses (Guo et al. 2012). Increased incidence might be associated with overuse of chemical fertilizers

and large-scale cultivation of high-yield disease-susceptible rice varieties (Guo et al. 2012). *U. virens* infects rice floral organs and then produces false smut balls in spikelets (Ikegami 1963; Fan et al. 2016). The pathogen generates a large amount of mycotoxins including ustiloxins and ustilaginoidins in RFS balls. Ustiloxins inhibit microtubule assembly and cytoskeleton formation in plant and animal cells (Koiso et al. 1994; Li et al. 1995), and hence pose health threats to humans and livestock. Therefore, the disease not only causes significant yield losses, but also severely reduces grain quality.

The first published genome of *U. virens* has been assembled using second generation Illumina short-read data (Zhang et al. 2014). The genome assembly has a size of 39.40 Mb with a high percentage of repeat

\* Correspondence: [wxs@cau.edu.cn](mailto:wxs@cau.edu.cn); [zhangkang01@caas.cn](mailto:zhangkang01@caas.cn)

<sup>1</sup>Department of Plant Pathology, The Ministry of Agriculture Key Laboratory of Pest Monitoring and Green Management, Joint Laboratory for International Cooperation in Crop Molecular Breeding, Ministry of Education, China Agricultural University, Beijing 100193, China  
Full list of author information is available at the end of the article



© The Author(s). 2021 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

sequences (~25%). Comparative genomics revealed that *U. virens* is evolutionarily close to *Metarhizium* spp., which are generally insect pathogens. Intraspecific genomic comparisons revealed that single nucleotide polymorphisms (SNPs) are extremely low (<0.1%) among the five strains collected from different regions in China. The draft genomes facilitate a better understanding of the molecular mechanisms of *U. virens* pathogenicity. For example, expression profiling during infection and sexual development of *U. virens* gave clues to genes contributing to pathogenicity and sexual conidiation, respectively (Han et al. 2015; Yu et al. 2016a). Single sequence repeat (SSR) markers have been developed and multiple pathogenicity genes have been identified with the assistance of genome information (Yu et al. 2016b; Qiu et al. 2019). A genome-wide protein interaction network has been constructed to boost functional genomic analyses in *U. virens* (Zhang et al. 2017). These studies have significantly advanced understanding of *U. virens* pathogenicity.

However, the first and recently-released *U. virens* genome assemblies are highly fragmented because of short sequencing reads or low-coverage PacBio data (Zhang et al. 2014; Kumagai et al. 2016). In particular, repeat sequences are difficult to be fully recovered in such assemblies, because identical repeat sequences might be collapsed during assembly (Alkan et al. 2011). These drawbacks exclude in-depth comparative genomic analysis for such assemblies at a genome level. In some important fungal pathogens including *Verticillium dahliae* (Faino et al. 2015), *Magnaporthe oryzae* (Bao et al. 2017), *Botrytis cinerea* (van Kan et al. 2017) and *Alternaria solani* (Wolters et al. 2018), high-quality genomes have been assembled using long sequencing reads obtained by third-generation sequencing technologies such as PacBio and Nanopore. Complete or near complete genome assemblies with accurate transposable element (TE) annotations greatly expedite comparative genomics in these fungi. Multiple studies on fungal genomes revealed that large-scale genomic polymorphisms such as structural variations and dispensable chromosomes are frequently involved in pathogenicity of fungal pathogens, and TEs play important roles in generating such genomic variations (Plissonneau et al. 2016; Möller and Stukenbrock 2017; Bertazzoni et al. 2018; Tsushima et al. 2019). Many filamentous pathogens have a two-speed genome architecture, highly conserved versus variable compartments (Raffaële and Kamoun 2012; Dong et al. 2015). The variable genomic compartments, usually TE-enriched, are considered as hotspots for structural variations, which greatly affect birth and death of effector genes and even shape the evolution of fungal genomes (Dong et al. 2015; Sánchez-Vallet et al. 2018). In *V. dahliae*, extensive chromosomal reshuffling and lineage-

specific regions (LSRs) are speculated to be created by TEs, which might drive genome evolution through introducing erroneous repair of double-strand breaks (de Jonge et al. 2013; Faino et al. 2016). In *M. oryzae*, Bao et al. (2017) revealed that virulence-associated secreted protein-encoding genes are enriched in LSRs, indicating a correlation between the variable regions and pathogenicity. In addition, the contributions of repeat-rich dispensable chromosomes to fungal pathogenicity have been well demonstrated in *Mycosphaerella graminicola* (Goodwin et al. 2011) and *Fusarium oxysporum* (Ma et al. 2010). Therefore, a chromosome-level genome assembly of *U. virens* is essential for illustrating structural variations and biological significance of TE in the genome evolution of this pathogen.

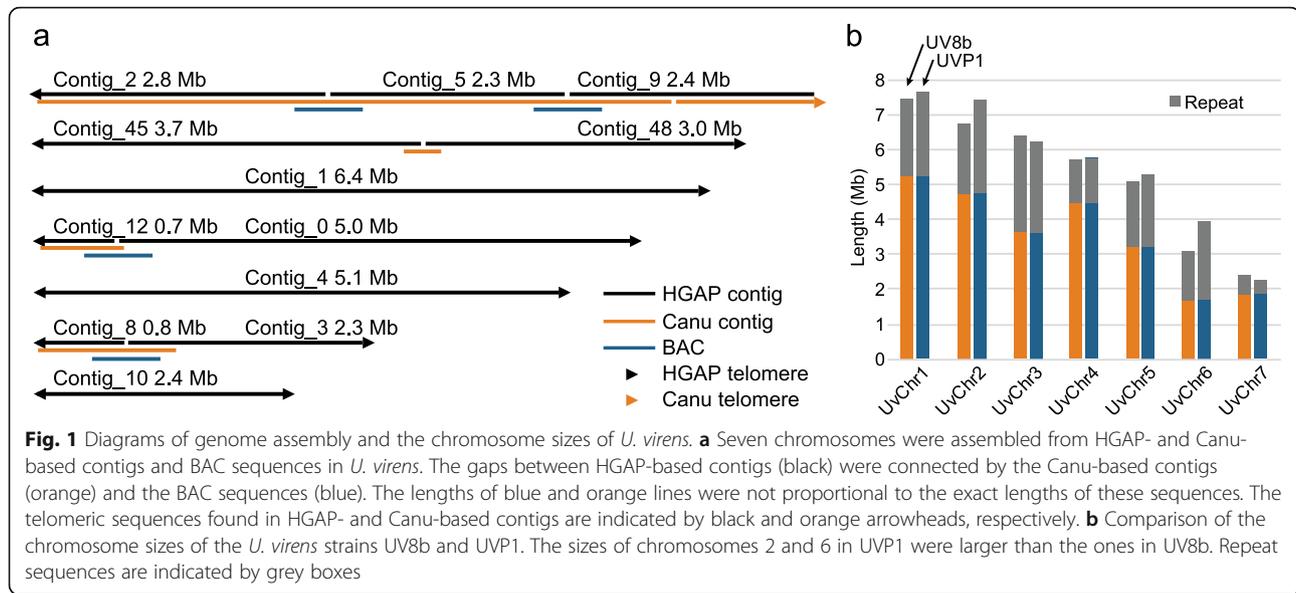
In this study, high-quality genomes were generated for the *U. virens* reference strain UV8b and a laboratory strain UVP1. Gene and TE annotations were updated based on the newly assembled UV8b genome. Comparative genomics among UV8b, UVP1 and eight other genome-sequenced strains illustrate the existence of massive structural variations and LSRs in *U. virens*, which are probably created by recently-inserted TEs. Furthermore, we assembled the mitochondrial genomes of eight *U. virens* strains, and uncovered extensive intra-specific sequence divergence in mitochondrial genomes of *U. virens*.

## Results

### A near-complete genome of *U. virens*

To obtain high-quality assembly of *U. virens* genome, single-molecule real-time (SMRT) sequencing data (~2.59 Gb) with a subread N50 length of 10.37 kb were generated for the strain UV8b. Using a hierarchical genome-assembly process (HGAP) (Chin et al. 2013), we constructed a draft assembly yielding a total size of 37.42 Mb, including 22 contigs with a contig N50 length of ~3.75 Mb. The longest 12 contigs covered more than 99% of the assembled genome, suggesting that the contigs could be further assembled into the chromosomes.

The telomere of fungal chromosomes is characterized by TTAGGG repetitive sequences (Qi et al. 2013). By scanning sequences in the assembly, 13 TTAGGG repetitive sequences were found and all at the ends of contigs. Three contigs having the telomeric repetitive sequences at both ends most likely represented entire chromosomes. To ascertain the telomeres in the assembled contigs, the Canu-based assembly was further constructed. Based on Canu- and HGAP-based assemblies, 14 telomeric sequences were recovered at the ends of 12 contigs (Fig. 1a). According to the overlaps between HGAP and Canu contigs, these 12 contigs were assembled into seven putative continuous chromosomes. Furthermore, the assembly was validated by multiple BAC



end sequences (BESs) (Wang et al. 2013). As shown in Fig. 1a, four out of five gaps in the chromosomes were bridged by more than ten BESs, further supporting the reliability of the assembly. Therefore, seven chromosomes with the TTAGGG repetitive sequences at both ends were constructed and numbered according to the sequence length. The total size of the assembled UV8b genome was 37.06 Mb, and the chromosomes ranged from 2.41 Mb to 7.47 Mb in length (Fig. 1b and Additional file 1: Table S1). The GC contents of these chromosomes varied from 47.81 to 52.54% (Additional file 1: Table S1). The chromosomes 3, 5 and 6 had low GC contents less than 50%, indicating that these chromosomes contain more repetitive sequences than the others (Additional file 1: Table S1). Quality assessment by BUSCO (Simão et al. 2015) showed that 98.62% of the BUSCO genes were predicted to exist in *U. virens* genome, indicating a high-quality genome assembly.

For intraspecific genome comparisons, another *U. virens* strain UVP1 was sequenced following the same procedure, and the genome was de novo assembled to chromosome-level with the total size of 38.70 Mb. The lengths of seven chromosomes in UVP1 were similar to the counterparts in UV8b (Fig. 1b), except for the chromosomes 2 and 6 which were significantly longer than those of UV8b (Additional file 1: Table S1).

### The high-quality genome assembly improves gene annotations

Based on the updated genome assembly, predicted TEs accounted for 32.63% of genomic sequences, which is much more than those in previous annotations (Zhang et al. 2014). The majority of TEs were categorized as long terminal repeat (LTR) retrotransposons. In

particular, TEs were enriched in the chromosomes 3, 5 and 6 (Additional file 1: Table S1), accounting for the low GC content in these chromosomes. A total of 282 repetitive DNA families were identified with a combined annotation method. Copia represented the most common family, followed by Gypsy (Table 1).

A total of 8297 protein-coding genes were predicted in the updated genome. Although the predicted genes were slightly reduced compared with the previous version, more functional genes including those encoding the PHI-base, CAZyme and secreted proteins were annotated (Table 1). In the new assembly, more protein domains were identified with Pfam and the percentage of single-

**Table 1** Basic characteristics of genome assembly of *Ustilagoideae virens*

Type	This version	Former version
Genome size	37.06	39.40
Contig N50	3,751,346	29,282
TE percentage	32.63%	25.25%
-Percentage of Copia LTR Family	19.43%	-
-Percentage of Gypsy LTR Family	8.18%	-
Gene number	8297	8426
# Proteins with domains	5697	5463
# Proteins with GO annotation	5832	5370
# PHI-base proteins	2159	2158 <sup>a</sup>
# CAZy proteins	230	213
# Secreted proteins	662	628
# Conventional effectors	256	193
# Non-conventional effectors	165	-

"-" indicates that the data is unavailable

<sup>a</sup>PHI-base proteins were re-predicted using the PHI database v4

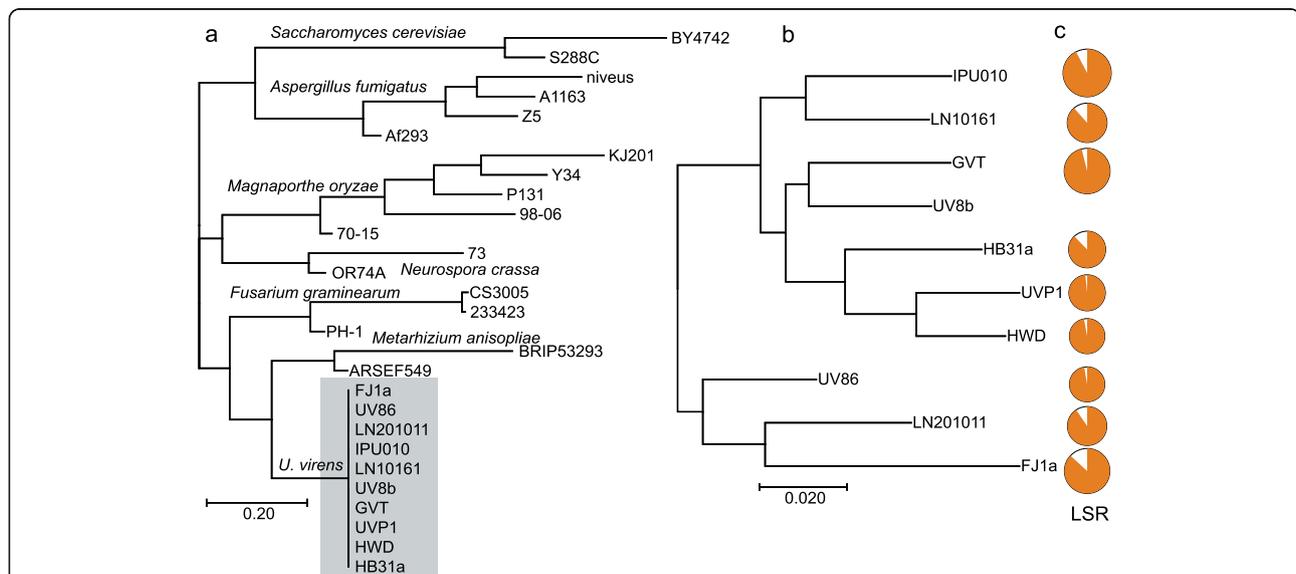
exon genes was reduced from 31.09 to 20.69%, indicating an improved genome assembly and gene prediction.

To identify more putative effectors in *U. virens*, the program EffectorP was used for the effector prediction based on a machine learning strategy (Sperschneider et al. 2016). A total of 256 genes were identified to encode putative effector proteins (Additional file 1: Table S2). It is notable that these predicted effector genes were located closer to adjacent TEs than other genes (Additional file 1: Table S3), suggesting that the higher genetic diversity of the effector genes might be partially generated from transposition events of neighboring TEs. In addition to conventional effector proteins carrying signal peptides, some effector proteins in fungi secreted by non-conventional pathways do not have evident signal peptides (Nombela et al. 2006). Using ApoplastP (Sperschneider et al. 2018), we uncovered 165 putative non-conventional effectors, which are short and Cys-rich peptides and are possibly located in apoplastic spaces of host cells (Additional file 1: Table S2). Notably, these non-conventional effector candidates share significantly lower similarities with their homologs in closely-related fungi including *Metarhizium anisopliae* and *Fusarium graminearum* than other non-effector proteins, which is a typical characteristic of effector proteins (Additional file 1: Table S4).

### The phylogenetic tree indicates a recent expansion of *U. virens*

To better illustrate the evolutionary characteristics of *U. virens* genome, a phylogenetic tree was constructed based on the sequences of single-copy homologous proteins in ten *U. virens* isolates from Japan, India and China, and also in multiple strains from six fungal species closely related with *U. virens* (Fig. 2a). The tree showed that phylogenetic distances among different *U. virens* strains were significantly shorter than those among distinct strains in other fungal species, indicating that the intraspecific protein sequence divergence in *U. virens* is much smaller than that in other investigated fungi (Fig. 2a). Since these *U. virens* strains have different geographic origins (Additional file 1: Table S5), the finding also suggests that *U. virens* might have recently spread over different regions worldwide. This hypothesis is consistent with the fact that rice false smut has emerged as one of the major rice diseases in recent decades, while it was previously considered a minor disease for centuries.

An intraspecific phylogenetic tree based on shared gene sequences of different *U. virens* strains demonstrated that these strains were clustered into two clades (Fig. 2b). The strains UV86, FJ1a and LN201011 were aggregated in one clade, while the remaining strains including UV8b and UVP1 were in the other. Intriguingly,



**Fig. 2** The phylogenetic relationship among various strains in *U. virens* and six related fungal species. **a** The phylogenetic distances among various *U. virens* strains (shaded) are significantly shorter than those among the strains in other fungi as revealed by the phylogenetic tree. The single-copy homologous protein sequences of different strains in the indicated fungal species were aligned using ClustalW, and the tree was constructed using the Neighbor-Joining method in MEGA X with 1000 bootstraps. **b** The intraspecific phylogenetic tree for multiple *U. virens* strains based on the shared gene sequences in these strains. The full-length gene sequences of all these strains were subjected to multiple alignments. The tree was constructed using the Neighbor-Joining method in MEGA X with 1000 bootstraps. **c** The relative size of lineage-specific regions (LSRs) and the percentage of repetitive elements accounting for LSRs in different *U. virens* strains. The relative size of LSRs is represented by the area of the circle, while the percentage of repetitive elements in LSRs is indicated by orange in circles

the strains collected from neighboring regions were not generally clustered together (Fig. 2b and Additional file 1: Table S5). For example, UV8b collected in Hubei Province of central China was more closely related to an Indian strain, GVT, than other strains such as UV86 (collected in Sichuan Province, southwest China). Similarly, IPU010 and LN10161 clustered together despite being collected in Japan and Liaoning Province of China, respectively. Besides, LN10161 and LN201011 were both collected in Liaoning Province, while they belonged to different clades according to this phylogenetic tree. These results did not reveal a significant relationship between the phylogenetic distances and geographical origins of these *U. virens* strains.

Although SNPs among different *U. virens* strains were extraordinarily low, the intronic sequences had a higher SNP ratio relative to conserved coding sequences (CDSs) (Additional file 2: Figure S1). The CDSs also exhibited much lower sequence polymorphisms compared with the intergenic sequences in *U. virens*. This phenomenon was consistently observed in different *U. virens* strains (Additional file 2: Figure S1). The difference in mutation frequency between coding and non-coding sequences might be a consequence of natural selection pressure. Notably, the effector genes had significantly more SNPs in CDSs than other genes, indicating that the effector genes might have evolved more rapidly under selection pressure from host plants than other genes (Additional file 1: Table S6).

### Comparison of two high-quality genomes reveals recently inserted LTR-retrotransposons

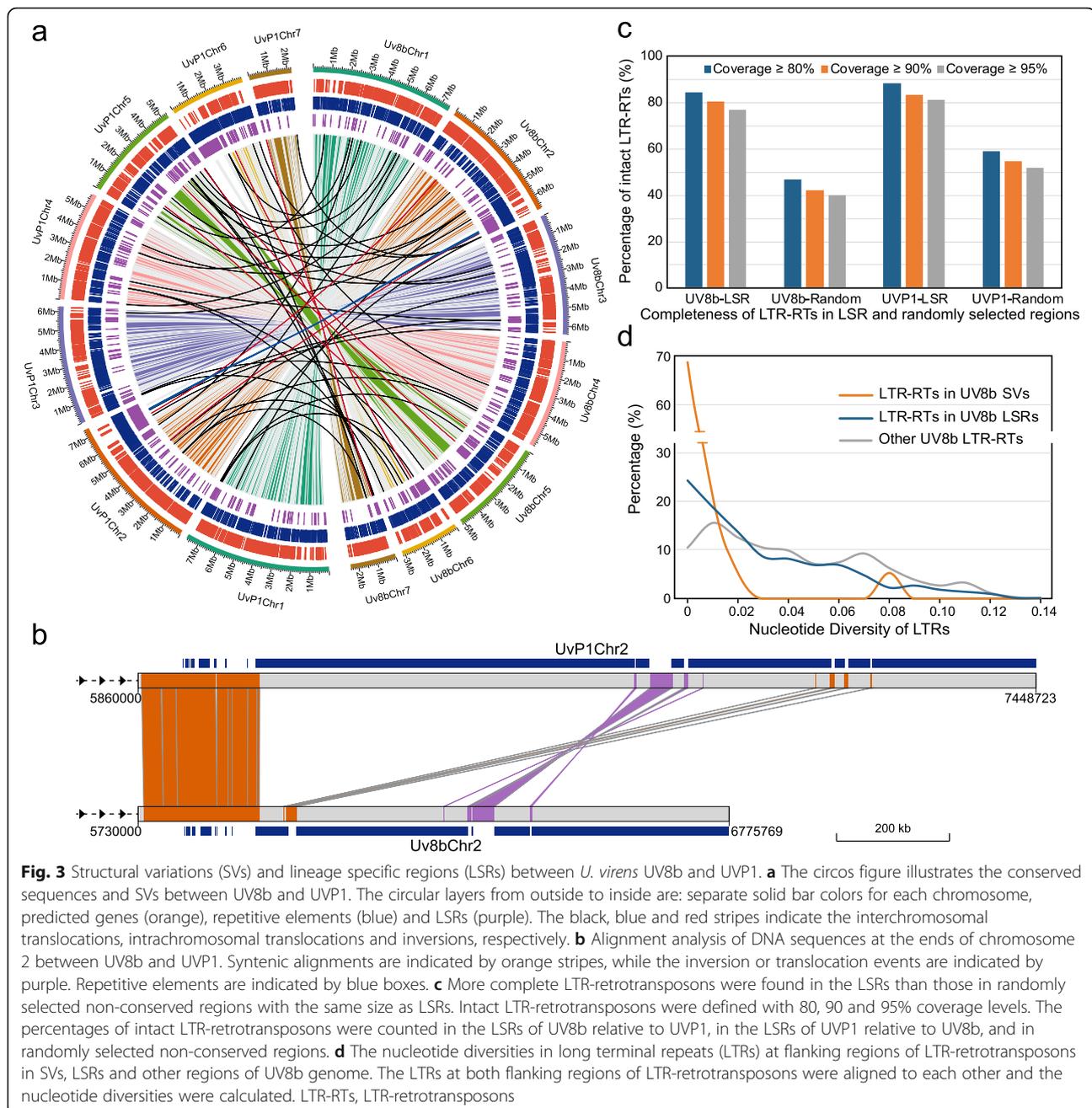
To identify intraspecific structural variations (SVs) in *U. virens*, the genomes of UV8b and UVP1 were compared (Fig. 3a). Even though the two genomes showed high conservation of synteny, 119 SVs including intra/inter-chromosome translocations and inversions were found to have varying sizes from 204 bp to 37.90 kb (Additional file 1: Table S7). Multiple inversion events occurred in the compartment near the telomere of chromosome 2, which was enriched with repeat elements and contained very few syntenic regions (Fig. 3b). Similarly, translocation events occurred between chromosome 7 of UV8b and chromosome 6 of UVP1 (Additional file 2: Figure S2). Despite the ultra-low genetic diversity in coding sequences between UV8b and UVP1, large-scale translocation and inversion events indicate that *U. virens* undergoes dynamic genomic alterations. We further found that almost all SV-involving sequences contained TEs (Additional file 1: Table S7). More interestingly, 87 out of 119 SVs overlapped exactly with one TE unit or coincided with one boundary of TEs (Additional file 1: Table S7), suggesting that genomic alterations are mainly caused by TE insertion.

Furthermore, a total of 4.13 Mb and 5.83 Mb of genome sequences were identified as LSRs in UV8b and UVP1 relative to each other, respectively. The majority of LSRs (98.34% in UV8b and 97.81% in UVP1) were repetitive elements, that is, 34.13 and 42.56% of repetitive elements in *U. virens* genome were categorized as UV8b- and UVP1-specific compared with each other, respectively. Furthermore, LSRs were enriched with intact LTR-retrotransposons. Approximately 39.76 and 48.24% of intact LTR-retrotransposons were located in LSRs of UV8b and UVP1, respectively. This enrichment was further verified by the fact that the percentage of intact LTR-retrotransposons in LSRs was much greater than that in the randomly selected non-conserved regions with the same sizes of LSRs (Fig. 3c). These results suggest that LSRs are predominantly generated by active TE translocation and duplication in *U. virens*.

A typical LTR-retrotransposon contains LTR elements at both flanking regions, which share identical sequences when this LTR-retrotransposon has just inserted into a new location. Therefore, the insertion time of LTR-retrotransposons could be estimated using the mutation rates between the flanking LTR sequences. We found that the curve for nucleotide diversity between the flanking LTRs peaked at zero, suggesting that most LTR-retrotransposons have recently inserted into *U. virens* genome (Fig. 3d). Intriguingly, recently inserted LTR-retrotransposons were enriched in the SVs and LSRs, since the percentages of LTR-retrotransposons with identical LTRs in the SVs and LSRs of UV8b and UVP1 were much higher than those in other genomic regions (Fig. 3d, see Additional file 2: Figure S3 for UVP1). These results indicate that genomic structural variations in *U. virens* genome are mainly driven by the recent propagation of TE, especially the LTR-retrotransposons.

### Massive intraspecies divergence is universal among *U. virens* strains

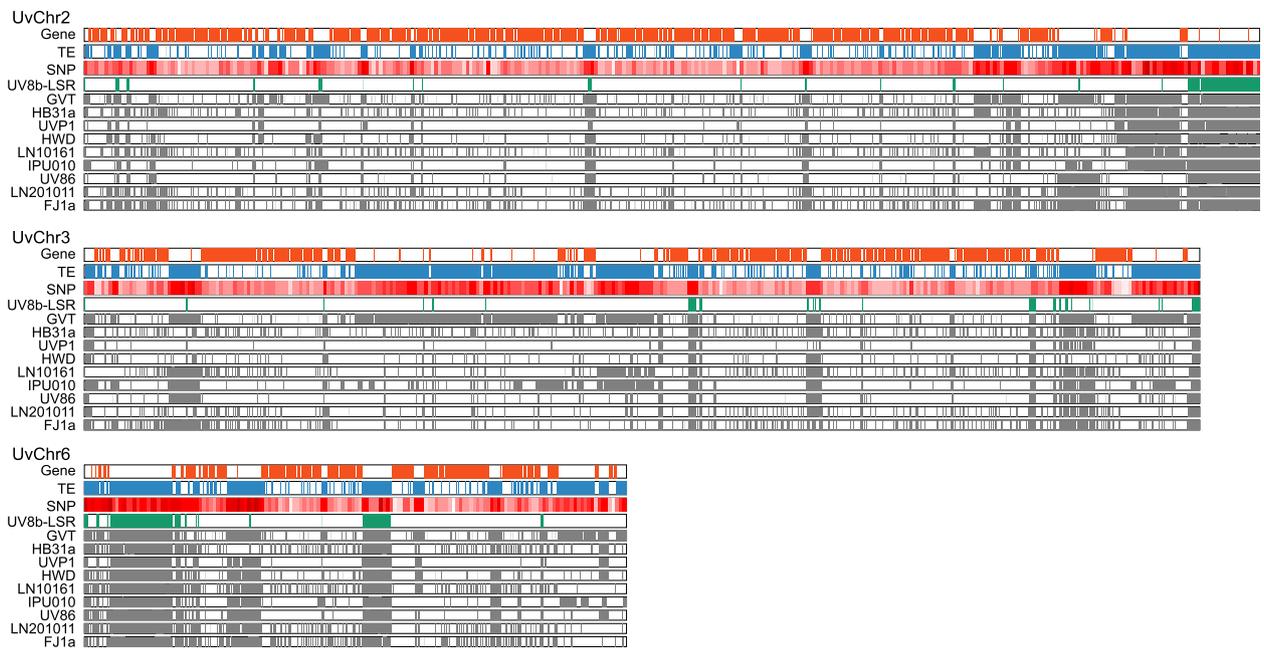
To investigate whether massive divergence between UV8b and UVP1 is a universal phenomenon in *U. virens* or is just a specific case, the LSRs of UV8b relative to FJ1a, HB31a, LN10161, LN201011, UV86 and HWD were identified and ranged from 4.09 Mb to 6.80 Mb based on the Illumina sequencing reads (Fig. 4 and Additional file 1: Table S8) (Zhang et al. 2014). Consistently, the percentages of TEs in these LSRs varied from 86.88 to 98.34%, indicating that LSR formation is largely dependent on TEs. For IPU010 and GVT without publicly available sequencing reads, we simulated the Illumina reads with 100× coverage according to their released genome assembly. The LSRs in UV8b relative to IPU010 and GVT were predicted to be 7.62 Mb and 6.90 Mb, and the repeat sequences accounted for 92.33 and 96.13% of these LSRs, respectively (Fig. 4 and



Additional file 1: Table S8). Furthermore, no significant correlation between LSR sizes and phylogenetic distances was established (Fig. 2c), implying that these LSRs emerge after strain differentiation rather than originating from their common ancestor.

UV8b-specific LSRs among all investigated strains were subsequently identified. The resultant LSRs was approximately 1.25 Mb in total, 98.31% of which was occupied by TE (Fig. 4 and Additional file 2: Figure S4). The

majority of the UV8b-specific LSRs were aggregated as two biggest clusters lying in the ends of chromosome 2 and 6, where TEs were highly enriched. Although almost all LSRs were found to be located in TE-enriched compartments, the long TE-enriched central region of chromosome 3 was highly conserved among these investigated strains except for GVT (Fig. 4). These findings indicate that the TE-enriched regions are not always hotspots for LSR.

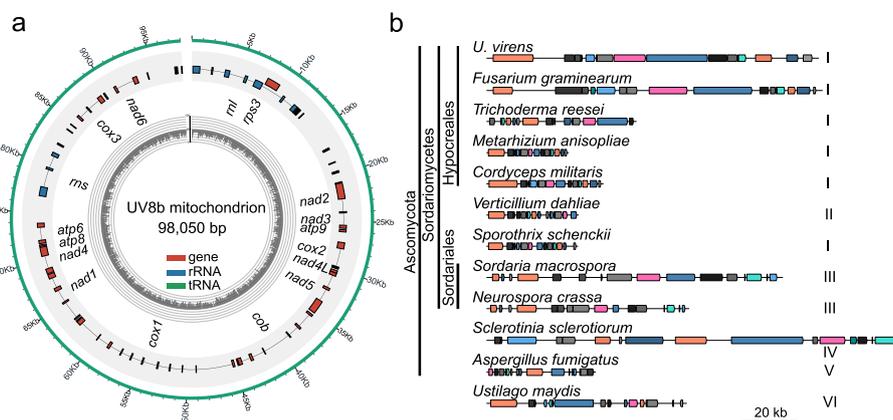


**Fig. 4** The distribution of the UV8b-specific lineage-specific regions (LSRs) relative to multiple *U. viresns* strains. The lines from top to bottom indicate the genes (orange), transposable elements (TEs) (blue), heatmap of SNPs compared with multiple strains (red), UV8b-specific LSRs that are absent in all other strains (green), and the LSRs relative to individual indicated strains (grey). LSRs relative to a specific strain were identified by aligning the sequencing reads of this strain to UV8b genome. The chromosomes 2, 3 and 6 were selected as examples, and other chromosomes were displayed in Additional file 2: Figure S4. UV8b-LSR, UV8b-specific LSR

**The large *U. viresns* mitochondrial genome is caused by the invasion of homing endonuclease genes**

Based on PacBio sequencing data, the complete mitochondrial genome of *U. viresns* UV8b containing 98,050 bp was successfully assembled with the GC content of 27.39%, which is much less than that of chromosomal

DNA. The mitochondrial genome was predicted to encode two ribosomal RNAs (*rns* and *rnl*) and 14 putative proteins involving in ATP production and oxidative phosphorylation. Besides, a ribosome-related gene *rps3* was predicted to lie in the intron region of *rnl* (Fig. 5a). The *U. viresns* mitochondrial gene repertoire is consistent



**Fig. 5** The characteristics, gene annotations and arrangement of mitochondrial genome in *U. viresns*. **a** The annotated genes (red), rRNAs (blue) and tRNAs (green) are shown in a circos diagram. The GC content is shown as an inner ring circular histogram. **b** The conservation of mitochondrial gene order and orientation among multiple fungal species. Mitochondrial genomes were drawn according to the exact lengths of DNA sequences. Bar length, 20 kb. The same colors indicate homologous genes in different species. The taxonomy of the investigated species is indicated in lines to the left, and Roman numerals on the right indicate the groups in which the fungi were categorized according to gene order

with that of most ascomycete fungi, indicating that these genes are essential in maintaining biological functions of mitochondrion.

Furthermore, comparisons of mitochondrial genomes among different fungal species showed that the mitochondrial genome of *U. virens* was larger than that of most fungi (Fig. 5b and Additional file 1: Table S9). Although the total lengths of exons (17–21 kb) did not vary significantly among these fungal species, size variations were largely attributable to different lengths of intronic (1–63 kb) and intergenic sequences (3–45 kb). In *U. virens* mitochondrial DNA, the total length of intronic and intergenic regions is ~ 54 kb and ~ 24 kb, respectively. To validate phylogenetic relationships among these investigated fungi, a phylogenetic tree was constructed based on mitochondrial protein sequences. The phylogenetic relationships revealed by the tree were generally consistent with the ones based on the chromosomal genes (Additional file 2: Figure S5). In addition, gene arrangement and orientation in mitochondrial genomes of the analyzed Hypocreales species were highly conserved (Fig. 5b). Intriguingly, *M. anisopliae*, an evolutionarily closely-related species of *U. virens*, contained no intron in the mitochondrial genes, implying that size variations might be caused by recent intron insertions or deletions.

Given the intron-rich nature of *U. virens* mitochondrial genes, the gene structures were further investigated. Most mitochondrial genes except *atp9*, *nad4L* and *atp8* contained 1–10 introns with an average length of 1,523 bp. A total of 37 putative open reading frames (ORFs) were predicted in the intronic regions. Interestingly, the majority (35/37) of these ORFs might be putative homing endonuclease genes (HEGs) including 27 LAGLID ADG type HEGs and eight GIY-YIG type HEGs (Additional file 1: Table S10). Among them, 11 were considered as pseudogenes due to lack of start/stop codons or premature terminations. This phenomenon suggests that accumulation of mutations in these HEGs leads to pseudogene degeneration after they occupied most mitochondrial intron regions in a specific *U. virens* population.

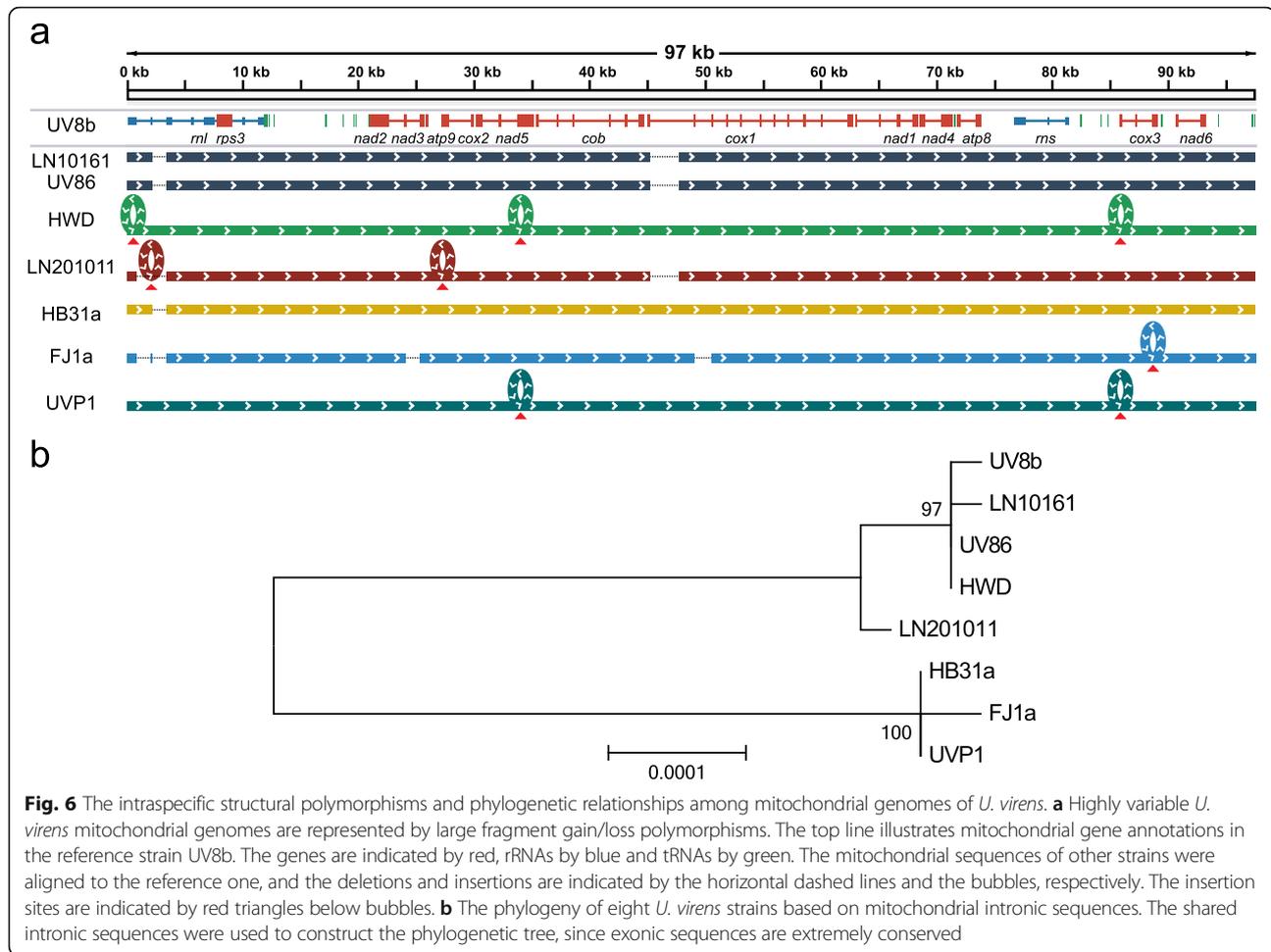
HEGs are known as highly invasive elements resembling TE (Roberts et al. 2003; Burt and Koufopanou 2004). During the investigation of the origination of HEGs in *U. virens*, we found that the majority of HEGs (21/35) in *U. virens* mitochondrial genome shared the highest protein sequence identities with the ones from the species in the same order (Hypocreales), while ten other HEGs shared the highest identities with those from other orders in the class Sordariomycetes (Additional file 1: Table S10). The remaining four HEGs were most closely related to ones in other classes in Ascomycota (Additional file 1: Table S10). For example, two

HEGs locating in the first and fifth introns of *cox1* shared the highest similarities with HEGs in *Sclerotinia sclerotiorum* in class Leotiomyces. These results reveal various evolutionary origins of HEGs in mitochondrial genome of *U. virens*.

#### Highly variable mitochondrial genomes among *U. virens* strains

To explore intraspecific sequence divergence in *U. virens* mitochondrial DNA, mitochondrial genomes were assembled for seven other strains of this species. These genomes ranged from ~ 94 kb to ~ 102 kb, showing high levels of length polymorphism (Fig. 6a and Additional file 1: Table S11). As compared with UV8b, these strains carried at least one large fragment gain/loss at intronic regions in mitochondrial genomes. The gain/loss sequence polymorphisms, mostly in *rnl*, *cox1* and *cox3*, resulted in substantial diversity in composition patterns of mitochondrial DNA. In total, these strains contained eight additional mitochondrial DNA fragments as compared with UV8b, and all of those fragments were annotated as putative HEGs. The large DNA fragments in UV8b mitochondrion that were missing from other strains also contained HEG-like ORFs. These results indicate that the size diversity in *U. virens* mitochondrial genomes is largely caused by the invasion or loss of HEGs.

Among these strains, LN10161 and UV86 shared an identical HEG distribution pattern in mitochondrial genomes, whereas the patterns in other strains were distinct due to different combinations of sequence gain/loss (Fig. 6a). Some strains shared several identical variations at specific sites. For instance, the second intron in *rnl* gene was missing in LN10161, UV86, LN201011, HB31a and FJ1a, and the partial deletion in the first intron of *cox1* occurred in LN10161, UV86 and LN201011. Therefore, we investigated whether the gain/loss patterns of HEGs in these strains correlate with the intraspecific phylogeny. Considering that mitochondrial exons were extremely conserved with only a few SNPs between strains, it was unlikely to distinguish these *U. virens* strains using the exonic sequences. Therefore, an intraspecific phylogenetic tree was constructed based on shared mitochondrial intronic sequences in these *U. virens* strains. As illustrated in Fig. 6b, eight strains were divided into two clades. Intriguingly, the strains in the same clade did not show a significant higher similarity in HEG distribution patterns than those in different clades. Eight HEG-associated polymorphisms were found in clade II strains (HB31a, FJ1a and UVP1), whereas only one polymorphism was shared by the three strains. Besides, UVP1 exhibited a highly similar HEG distribution pattern with HWD in the other clade. The two strains contained two exclusive HEG insertion



events that did not exist in other strains (Fig. 6a). These results suggest that HEG invasion and deletion might be independent from mitochondrial evolution in different *U. virens* strains.

Subsequently, the intraspecific phylogenetic trees constructed using chromosomal and mitochondrial genomes were compared (Figs. 2b and 6b). Although there were two clades in both trees, the phylogenetic relationships of these strains indicated by the trees were significantly different. In the mitochondrial tree, the strain FJ1a was clustered with HB31a and UVP1 to form clade II, whereas FJ1a was clustered with UV86 and LN201011 according to the chromosomal tree. The substantial difference indicates that mitochondrial evolution, which might be influenced by the interchange or by the recombination of mitochondrial DNA, is independent from chromosomal evolution in *U. virens*.

## Discussion

A high-quality *U. virens* genome is a prerequisite not only for better understanding of virulence and

pathogenicity but also for interspecific and intraspecific comparative genomics. In this study, we report chromosome-level genome assemblies and reveal nuclear and mitochondrial genome evolution characteristics in *U. virens*.

### A high-quality repeat-rich fungal genome assembly using a non-hybrid method

In previous studies, hybrid assembly strategies to integrate high-depth Illumina short-read data and low-coverage PacBio long-read data were preferentially used to balance quality and cost (Faino and Thomma 2014; Kumar et al. 2017). Hence, we attempted to assemble *U. virens* genome with both Illumina and PacBio sequencing data using different hybrid methods including SPAdes (Bankevich et al. 2012) and DBG2OLC (Ye et al. 2016). Intriguingly, we found that in our case the non-hybrid assembling method generated a larger contig N50 size and greater completeness than hybrid ones, even with relatively low-coverage PacBio sequencing data (Additional file 1: Table S12). In addition, the HGAP-based assembly for *U. virens* genome showed a better

performance in sequence continuity than the Canu-based one when PacBio sequencing data were increasing (Additional file 1: Table S12), indicating that HGAP method might be preferable to Canu when more than 50× coverage PacBio data are available. This finding is consistent with the previous study in *V. dahliae* genome assembly (Faino et al. 2015). The optimized method generated a chromosome-level genome assembly even though *U. virens* genome contains a much higher percentage of repetitive sequences compared with other ascomycete pathogens, such as *M. oryzae* (Bao et al. 2017) and *F. graminearum* (Cuomo et al. 2007) (Fig. 1). The high-quality assembly enables improved annotations of functional genes and repetitive sequences. Particularly, the effector gene repertoire was greatly enlarged, although some putative effectors with the length of more than 400 amino-acid residues such as glycoside hydrolase-type effectors were excluded from this repertoire (Additional file 1: Table S2) (Fiorin et al. 2018; Sun et al. 2020).

#### The phylogenetic relationships among *U. virens* strains are independent of geographical origins

Previously, we revealed that nucleotide diversity among *U. virens* strains is relatively low (Zhang et al. 2014). Here, we compared raw sequencing data and genome assemblies of several additional *U. virens* strains to further illustrate intraspecific variations. These strains were collected from different countries including China, Japan, the United States and India (Additional file 1: Table S5). The phylogenomic analysis showed that evolutionary distances among *U. virens* strains are negligible compared with those among other pathogenic fungal strains such as *M. anisopilae* and *F. graminearum*, indicating a recent expansion of *U. virens* (Fig. 2a). Consistently, the phylogenetic trees constructed with mitochondrial and nuclear genomic sequences suggest that the evolutionary distances of *U. virens* strains are independent of their geographical origins (Figs. 2b and 6b). This speculation needs to be further confirmed by large-scale re-sequencing of *U. virens* strains from diverse regions.

The interspecific relationships of *U. virens* to other investigated fungal species indicated by mitochondrial and nuclear phylogenetic trees are highly similar (Fig. 2a and Additional file 2: Figure S5). However, the intraspecific relationships among *U. virens* strains revealed by these trees are significantly different. One possible explanation is that nuclear and mitochondrial genomes in *U. virens* evolve independently. As reviewed previously, mitochondrial inheritance patterns vary in different fungi, and most filamentous ascomycetes inherit mitochondria from recipient gametes (Basse 2010; Xu and Li 2015). Large-scale intraspecific structural variations are often generated from chromosome shuffling and

recombination during sexual production (Croll et al. 2013). However, genetic inheritance of mitochondrial DNA follows maternal inheritance law. Therefore, the difference in nuclear and mitochondrial inheritance might be explained by the fact that *U. virens* is heterothallic (Zhang et al. 2014). A better understanding of the mitochondrial inheritance pattern in *U. virens* will facilitate evaluating its potential influence on mitochondrial genomic evolution.

#### Widespread LSRs in *U. virens* strains

Despite relatively low incidence of SNPs, SVs and LSRs are widely present in different *U. virens* strains (Figs. 3a and 4). We speculate that massive intraspecific structural variations are driven by active TEs, especially LTR-retrotransposons, which have been recently inserted into *U. virens* genome (Fig. 3d). This hypothesis is supported by the following results. First, these SVs are predominantly attributable to transposition of TEs, because they almost completely overlap with annotated TEs. Those that cannot be precisely covered by one annotated TE might be the consequence of nested TEs, in which the intact TEs have been destroyed by subsequent insertion of another TE. Second, TEs, particularly intact LTR-retrotransposons, account for the majority of LSRs between UV8b and UVP1, indicating that TEs contribute mostly to LSR formation. Third, *U. virens* strains differ from each other in terms of sizes and locations of LSRs, and the LSR variation is independent of their phylogenetic distances (Figs. 2 and 4). The findings indicate that LSRs are formed after differentiation of *U. virens* strains rather than being inherited from a common ancestor. Fourth, LTR-retrotransposons in *U. virens* undergo a recent burst, especially for those in SVs and LSRs (Fig. 3d). This result coincides with the recent formation of SVs and LSRs, and provides a convincing explanation for emergence of genomic variations. Collectively, these findings suggest that the burst of LTR-retrotransposons drives genomic evolution and reshapes genome structure of *U. virens*. The genomic plasticity, referring to capacity to rapidly evolve, might help pathogens escape host defense by introducing new pathogenicity genes such as effector genes. The hypothesis is supported by the following example. The putative effector genes *UV8b\_05173* and *UV8b\_07599* that may contribute to *U. virens* virulence and pathogenicity are located in the LSRs of UV8b relative to UVP1 (Additional file 2: Figure S6). The results suggest that genomic plasticity created by LSRs might play an important role in the rice-*U. virens* co-evolutionary interactions and special adaptation.

Although the delimitation of LSR might be influenced by several factors including uneven distributions of reads mapped to genome, incomplete assemblies and absence of raw sequencing data for GVT and IPU010, our results

strongly indicate that LSR hotspots exist in *U. virens* genome, such as telomere-proximal TE-enriched regions of chromosomes 2 and 6. Meanwhile, TE-enriched compartments are not always LSR hotspots (Fig. 4). These specific hotspots might be closely related to limited chromatin accessibility and low meiotic recombination rates, as reviewed previously (Dimitri and Junakovic 1999; Kent et al. 2017). In the future, Hi-C and/or ATAC-seq assays, through which the interaction and accessibility of chromatin could be effectively detected, can be conducted to better understand how LSRs form and evolve.

#### Highly conserved coding sequences versus dynamic mobile introns in *U. virens* mitochondrial genome

The large size of *U. virens* mitochondrial genome is mainly attributable to ubiquitous HEG-containing introns in mitochondrial genes and rRNAs (Additional file 1: Table S10). The HEG-containing introns have been reported in other fungal species including *Fusarium*, *Aspergillus* and *Penicillium* species (Al-Reedy et al. 2012; Joardar et al. 2012; Brankovics et al. 2017). By contrast, no intron was found in mitochondrial genes or rRNAs in *M. anisopliae*, which is one of the closest species of *U. virens* among the genome-sequenced species. The finding suggests that HEG gain/loss in mitochondrial genomes occurs after differentiation of these two species. Besides, we revealed that the distribution patterns of HEGs are different among *U. virens* and *Fusarium* species (Al-Reedy et al. 2012; Brankovics et al. 2017). Further analysis found that HEG occurrences in different *U. virens* strains were independent of the intraspecific phylogenetic relationships, suggesting that at least some HEGs invade *U. virens* mitochondrial genome after strain diversification. Given the extremely low genetic diversity in *U. virens* mitochondrial exon sequences, we speculate that these HEG-containing mobile introns are highly active and have inserted into *U. virens* mitochondrial genomes recently. As reviewed previously, accumulation of mutations in HEGs causes them to lose primary functions and eventually to vanish after invasion and colonization in mitochondrial genome (Burt and Koufopanou 2004). Consistently, many HEGs in *U. virens* lose their functions due to missing start/stop codons and premature terminations. For example, the *cob* gene, which is subject to frequent gain and loss of HEGs in introns in many fungal species (Yin et al. 2012; Guha et al. 2018), contained no significant structural variations among *U. virens* strains. This phenomenon is probably attributable to the fact that most of the HEGs in *cob* are functionally incomplete in *U. virens* (Additional file 1: Table S10). These results suggest that various HEGs in *U. virens* exist in a dynamic state, representing different stages in the lifecycle of HEGs. These characteristics of

HEGs in *U. virens* make it an excellent model to understand HEG evolution in filamentous fungi.

#### Conclusions

In short, near-complete genome assemblies of *U. virens* facilitate intraspecific genomic comparisons at both nucleotide and structural levels. Nucleotide sequences are highly conserved in different *U. virens* strains, whereas massive structural variations generated by TEs provide substantial evolutionary plasticity. The plasticity might promote rapid evolution of *U. virens* under intensive selection stresses, and as a consequence, increases the difficulty of developing strategies for management of this important rice disease.

#### Methods

##### DNA extraction and sequencing

The *U. virens* strains, UV8b and UVP1, were cultured on PSA media, and genomic DNA was extracted from mycelia using a CTAB (cetyltrimethyl ammonium bromide) method (Rogers and Bendich 1994). Four libraries with 20 kb insertion size were constructed using the isolated DNA and were then sequenced on a PacBio RS II instrument with P4-C6 chemistry.

##### Genome assembly

De novo genome assembly was performed using the HGAP version 3 in SMRT Analysis v2.3.0. The resultant contigs were polished using Quiver (Chin et al. 2013) and Pilon (Walker et al. 2014) based on the PacBio subreads and Illumina sequence reads (Zhang et al. 2014), respectively. Alternatively, the PacBio subreads were assembled using Canu (Koren et al. 2017) followed by a similar polishing process. BAC sequences from a previous study (Wang et al. 2013) were aligned to the contigs with nucmer (Kurtz et al. 2004). The assembly was manually corrected based on subread mapping, BAC sequences and the overlaps of contigs generated from HGAP and Canu methods. The UVP1 genome was assembled and polished separately following a similar strategy.

##### Genome annotation

The genomic sequences of UV8b and UVP1 were combined for repeat annotation. Repeat elements were predicted by a combination of RepeatModeler (<http://www.repeatmasker.org/RepeatModeler/>) and LTRharvest (Ellinghaus et al. 2008). Specifically, the intact LTR-retrotransposons from LTRharvest were BLASTed (Schäffer et al. 2001) against the assembled genome to remove those with less than five copies, and the remaining ones were merged into the repeat library generated from RepeatModeler. The resultant repeat libraries were merged into non-redundant consensus clusters

(repeat families) using USEARCH (Edgar 2010). These families were annotated and classified using RepeatClassifier module of RepeatModeler, and the unidentified families were annotated by aligning them to RepBase (Bao et al. 2015).

Gene annotation of UV8b genome was performed by combining *ab initio* gene predictions from Augustus (Stanke et al. 2008) and GeneMark-ES (Ter-Hovhannissyan et al. 2008) with homolog mapping from the related species, and with evidence from transcriptome analyses as well (Zhang et al. 2014). The gene ID for the updated gene annotation was assigned as UV8b\_00000 with five digits to distinguish it from the previous version with four digits. The secreted proteins, conventional effector proteins, and PHI-base proteins were predicted as described previously (Zhang et al. 2014). Specifically, conventional effector proteins were replenished using EffectorP based on a machine learning method (Sperschneider et al. 2016). The non-conventional effectors without putative signal peptides that might be located in the apoplast were predicted using ApoplastP (Sperschneider et al. 2018). The resultant proteins with less than 400 amino acids and more than four cysteine residues were considered as potential non-conventional effectors.

### Phylogenetic analysis

The interspecific and intraspecific nuclear phylogenetic trees were constructed based on single-copy homologous proteins in selected fungal species and single-copy genes that were full-length in all genomic assemblies of investigated *U. virens* strains, respectively. The mitochondrial protein sequences of selected fungal species were used for the interspecific mitochondrial tree. Considering that protein sequences were extremely conserved and that dynamic HEGs were widely distributed in intronic regions in *U. virens* mitochondrial genomes, only shared intronic loci in all investigated strains were used to construct an intraspecific tree. DNA and protein sequences were subjected to sequence alignments using ClustalW (Thompson et al. 1994), and were then linked together end-to-end to form a concatenated sequence for tree construction, respectively. Finally, phylogenetic trees were constructed by the Neighbor-Joining method in MEGA X (Kumar et al. 2018) with 1000 bootstraps.

### Comparative genomics

The genomes of UV8b and UVP1 were compared using nucmer (Kurtz et al. 2004). Translocations and inversions were defined as DNA fragments with the alignment length of  $\geq 200$  bp and the identity of  $\geq 95\%$  in different positions and orientations. The DNA regions that were not covered by alignments were considered as potential LSRs. Alternatively, these two genomes were aligned to each other using Mauve (Darling et al. 2004),

which adopts an integrated and stricter algorithm to identify conserved segments. By comparison of results from nucmer and Mauve, the LSRs between UV8b and UVP1 were predicted and further manually polished. For the simulation of LSRs, genomic segments with the same length compared to LSRs were randomly selected from non-conserved regions identified through Mauve. The percentages of intact LTR-retrotransposons located in the predicted and simulated LSRs were compared to evaluate enrichment of LTR-retrotransposon in LSRs. The nucleotide diversity of LTRs was calculated to estimate the insertion time of LTR-retrotransposons. Briefly, a pair of LTRs were extracted from both flanking regions of intact LTR-retrotransposons and were then aligned with each other using Muscle (Edgar 2004), and the SNPs and InDels revealed by alignments were counted to evaluate sequence diversity.

To detect potential LSRs in UV8b compared with other strains, raw Illumina data were aligned to the genome sequence of UV8b. Raw sequencing reads for HWD were obtained from the NCBI SRA database (accession No. SRP051560). For the strains without raw sequencing data (IPU010 and GVT), we simulated paired-end reads with 100 $\times$  coverage and 150 bp in length using ART (Huang et al. 2012) with default parameters. Assemblies for GVT and IPU010 were obtained from the NCBI assembly database (accession No. GCA\_002939685.2 and GCA\_000965225.2). The LSRs were identified through the local Perl script (available upon request). Specifically, the regions with extremely low coverage or no coverage were extracted, and adjacent regions were merged. The merged regions were then filtered to remove small fragments ( $< 100$  bp), which could be caused by the uneven distribution of mapped reads. Only the regions that were found as LSRs relative to all other nine strains were defined as putative UV8b-specific LSRs.

### Assembly, annotation and intraspecific comparison of mitochondrial genomes

Mitochondrial gene annotations of fungal species were downloaded from the MiToFun database (<http://mitofun.biol.uoa.gr/>). The mitochondrial DNA contigs were extracted by aligning mitochondrial genes in *Fusarium* species to assembled contigs of *U. virens*. The mitochondrial genes were annotated by Mfannot (<http://megasun.bch.umontreal.ca/RNAweasel/>) and via alignments with the mitochondrial genes in phylogenetically related species followed by manual corrections. Mitochondrial tRNA genes were predicted through RNAweasel (<http://megasun.bch.umontreal.ca/RNAweasel/>). The ORFs in introns of mitochondrial genes were predicted through Mfannot, and were then subjected to BLAST and Pfam (Finn et al. 2014) searches for correction of the boundaries and prediction of functional domains.

The mitochondrial genomes of FJ1a, HB31a, LN201011, LN10161 and UV86 were constructed using the previously assembled contigs (Zhang et al. 2014). The raw sequencing data of the strain HWD were assembled using SPAdes (Bankevich et al. 2012). The mitochondrial DNA contigs were extracted from assemblies and were connected according to their alignments with UV8b mitochondrial genome. The resultant mitochondrial genomes were further polished using Pilon. The mitochondrial genes for these strains were annotated by aligning with the predicted genes in UV8b using Exonerate (Slater and Birney 2005), followed by manual correction.

#### Abbreviations

BES: BAC end sequence; CDS: Conserved coding sequence; CTAB: Cetyltrimethyl ammonium bromide; *F. graminearum*: *Fusarium graminearum*; HEG: Homing endonuclease gene; HGAP: Hierarchical genome-assembly process; LSR: Lineage-specific region; LTR: Long terminal repeat; *M. anisopliae*: *Metarhizium anisopliae*; *M. oryzae*: *Magnaporthe oryzae*; ORF: Open reading frame; RFS: Rice false smut; SMRT: Single-molecule real-time; SNP: Single nucleotide polymorphism; SSR: Single sequence repeat; SV: Structural variation; TE: Transposable element; *U. virens*: *Ustilagoidea virens*; *V. dahliae*: *Verticillium dahliae*

#### Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s42483-021-00086-x>.

**Additional file 1: Table S1.** Chromosome lengths in genome assemblies of *U. virens* strains UV8b and UVP1. **Table S2.** List of candidate effector proteins in *U. virens*. **Table S3.** The average distance of effector genes apart from adjacent transposable elements in *U. virens* genome. **Table S4.** Conservation analysis of candidate effector proteins in *U. virens* compared with the closely related fungal species *M. anisopliae* and *F. graminearum*. **Table S5.** Location of the *U. virens* strains collected in this study. **Table S6.** SNP frequencies in coding sequences of candidate effector- and secreted protein-encoding genes in different *U. virens* strains compared with the reference genome. **Table S7.** The predicted structural variations between *U. virens* UV8b and UVP1 genomes. **Table S8.** Lineage specific regions in UV8b genome identified through comparisons with other *U. virens* strains. **Table S9.** The sizes of fungal mitochondrial genomes collected from MiToFun database. **Table S10.** The putative open reading frames in intronic regions of the mitochondrial genes in UV8b. **Table S11.** The sizes of mitochondrial genomes in different *U. virens* strains. **Table S12.** The quality of genome assemblies using non-hybrid and hybrid assembling methods.

**Additional file 2: Figure S1.** SNP frequencies in coding, intronic and intergenic sequences in different *U. virens* strains relative to UV8b genome. The strains are indicated by different colors. Genomic sequences of these strains were aligned to the reference genome with nucmer, and SNPs were called with "show-snps", a tool in Mummer package. The genes exhibiting full-length gene sequences in the assemblies of all these strains were subjected to SNP calculation. For each strain, lineage-specific regions were excluded from the intergenic regions to avoid the potential miscalculation. **Figure S2.** Potential inter-chromosomal translocation events illustrated by alignments of DNA sequences between UV8b and UVP1 genomes. Syntenic alignments are indicated by blue stripes. The inter-chromosomal translocation region is indicated by red box. Alignment analysis between these two genomes was conducted with nucmer. The structure variations were identified with the thresholds of the alignment length of  $\geq 200$  bp and the identity of  $\geq 95\%$ . **Figure S3.** The nucleotide diversities in long terminal repeats (LTRs) at flanking regions of LTR-retrotransposons in SVs, LSRs and other regions of UVP1 genome. The LTRs at both flanking regions of LTR-

retrotransposons were aligned to each other and the nucleotide diversities were calculated. LTR-RTs, LTR-retrotransposons. **Figure S4.** The distribution of the lineage-specific regions (LSRs) of UV8b relative to multiple *U. virens* strains. The lines from top to bottom indicate the genes (orange), transposable elements (TEs) (blue), heatmap of SNPs compared with multiple strains (red), UV8b-specific LSRs that are absent in all other strains (green), and LSRs compared with the indicated strains (grey). LSRs relative to a specific strain were identified by aligning the sequencing reads of this strain to UV8b genome. For IPU010 and GVT without publicly available sequencing reads, the Illumina reads with 100x coverage were simulated based on their assemblies. The other three chromosomes (2, 3, 6) were displayed in Fig. 4. UV8b-LSR, UV8b-specific LSRs. **Figure S5.** The phylogenetic relationship among *U. virens* and the related fungal species based on mitochondrial protein sequences. Homologous mitochondrial protein sequences in these indicated fungal species were aligned using ClustalW, and the tree was constructed using the Neighbor-Joining method in MEGA X with 1000 bootstraps. **Figure S6.** Certain candidate effector genes are located at the lineage-specific regions of UV8b relative to UVP1. Syntenic alignments are indicated by blue stripes. The effector genes are indicated by red boxes, while other genes are indicated by grey boxes. Alignment analysis was performed as described in Figure S2.

#### Acknowledgments

We are grateful to Dr. Myco Umemura at Technology Research Association of Highly Efficient Gene Design, Sapporo, Hokkaido, Japan for providing genomic data of IPU010.

#### Authors' contributions

WS and KZ designed the study. KZ analyzed the data. ZZ (Zaixu Zhao) and ZZ (Ziding Zhang) participated in data analysis and interpretation. YL cultured the strains and carried out the DNA extraction experiment. SL and NY provided the strains and sequencing data. WS, KZ and TH wrote this manuscript. All authors read and approved the final manuscript.

#### Funding

The work is supported by the National Natural Science Foundation of China grants 31630064 and U19A2027, the Natural Science Foundation of Beijing, China (6181001) and the 111 project B13006 to Wenxian Sun.

#### Availability of data and materials

The assemblies for the UV8b and UVP1 strains generated in this study have been deposited in the NCBI database under BioProject accessions PRJNA606646 and PRJNA610522, respectively.

#### Declarations

##### Ethics approval and consent to participate

Not applicable.

##### Consent for publication

Not applicable.

##### Competing interests

The authors declare that they have no competing interests.

##### Author details

<sup>1</sup>Department of Plant Pathology, The Ministry of Agriculture Key Laboratory of Pest Monitoring and Green Management, Joint Laboratory for International Cooperation in Crop Molecular Breeding, Ministry of Education, China Agricultural University, Beijing 100193, China. <sup>2</sup>Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Key Laboratory of Biology and Genetic Improvement of Horticultural Crops of the Ministry of Agriculture, Sino-Dutch Joint Laboratory of Horticultural Genomics, No.12 Zhongguancun South Street, Haidian District, Beijing 100081, China. <sup>3</sup>State Key Laboratory of Agrobiotechnology, College of Biological Sciences, China Agricultural University, Beijing 100193, China. <sup>4</sup>Institute of Microbiology, Chinese Academy of Sciences, Beijing 100101, China. <sup>5</sup>State Key Laboratory of Biocontrol, Guangdong Provincial Key Laboratory of Plant Resources, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China.

<sup>6</sup>School of Environmental Sciences, University of Guelph, Guelph, Ontario N1G 2W1, Canada. <sup>7</sup>College of Plant Protection, Jilin Agricultural University, 2888 Xincheng Street, Changchun 130118, Jilin, China.

Received: 9 February 2021 Accepted: 13 April 2021  
Published online: 28 April 2021

## References

- Alkan C, Sajjadian S, Eichler EE. Limitations of next-generation genome sequence assembly. *Nat Methods*. 2011;8(1):61–5. <https://doi.org/10.1038/nmeth.1527>.
- Al-Reedy RM, Malireddy R, Dillman CB, Kennell JC. Comparative analysis of *Fusarium* mitochondrial genomes reveals a highly variable region that encodes an exceptionally large open reading frame. *Fungal Genet Biol*. 2012; 49(1):2–14. <https://doi.org/10.1016/j.fgb.2011.11.008>.
- Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. *J Comput Biol*. 2012;19(5):455–77. <https://doi.org/10.1089/cmb.2012.0021>.
- Bao J, Chen M, Zhong Z, Tang W, Lin L, Zhang X, et al. PacBio sequencing reveals transposable elements as a key contributor to genomic plasticity and virulence variation in *Magnaporthe oryzae*. *Mol Plant*. 2017;10(11):1465–8. <https://doi.org/10.1016/j.molp.2017.08.008>.
- Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. *Mob DNA*. 2015;6(1):11. <https://doi.org/10.1186/s13100-015-0041-9>.
- Basse CW. Mitochondrial inheritance in fungi. *Curr Opin Microbiol*. 2010;13(6): 712–9. <https://doi.org/10.1016/j.mib.2010.09.003>.
- Bertazzoni S, Williams AH, Jones DA, Syme RA, Tan K-C, Hane JK. Accessories make the outfit: accessory chromosomes and other dispensable DNA regions in plant-pathogenic fungi. *Mol Plant-Microbe Interact*. 2018;31(8):779–88. <https://doi.org/10.1094/MPMI-06-17-0135-FI>.
- Brankovics B, van Dam P, Rep M, de Hoog GS, van der Lee TAJ, Waalwijk C, et al. Mitochondrial genomes reveal recombination in the presumed asexual *Fusarium oxysporum* species complex. *BMC Genomics*. 2017;18(1):735. <https://doi.org/10.1186/s12864-017-4116-5>.
- Burt A, Koufopanov V. Homing endonuclease genes: the rise and fall and rise again of a selfish element. *Curr Opin Genet Dev*. 2004;14(6):609–15. <https://doi.org/10.1016/j.gde.2004.09.010>.
- Chin C-S, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat Methods*. 2013;10(6):563–9. <https://doi.org/10.1038/nmeth.2474>.
- Cooke MC. Some extra-European fungi. *Grevillea*. 1878;7:13–5.
- Croll D, Zala M, McDonald BA. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. *PLoS Genet*. 2013;9(6):e1003567. <https://doi.org/10.1371/journal.pgen.1003567>.
- Cuomo CA, Güldener U, Xu J-R, Trail F, Turgeon BG, Pietro AD, et al. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*. 2007;317(5843):1400–2. <https://doi.org/10.1126/science.1143708>.
- Darling ACE, Mau B, Blattner FR, Perna NT. Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res*. 2004;14(7): 1394–403. <https://doi.org/10.1101/gr.2289704>.
- de Jonge R, Bolton MD, Kombrink A, van den Berg GCM, Yadeta KA, Thomma BPHJ. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res*. 2013;23(8):1271–82. <https://doi.org/10.1101/gr.152660.112>.
- Dimitri P, Junakov N. Revising the selfish DNA hypothesis: new evidence on accumulation of transposable elements in heterochromatin. *Trends Genet*. 1999;15(4):123–4. [https://doi.org/10.1016/S0168-9525\(99\)01711-4](https://doi.org/10.1016/S0168-9525(99)01711-4).
- Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr Opin Genet Dev*. 2015;35:57–65. <https://doi.org/10.1016/j.gde.2015.09.001>.
- Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32(5):1792–7. <https://doi.org/10.1093/nar/gkh340>.
- Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*. 2010;26(19):2460–1. <https://doi.org/10.1093/bioinformatics/btq461>.
- Elinghaus D, Kurtz S, Willhoeft U. *LTRharvest*, an efficient and flexible software for *de novo* detection of LTR retrotransposons. *BMC Bioinf*. 2008;9(1):18. <https://doi.org/10.1186/1471-2105-9-18>.
- Faino L, Seidl MF, Datema E, van den Berg GCM, Janssen A, Wittenberg AHJ, et al. Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *mBio*. 2015;6:e00936–15.
- Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCM, Wittenberg AHJ, et al. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. *Genome Res*. 2016;26(8):1091–100. <https://doi.org/10.1101/gr.204974.116>.
- Faino L, Thomma BPHJ. Get your high-quality low-cost genome sequence. *Trends Plant Sci*. 2014;19(5):288–91. <https://doi.org/10.1016/j.tplants.2014.02.003>.
- Fan J, Yang J, Wang Y-Q, Li G-B, Li Y, Huang F, et al. Current understanding on *Villosiclava virens*, a unique flower-infecting fungus causing rice false smut disease: a profile on the rice false smut pathogen. *Mol Plant Pathol*. 2016; 17(9):1321–30. <https://doi.org/10.1111/mpp.12362>.
- Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. *Nucleic Acids Res*. 2014;42(Database issue):D222–30. <https://doi.org/10.1093/nar/gkt1223>.
- Fiorin GL, Sánchez-Vallet A, de Toledo Thomazella DP, do Prado PFV, do Nascimento LC, de Oliveira Figueira AV, et al. Suppression of plant immunity by fungal chitinase-like effectors. *Curr Biol*. 2018;28:3023–3030.e5.
- Goodwin SB, M'Barek SB, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensable structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet*. 2011;7(6):e1002070. <https://doi.org/10.1371/journal.pgen.1002070>.
- Guha TK, Wai A, Mullineux S-T, Hausner G. The intron landscape of the mtDNA *cytb* gene among the Ascomycota: introns and intron-encoded open reading frames. *Mitochondrial DNA A*. 2018;29(7):1015–24. <https://doi.org/10.1080/24701394.2017.1404042>.
- Guo X, Li Y, Fan J, Li L, Huang F, Wang W. Progress in the study of false smut disease in rice. *J Agric Sci Technol A*. 2012;2:1211–7.
- Han Y, Zhang K, Yang J, Zhang N, Fang A, Zhang Y, et al. Differential expression profiling of the early response to *Ustilaginoidea virens* between false smut resistant and susceptible rice varieties. *BMC Genomics*. 2015;16(1):955. <https://doi.org/10.1186/s12864-015-2193-x>.
- Huang W, Li L, Myers JR, Marth GT. ART: a next-generation sequencing read simulator. *Bioinformatics*. 2012;28(4):593–4. <https://doi.org/10.1093/bioinformatics/btr708>.
- Ikegami H. Occurrence and development of sclerotia of the rice false smut fungus. *Res Bull Fac Agric*. 1963;18:47–53.
- Joardar V, Abrams NF, Hostetler J, Paukstelis PJ, Pakala S, Pakala SB, et al. Sequencing of mitochondrial genomes of nine *Aspergillus* and *Penicillium* species identifies mobile introns and accessory genes as main sources of genome size variability. *BMC Genomics*. 2012;13(1):698. <https://doi.org/10.1186/1471-2164-13-698>.
- Kent TV, Uzunović J, Wright SI. Coevolution between transposable elements and recombination. *Philos Trans R Soc B Biol Sci*. 2017;372(1736):20160458. <https://doi.org/10.1098/rstb.2016.0458>.
- Koiso Y, Li Y, Iwasaki S, Hanaoka K, Kobayashi T, Sonoda R, et al. Ustiloxins, antimicrobial cyclic peptides from false smut balls on rice panicles caused by *Ustilaginoidea virens*. *J Antibiot*. 1994;47(7):765–73. <https://doi.org/10.7164/antibiotics.47.765>.
- Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive *k*-mer weighting and repeat separation. *Genome Res*. 2017;27(5):722–36. <https://doi.org/10.1101/gr.215087.116>.
- Kumagai T, Ishii T, Terai G, Umemura M, Machida M, Asai K. Genome sequence of *Ustilaginoidea virens* IPU010, a rice pathogenic fungus causing false smut. *Microbiol Resour Announce*. 2016;4:e00306–16.
- Kumar A, Pandey V, Singh M, Pandey D, Saharan MS, Marla SS. Draft genome sequence of Karnal bunt pathogen (*Tilletia indica*) of wheat provides insights into the pathogenic mechanisms of quarantined fungus. *PLoS One*. 2017; 12(2):e0171323. <https://doi.org/10.1371/journal.pone.0171323>.
- Kumar S, Stecher G, Li M, Knyaz C, Tamura K. MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol Biol Evol*. 2018;35(6):1547–9. <https://doi.org/10.1093/molbev/msy096>.
- Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. *Genome Biol*. 2004;5(2):R12. <https://doi.org/10.1186/gb-2004-5-2-r12>.

- Li Y, Koiso Y, Kobayashi H, Hashimoto Y, Iwasaki S. Ustiloxins, new antimitotic cyclic peptides: interaction with porcine brain tubulin. *Biochem Pharmacol.* 1995;49(10):1367–72. [https://doi.org/10.1016/0006-2952\(95\)00072-8](https://doi.org/10.1016/0006-2952(95)00072-8).
- Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature.* 2010;464(7287):367–73. <https://doi.org/10.1038/nature08850>.
- Möller M, Stukenbrock EH. Evolution and genome architecture in fungal plant pathogens. *Nat Rev Microbiol.* 2017;15(12):756–71. <https://doi.org/10.1038/nrmicro.2017.76>.
- Nombela C, Gil C, Chaffin WL. Non-conventional protein secretion in yeast. *Trends Microbiol.* 2006;14(1):15–21. <https://doi.org/10.1016/j.tim.2005.11.009>.
- Plissonneau C, Stürchler A, Croll D. The evolution of orphan regions in genomes of a fungal pathogen of wheat. *mBio.* 2016;7:e01231–16.
- Qi X, Li Y, Honda S, Hoffmann S, Marz M, Mosig A, et al. The common ancestral core of vertebrate and fungal telomerase RNAs. *Nucleic Acids Res.* 2013;41(1):450–62. <https://doi.org/10.1093/nar/gks980>.
- Qiu J, Meng S, Deng Y, Huang S, Kou Y. *Ustilagoidea virens*: a fungus infects rice flower and threatens world rice production. *Rice Sci.* 2019;26(4):199–206. <https://doi.org/10.1016/j.rsci.2018.10.007>.
- Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat Rev Microbiol.* 2012;10(6):417–30. <https://doi.org/10.1038/nrmicro.2012.79>.
- Roberts RJ, Belfort M, Bestor T, Bhagwat AS, Bickle TA, Bitinaite J, et al. A nomenclature for restriction enzymes, DNA methyltransferases, homing endonucleases and their genes. *Nucleic Acids Res.* 2003;31(7):1805–12. <https://doi.org/10.1093/nar/gkg274>.
- Rogers SO, Bendich AJ. Extraction of total cellular DNA from plants, algae and fungi. In: Gelvin SB, Schilperoort RA, editors. *Plant molecular biology manual*. Dordrecht: Springer; 1994. p. 183–90.
- Sánchez-Vallet A, Fouché S, Fudal I, Hartmann FE, Soyer JL, Tellier A, et al. The genome biology of effector gene evolution in filamentous plant pathogens. *Annu Rev Phytopathol.* 2018;56(1):21–40. <https://doi.org/10.1146/annurev-phyto-080516-035303>.
- Schäffer AA, Aravind L, Madden TL, Shavirin S, Spouge JL, Wolf YI, et al. Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.* 2001;29(14):2994–3005. <https://doi.org/10.1093/nar/29.14.2994>.
- Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics.* 2015;31(19):3210–2. <https://doi.org/10.1093/bioinformatics/btv351>.
- Slater GSC, Birney E. Automated generation of heuristics for biological sequence comparison. *BMC Bioinf.* 2005;6(1):31. <https://doi.org/10.1186/1471-2105-6-31>.
- Sperschneider J, Dodds PN, Singh KB, Taylor JM. ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytol.* 2018;217(4):1764–78. <https://doi.org/10.1111/nph.14946>.
- Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, et al. EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytol.* 2016;210(2):743–61. <https://doi.org/10.1111/nph.13794>.
- Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntentically mapped cDNA alignments to improve *de novo* gene finding. *Bioinformatics.* 2008;24(5):637–44. <https://doi.org/10.1093/bioinformatics/btn013>.
- Sun W, Fan J, Fang A, Li Y, Tariqjaveed M, Li D, et al. *Ustilagoidea virens*: insights into an emerging rice pathogen. *Annu Rev Phytopathol.* 2020;58(1):363–85. <https://doi.org/10.1146/annurev-phyto-010820-012908>.
- Tanaka E, Ashizawa T, Sonoda R, Tanaka C. *Villosiclava virens* gen. nov., comb. nov., teleomorph of *Ustilagoidea virens*, the causal agent of rice false smut. *Mycotaxon.* 2008;106:491–501.
- Ter-Hovhannisyantsyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res.* 2008;18(12):1979–90. <https://doi.org/10.1101/gr.081612.108>.
- Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* 1994;22(22):4673–80. <https://doi.org/10.1093/nar/22.22.4673>.
- Tsushima A, Gan P, Kumakura N, Narusaka M, Takano Y, Narusaka Y, et al. Genomic plasticity mediated by transposable elements in the plant pathogenic fungus *Colletotrichum higginsianum*. *Genome Biol Evol.* 2019;11(5):1487–500. <https://doi.org/10.1093/gbe/evz087>.
- van Kan JAL, Stassen JHM, Mosbach A, van der Lee TAJ, Faino L, Farmer AD, et al. A gapless genome sequence of the fungus *Botrytis cinerea*. *Mol Plant Pathol.* 2017;18(1):75–89. <https://doi.org/10.1111/mpp.12384>.
- Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One.* 2014;9(11):e112963. <https://doi.org/10.1371/journal.pone.0112963>.
- Wang X, Liu Q, Wang H, Luo C-X, Wang G, Luo M. A BAC based physical map and genome survey of the rice false smut fungus *Villosiclava virens*. *BMC Genomics.* 2013;14(1):883. <https://doi.org/10.1186/1471-2164-14-883>.
- Wolters PJ, Faino L, van den Bosch TBM, Evenhuis B, Visser RGF, Seidl MF, et al. Gapless genome assembly of the potato and tomato early blight pathogen *Alternaria solani*. *Mol Plant-Microbe Interact.* 2018;31(7):692–4. <https://doi.org/10.1094/MPMI-12-17-0309-A>.
- Xu J, Li H. Current perspectives on mitochondrial inheritance in fungi. *Cell Health Cytoskeleton.* 2015;7:143–54.
- Ye C, Hill CM, Wu S, Ruan J, Ma Z. DBG2OLC: efficient assembly of large genomes using long erroneous reads of the third generation sequencing technologies. *Sci Rep.* 2016;6(1):31900. <https://doi.org/10.1038/srep31900>.
- Yin L-F, Hu M-J, Wang F, Kuang H, Zhang Y, Schnabel G, et al. Frequent gain and loss of introns in fungal cytochrome b genes. *PLoS One.* 2012;7(11):e49096. <https://doi.org/10.1371/journal.pone.0049096>.
- Yu J-J, Yu M-N, Nie Y-F, Sun W-X, Yin X-L, Zhao J, et al. Comparative transcriptome analysis of fruiting body and sporulating mycelia of *Villosiclava virens* reveals genes with putative functions in sexual reproduction. *Curr Genet.* 2016a;62(3):575–84. <https://doi.org/10.1007/s00294-015-0563-1>.
- Yu M, Yu J, Li H, Wang Y, Yin X, Bo H, et al. Survey and analysis of simple sequence repeats in the *Ustilagoidea virens* genome and the development of microsatellite markers. *Gene.* 2016b;585(1):28–34. <https://doi.org/10.1016/j.gene.2016.03.016>.
- Zhang K, Li Y, Li T, Li Z-G, Hsiang T, Zhang Z, et al. Pathogenicity genes in *Ustilagoidea virens* revealed by a predicted protein–protein interaction network. *J Proteome Res.* 2017;16(3):1193–206. <https://doi.org/10.1021/acs.jproteome.6b00720>.
- Zhang Y, Zhang K, Fang A, Han Y, Yang J, Xue M, et al. Specific adaptation of *Ustilagoidea virens* in occupying host florets revealed by comparative and functional genomics. *Nat Commun.* 2014;5(1):3849. <https://doi.org/10.1038/ncomms4849>.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

