Genome Biology and Evolution Advance Access published July 24, 2014 doi:10.1093/gbe/evu148

Gene loss rather than gene gain is associated with a host jump from monocots to dicots in the smut fungus *Melanopsichium pennsylvanicum*

Rahul Sharma [1,2,3,4], Bagdevi Mishra [1,2,3], Fabian Runge [5], Marco Thines [1,2,3,4*]

[1] Biodiversity and Climate Research Centre (BiK-F), Georg-Voigt-Str. 14-16, 60325 Frankfurt (Main), Germany.

[2] Institute of Ecology, Evolution and Diversity, Goethe University, Max-von-Laue-Str. 9, 60323 Frankfurt (Main), Germany.

[3] Senckenberg Gesellschaft für Naturforschung, Senckenberganlage 25, 60325 Frankfurt (Main), Germany.

[4] Center for Integrative Fungal Research (IPF), Georg-Voigt-Str. 14-16, 60325 Frankfurt (Main), Germany.

[5] University of Hohenheim, Institute of Botany 210, 70593 Stuttgart, Germany.

Corresponding author:

Marco Thines

Integrative Fungal Research (IPF), Biodiversity and Climate Research Centre (BiK-F),

Senckenberganlage 25, D-60325

Frankfurt am Main, Germany

Tel: +49 (0)69 754 21833

Fax: +49 (0)69 798 24780

marco.thines@senckenberg.de

Running title: Biotroph pathogen host jumps trigger gene loss

Keywords: Comparative Genomics, Effector Genes, Evolutionary Biology, Genome Assembly, Host Jump, Positive Selection, Smut Fungi.

Abstract

Smut fungi are well-suited to investigate the ecology and evolution of plant pathogens, as they are strictly biotrophic, yet cultivable on media. Here we report the genome sequence of *Melanopsichium pennsylvanicum*, closely related to *Ustilago maydis* and other Poaceae-infecting smuts, but parasitic to a dicot plant. To explore the evolutionary patterns resulting from host adaptation after this huge host jump, the genome of *M. pennsylvanicum* was sequenced and compared to the genomes of *Ustilago maydis*, *Sporisorium reilianum*, and *Ustilago hordei*. While all four genomes had a similar completeness in CEGMA analyses, gene absence was highest in *M. pennsylvanicum*, and most pronounced in putative secreted proteins, which are often considered as effector candidates. In contrast, the amount of private genes was similar among the species, highlighting that gene loss rather than gene gain is the hallmark of adaptation after the host jump to the dicot host. Our analyses revealed a trend of putative effectors to be next to another putative effector, but the majority of these are not in clusters and thus the focus on pathogenicity clusters might not be appropriate for all smut genomes. Positive selection studies revealed that *M. pennsylvanicum* has the highest number and proportion of genes under positive selection. In general, putative effectors showed a higher proportion of positively selected genes than non-effector candidates. The 248 putative secreted effectors found in all four smut genomes might constitute a core set needed for pathogenicity, while those 92 that are found in all grass-parasitic smuts, but have no ortholog in *M. pennsylvanicum* might constitute a set of effectors important for successful colonization of grass hosts.

Introduction

Melanopsichium pennsylvanicum is a non-obligate biotrophic pathogen and is responsible for gall smut of *Persicaria* species (Hirschhorn, 1941), forming sturdy lobe-shaped smut galls on the host plant, like other *Melanopsichium* species (Fischer, 1953; McAlpine, 1910). Most of the members of Ustilaginaceae are parasitic to Poales and Cyperales, and some are responsible for infecting economically important cereal crops like maize, barley, wheat, and oat (Vánky, 1994). As an exception among Ustilaginaceae, *M. pennsylvanicum* colonizes the dicot genus *Persicaria* (Begerow et al., 2000; Begerow et al., 2006). Mature galls are often covered with black material, which hardens upon desiccation (Barbe, 1962; Vánky, 2002), in contrast to most smuts, which liberate a powder of dark-coloured spores from their galls. Molecular phylogenetic studies have revealed that *M. pennsylvanicum* is embedded in Ustilago s.l. infecting *Poaceae* (Begerow et al., 2004; Stoll et al., 2005; Weiß M, 2004). Other, more distantly related species of the Ustilaginaceae are parasitic to the monocot Cyperaceae and Juncaceae (Begerow et al., 2000; Begerow et al., 2004).

Hemibiotrophic and biotrophic filamentous plant pathogens manipulate their hosts with a suite of effector proteins, which are secreted by the pathogens and function in the apoplast or are translocated into the host plant cell, where they exert their function. Past studies have characterized several effectors secreted by fungal and oomycete plant pathogens (Birch et al., 2008; Djamei et al., 2011; Doehlemann et al., 2009; Kamoun, 2006; Tyler, 2009). Effector proteins generally have a conserved N-terminal signal domain that directs the effector proteins to the host and a C-terminal domain, which is often under strong selection pressure and is responsible for the virulence effects on the host tissues (Win et al., 2007). A huge amount of putative secreted effector proteins (PSEPs) has been reported in the genomes of the smuts *Ustilago hordei* (Laurie et al., 2012), *Sporisorium reilianum* (Schirawski et al., 2010) and

Ustilago maydis (Kämper et al., 2006). In general, PSEPs show higher sequence divergence and less sequence conservation than the non-effector proteins (Schirawski et al. 2010). Many of these secreted effectors have been reported to be organized into pathogenicity clusters in the *U. maydis* genome (Kämper et al., 2006) and comparative studies have been performed to estimate the conservation of these clusters within the other smut genomes (Laurie et al., 2012; Schirawski et al., 2010).

However, all the currently available smut genomes are from hosts within Poaceae, which makes it difficult to identify the core set of conserved effectors that are needed for plant colonization and the more variable effector complement that is needed to exploit a certain group of hosts. Thus, *Melanopsichium* species, which evolved as the result of a host jump to dicots, offer the possibility to address several major questions in plant pathogen evolution. These include the following: What general changes can be observed in the genomes after a long-range host jump? Is the adaptation to the new host associated with gene gain or gene loss? Is there a suit of core pathogenicity effector genes? To what extent are also non-effector genes affected by the adaptation process?

To address the above-mentioned questions, whole genome sequencing, assembly and annotation of the *M. pennsylvanicum* strain 4 (Mp4) was performed using high-throughput sequencing technologies and bioinformatic tools. The bioinformatic analyses presented in this study shed light on several evolutionary events after long-range host jumps and provide a basis for future functional investigations into the biology and function of effectors of smut fungi.

The present study was conducted using available bioinformatics tools and newly developed shell/perl scripts. Solely computational approaches were used to perform the analyses. Thus, even though multiple computational approaches were applied to crosscheck the outcome of

a single tool, the findings of our study should be substantiated by experimental data in future studies.

Results

Illumina genome sequencing, assembly and completeness estimation

After filtering out reads with N's and low quality reads (quality <26), 44,636,214 reads were used for genome assemblies, reaching an average coverage depth of 339.23 at an assumed haploid genome size of 20 Mb. Velvet (Zerbino and Birney, 2008) generated 1746 contigs with an N50 contig size of 43.37 kbps, and the largest contig was of 218.8 kbps. The final scaffolded nuclear genome assembly was of 19,156,659 bps and had an N50 scaffold size of 121.67 kbps; the largest scaffold was of 690.5 kbps, with 434 scaffolds in total. The mitochondrial genome was of 74,914 bps. Assembly quality was estimated by computing the size and number of scaffolds in respective N-classes (Figure 1A). After generating the final assemblies, all reads initially obtained by Illumina sequencing were mapped back onto the generated scaffolds, and 99.6% of the reads were successfully mapped back.

Completeness and continuity of the assembly was tested using the CEGMA pipeline (Parra et al., 2007) and compared with the completeness and continuity of the other three Ustilaginaceae genomes available. The completeness and continuity of the gene space of *Melanopsichium pennsylvanicum* was found to be comparable with the three available genomes: 95.16, 94.76, 95.97 and 95.16% of highly conserved genes were found in *M. pennsylvanicum* (*Mp4*), *Ustilago hordei* (*Uh*), *Sporisorium reilianum* (*Sr*) and *U. maydis* (*Uh*), respectively (Figure 1B). The genome of *M. pennsylvanicum* was aligned with the other three smut genomes

using Mummer3 (Kurtz et al., 2004) and mapped well on the other three genomes (Supplementary Figure 1) and on all 23 chromosomes of both *U. maydis* and *S. reilianum*.

Repeat elements and Gene predictions

A total of 3.15% of the genome consisted of masked repeat elements. After masking all repeat elements, genes were predicted using both *ab-initio* and homology based methods (See Material and Methods). A total of 6280 genes were identified in the Mp4 genome including 107 transfer RNAs.

To estimate the number of putative secreted effectors proteins (PSEPs) both SignalP v4 (Petersen et al., 2011) and TargetP v1 (Emanuelsson et al., 2007) were used on all of the four Ustilaginales genomes. These predictions generated 418 PSEP-encoding gene in the genome of *M. pennsylvanicum*. When applying the same methodology on the other three smut genomes 545, 633 and 629 PSEP-encoding genes were identified in the genomes of *U. hordei*, *S. reilianum* and *U. maydis*, respectively.

Orthologous genes

Orthologs and paralogs were identified using OrthoMCL (Li et al., 2003) and inparanoid (Ostlund et al., 2010) tools, further confirmations were done using BlastP and tBlastn searches (See methods). In total 5277 orthologs were found that were present in all four species, out of these 5200 were 1:1 orthologs (Figure 2A). In total, 623 genes present only in *M. pennsylvanicum* had no ortholog in the other smut genomes. Similarly 772, 449 and 580 were present only in the genomes of *U. hordei*, *S. reilianum* and *U.* maydis, respectively. Interestingly, 429 orthologs were present in *U. hordei*, *S. reilianum* and *U. maydis*, but absent in *M. pennsylvanicum.* In contrast 147, 37 and 61 orthologous genes were absent in *U. hordei*, *S. reilianum* and *U.* maydis, respectively, but present in the corresponding other three genomes.

Similar predictions including only PSEP-encoding genes showed that 248 PSEPs were present in all genomes, and 92 PSEPs were present in the all genomes except for *M. pennsylvanicum.* In contrast 36, 17 and 15 PSEPs were absent in *U. hordei*, *S. reilianum* and *U. maydis*, respectively, but present in all other corresponding genomes (Figure 2B).

Ortholog prediction also identified the orthologs of genes corresponding to the mating-type loci of *U. maydis*, these genes were *mp02686* (bE1), *mp02694* (bW1) and *mp02947* (Pheromone receptor 1) in *M. pennsylvanicum*.

Gene gain and gene loss

For assuring also the complete absence of similar genes, which did not fulfill the criteria for ortholgy but still show limited similarity, BlastP (Altschul et al., 1990) searches for the orthologs that were found in three genomes and were absent in the fourth were performed locally using standalone Blast (e-value less than 0.1 and percentage identity greater than 35%). In addition, also the intergenic regions were again scanned for genes that might have been missed in the annotations. These searches, including the ten stretches of intergenic sequences found by blast that were not predicted as genes, resulted in 292 genes present in all other species but absent in *M. pennsylvanicum*, 99 in *U. hordei*, 17 in *S. reilianum* and 47 in *U. maydis*. The list and functional annotations of the 292 genes lost in *M. pennsylvanicum* but present in the grass-infecting smuts are given in Supplementary Table 1. To screen for gene remnants in intergenic regions of all genomes, relaxed tBlastn searches were performed using a percentage cutoff of 35% and an alignment length of at least 30% of the query sequence length. These predictions resulted in genes present in one genome, but with no gene remnants or distantly similar genes in the other three species. There were 136 such genes in *M. pennsylvanicum*, 69 in *U. hordei*, 14 in *S. reilianum* and 42 in *U. maydis* (Figure 3). The Functional annotations of these lost genes are given in Supplementary Table 1.

In terms of PSEP-encoding genes, 57 (Supplementary Table 1) were absent in the genome of *M. pennsylvanicum*, 17, 3 and 2 were absent in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively (Figure 3). Of these genes 44, 13, 1, 2 of *M. pennsylvanicum*, *U. hordei*, *S. reilianum*, and *U. maydis*, respectively, had no gene remnants or distantly similar genes in the other three genomes.

The 623, 772, 449, and 580 genes found in *M. pennsylvanicum*, *U. hordei*, *S. reilianum*, and *U. maydis*, respectively, which were not having a predicted ortholog in the corresponding other three genomes were further analyzed as described for gene losses. These searches revealed that 324, 510, 335, and 422 genes of *M. pennsylvanicum*, *U. hordei*, *S. reilianum* and *U. maydis*, respectively, had no similar gene in the corresponding other three genomes and were thus considered as gene gains. Regarding PSEP-encoding genes, only 40 were gained in the genome of *M. pennsylvanicum*, but 75, 93 and 75 were gained in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively (Figure 3). When including searches for gene remnants and distantly similar genes, the respective figures were 291, 461, 274, and 395 for *M. pennsylvanicum*, *U. hordei*, *S. reilianum* and *U. maydis*, respectively, for all genes, of which 39, 69, 82, and 67 were encoding for PSEPs, respectively. Thus, while *M. pennsylvanicum* had the highest number of PSEP-encoding genes lost, it had the lowest number of PSEP-encoding genes gained among the four smut genomes.

Distribution of pseudogenes

Recently developed pseudogenes were predicted by first extracting the intergenic sequences from all the four species and then searching a local protein database containing the predicted genes from all for species with tBlastn and Exonerate (Slater and Birney, 2005) (See Material and Methods).

This approach led to the discovery of new genes with intact open reading frames that were not picked up by previous annotations. In total, 14 putative new genes were found in *M. pennsylvanicum*, 55 in *U. hordei*, 3 in *S. reilianum*, and 23 in *U. maydis*. Using tBlastn no pseudogene was found in *M. pennsylvanicum*, 142 pseudogenes were found in *U. hordei* and 6 in *S. reilianum*, and 2 in *Ustilago maydis*. Using Exonerate 3, 160, 2, and 9 pseudogenes were observed in the genomes of *M. pennsylvanicum*, *U. hordei*, *S. reilianum* and *Ustilago maydis*, respectively. It should be noted, however, that the approaches only detect recently developed pseudogenes as a measurement of gene turnover and not for identifying genes deteriorated as a result of adaptation to the specific hosts.

Divergence of four smut species

To infer the phylogenetic relationships among the four smut fungal genomes, RAxML (Stamatakis, 2006; Stamatakis et al., 2005) was run with 1000 bootstrap iterations on multiple sequence alignments of the 1:1 orthologs of the four smut fungi and *Malassezia globosa*, which was used as outgroup. A total of 2979 1:1 orthologs were found among the five genomes and subjected to alignments and phylogenetic analysis. Thereby a sister-group relationship of *U. maydis* and *S. reilianum* was found, without significant support. *Ustilago hordei* and *M. pennsylvanicum* were found to group together, with maximum bootstrap support. The genetic distance between the four smut fungi was similar (Figure 3).

Protein subcellular localizations

The ProtComp9 package (www.softberry.com) was used for the identification of protein subcellular localization in the four genomes. In total 1445, 1455, 1582, and 1470 mitochondria targeted proteins were found in the genomes of *M. pennsylvanicum, U. hordei, S. reilianum* and *U. maydis*, respectively (Supplementary Figure 2). Further, 1888 nuclear proteins were predicted in *M. pennsylvanicum*, 2139 in *U. hordei*, 1910 in *S. reilianum* and 1992 in *U. maydis*. In

addition, 1351, 1546, 1271 and 1323 proteins in the genome of *M. pennsylvanicum*, *U. hordei*, *S. reilianum* and *U. maydis*, respectively, were predicted as cytoplasmic.

To check the conservation and positive selection on the genes coding for proteins targeted to different cellular organelles, all 1:1 orthologs were compared and their percentage identity was calculated using BlastP. Positively selected genes and dN/dS ratio information was inferred using the codeml Branch-Site model of PAML at a 1% level of significance of false discovery rate (FDR) with Bayes Empirical Bayes (BEB) >95%. These analyses revealed the lowest sequence conservation and highest proportion of genes under positive selection in the PSEP-encoding genes, while peroxisome-targeted genes were most conserved (Figure 4).

Patterns of positive selection among four smut species

To detect the genes under selection pressure, the codeml program of PAML was used on all 1:1 orthologs within the four smut genomes. In further tests, multiple hypothesis testing was done using the Bonferroni corrections (BC) and false discovery rate (FDR) tests at 5%, 1% and 0.1% level of significance. All the three methods of hypothesis testing generated the highest percentage and proportion of positively selected genes in the genome of *M. pennsylvanicum* compared to other three species genes (Figure 5A). Detailed predictions with a FDR at 1% level of significance and considering only genes with at least one positively selected site with more than >95% BEB confidence revealed that *M. pennsylvanicum* has by far the highest percentage of genes under positive selection (Figure 5B).

Positively selected putative secreted effectors and non-secreted protein encoding genes

To compare the percentage of PSEPs and non-secreted protein encoding genes positively selected with a high level of confidence, only those genes were used that had >95% Bayes Empirical Bayes (BEB) support and a false discovery rate (FDR) at 1% level of significance.

Notably, when the number of positively selected genes was compared, the percentage of genes encoding PSEPs was comparatively higher (Figure 5B). It was revealed that 18.47% of the non-secreted protein encoding genes and 22.00% secreted protein encoding genes of *M. pennsylvanicum* were positively selected. *Ustilago hordei* showed 9.66% and 13.79%, *S. reilianum* 6.48% and 8.35%, and *U. maydis* 3.08% and 8.17% of positively selected non-secreted and secreted protein encoding genes, respectively.

We further assessed the percentage of positively selected sites among the tested genes that passed the BEB >95% confidence threshold at a FDR <1%. A higher percentage of selected sites was found in the PSEP-encoding genes (Figure 5C), compared to the non-effector genes in all four genomes.

Candidate pathogenicity clusters

Using the "three direct neighbor" (TDN) approach (See methods) 23 new candidate pathogenicity clusters were defined in *Ustilago maydis*, the 12 clusters published already were also retrieved. For the other species this method generated 37 candidate pathogenicity clusters in *S. reilianum*, 19 in *U. hordei* and 17 in *M. pennsylvanicum*.

A sliding window of 3 kb (Supplementary Figure 3) was found to generate a similar amount of candidate pathogenicity clusters when compared with the TDN method. By this method, a total of 22, 27, 53 and 43 candidate clusters where found in *M. pennsylvanicum*, *U. hordei*, *S. reilianum* and *U. maydis* respectively. By combining the output of these two methods, a total of 24, 29, 55 and 46 candidate clusters where found in *M. pennsylvanicum*, *U. hordei*, *S. reilianum* and *U. maydis* respectively (Table 1). Supplementary Table 2 lists the orthologs of the *U. maydis* PSEPs-encoding genes in the previously reported (Kämper et al., 2006) 12 *U. maydis* pathogenicity clusters. The novel 34 candidate clusters of *U. maydis* with their ortholog information are in Supplementary Table 3, orthologs of the PSEP-encoding genes in the

candidate pathogenicity clusters of *S. reilianum*, *U. hordei* and *M. pennsylvanicum* are listed in the Supplementary Tables 4, 5, and 6, respectively.

Pathogenicity cluster conservation

While checking the conservation of candidate pathogenicity clusters within the species (Table 2), a high degree of conservation of candidate pathogenicity clusters among the graminicolous species was observed, conservation was lower in *M. pennsylvanicum*. This is also apparent for the cluster 19A of *U. maydis*, where the absence of several putative secreted effectors in *M. pennsylvanicum* could be observed, as well as the proliferation of genes encoding PSEPs in *U. maydis* and *S. reilianum* (Figure 6).

Genome architecture comparisons

To investigate the compactness of four genomes with respect to the gene distributions, the length of the 5' and 3'gene flanking regions were computed for all genes of the four genomes. It was revealed that all genomes had a similar degree of compactness, with average intergenic distances ranging from 921.53 bp and 920.38 bp for the 5' and 3' distance, respectively, in *S. reilianum* to 1060.17 bp and 1064.38 bp for the 5' and 3' distance, respectively, in *M. pennsylvanicum* (Figure 7 A-D)**.**

Distances to neighboring genes for the genes encoding PSEPs were also assessed (Figure 7 E-H)**.** These results showed that the mean of 5' and 3' distances of PSEP-encoding genes ranged from 1004.62 bp and 1060.11 bp, respectively, in *S. reilianum* to 1387.74 bp and 1360.53 bp, respectively in *M. pennsylvanicum*. PSEP-encoding genes thus on average reside in slightly more gene-sparse regions of the smut genomes.

Also the frequency of the genes in relation to the average lengths of the 5' and 3' end flanking region for all genes (Supplementary Figure 4 A-D) and PSEP-encoding genes (Supplementary Figure 4 E-H) revealed a similar pattern.

Analyses regarding the enrichment of PSEP-encoding genes in clusters revealed that there were 46 cases, where a PSEP-encoding gene was next to another PSEP-encoding gene in the *M. pennsylvanicum* genome, 59, 154, and 158 occurrences were observed, in the genomes of *U. hordei*, *S. reilianum*, and *U. maydis*, respectively. Combinatorial expectancies of these occurrences were 28, 42, 60, and 58, in the same order as above. It was thus revealed that, while most PSEP-encoding genes are not residing in clusters, there is a trend towards the clustering of these genes, which is especially pronounced in *U. maydis* and *S. reilianum*.

Discussion

Genome assembly and gene calls

In this study, Illumina sequencing was used to generate the genome sequence of *Melanopsichium pennsylvanicum,* a non-obligate biotrophic pathogen, from the family Ustilaginaceae. *Melanopsichium pennsylvanicum* is responsible for gall smut on *Polygonum pennsylvanicum* (Barbe 1962) and is thus unusual among the Ustilaginaceae in having a dicot host, with its closest relatives being pathogens of the monocot Poaceae (Begerow et al., 2004; McTaggart et al., 2012; Stoll et al., 2005; Weiß M, 2004). Extensive comparative studies have been done previously on the genomes of *Ustilago maydis*, *S. reilianum*, and *U. hordei* to shed light on their pathogenic behavior, evolution and genomic makeups (Kämper et al., 2006; Laurie et al., 2012; Schirawski et al., 2010). However, comparative studies with the aim to identify genes that might be required for pathogenicity on grasses or required for pathogenicity in general where not feasible to date, as all of these species parasitize hosts in Poaceae. To shed light on this topic, which is crucial for understanding the pathogenicity of smuts in particular and biotrophic plant pathogens in general, a comparative genomics approach was taken in this study involving the three smut genomes published so far and the genome of the dicot-infecting *M. pennsylvanicum*.

Genome assemblies of *M. pennsylvanicum* generated 434 scaffolds for the nuclear genome of 19.15 Mb and one scaffold for the mitochondrial genome of 74.91 kb, which is comparable to the assembled genome published for *U. hordei* (Laurie et al., 2012). The assembled genome completeness with respect to the core eukaryotic genes was almost identical for all genomes, including *M. pennsylvanicum*,. It can thus be concluded that although the genome architecture is much better resolved in *U. maydis* and *S. reilianum* (Kämper et al. 2006;

Schirawski et al. 2010), in comparison to *U. hordei* (Laurie et al., 2012) and *M. pennsylvanicum*, the gene space is equally covered in all four genomes.

The genome of *M. pennsylvanicum* encodes 6280 genes, which is substantially less than in the other Ustilaginaceae sequenced to date, as *Ustilago hordei* has 7111 protein coding genes, while *S. reilianum* has 6673 and *U. maydis* 6787 protein coding genes. To cross-check this noticeable result, despite the evidence for good gene-calling in all four genomes as inferred from the CEGMA analyses, tBlastn searches were carried out to ensure that not a significant number of genes were missed in any of the species. While a few additional potential genes were found in all four genomes, their amount was comparable and generally low. The lower amount of genes in *M. pennsylvanicum* might be the result of the host jump to a dicot plant, as many genes that are related to colonizing grass hosts will not produce effector proteins that match the divergent targets in the new host environment and are thus prone to be lost. As expected, especially the PSEPs seem to be affected. While *M. pennsylvanicum* contains 7.5% less non-secreted proteins than *U. maydis*, it harbors 33.6 % fewer PSEPs. When considering the average of the three published smut genomes of *U. maydis*, *S. reilianum*, and *U. hordei*, *M. pennsylvanicum* has only 8.4 % less non-secreted proteins, while it contains 30.6 % fewer PSEPs.

Non-secreted protein encoding genes also regulate the expression of effector genes and other processes in the infection process. Thus, the loss of several additional genes from the nuclear genome might also be associated with this huge host jump. The morphology and disease symptoms are highly different from the other three species, which led to the description of the genus *Melanopsichium* for the smut pathogens of Polygonaceae. However, phylogenetic investigations have shown that *M. pennsylvanicum* is embedded within the *Ustilago* s.l. clades (Begerow et al., 2006; McTaggart et al., 2012; Stoll et al., 2005). As the regulation of the

morphology and disease symptoms of smut fungi is still poorly understood, it remains uncertain, if the gene losses observed in the genome also contribute to the differences in morphology.

Phylogenetic position

Phylogenetic studies in the Ustilaginales, on the basis of the internal transcribed spacer (ITS) and RNA sequences, have already been reported (Begerow et al., 2006; Suh and Sugiyama, 1993), but generally exhibited a low resolution on the backbone of *Ustilago* in the broad sense. The phylogenetic relatedness of the four genomes was assessed on the basis of all orthologous nuclear genes found in the smut fungi and the more distantly related ustilaginomycete *Malassezia globosa*. Even though a sister-group relationship of *U. maydis* and *S. reilianum* was not supported, *U. hordei* and *M. pennsylvanicum* were grouped together with maximum support, confirming the nestedness of the latter in *Ustilago* in the broad sense. Thus, although there are some morphological differences in some lineages of the *Ustilago* s.l. complex, it could be a practical solution to merge some of the segregate genera again with *Ustilago*, also to enable the continued use of the genus name *Ustilago* for *U. maydis*.

Positively selected genes and sites

Positive selection or natural selection within genes is the main evolutionary event for adaptation and has thus been an important focus of several comparative genomics studies (Haas et al., 2009; Kemen and Jones, 2012; Schirawski et al., 2010). For efficient colonization of a new host, it can be expected that most of the effectors that are still able to operate a target in the new host have to adapt to the very different host environment and thus should show relatively strong signatures of positive selection. It has been observed that the genes encoding PSEPs are generally under stronger selection pressure than the genes encoding non-effector proteins (Kemen et al., 2011; Schirawski et al., 2010). These PSEP-encoding genes are believed to be under high selection pressure due to their highly evolving counterparts (resistance genes) in the host, but

after host jumps, the adaptation to new targets arguably is the most important driver of positive selection. Supporting this hypothesis, *M. pennsylvanicum* showed the highest percentage of PSEPs under positive selection, 59.5 to more than two-fold higher than in the other genomes at >95 % BEB (Bayes Empirical Bayes) support and <1% FDR (False Discovery Rate).

Patterns associated with the adaptation to a new host and a reappraisal of the pathogenicity cluster concept

Host jump events are expected to be associated with several changes in the genome of the pathogen, like genome rearrangements, positive selection, gene losses and gene gains. Although some comparative genomics studies (Kemen et al., 2011; Laura Baxter, 2010; Raffaele et al., 2010) and a few experimental works (Baxter et al. 2010; Dong, et al. 2014) have previously been done in hemibiotrophic oomycetes, no detailed investigations on the effects of host jumps of biotrophic oomycete or fungal pathogens to largely unrelated hosts have been carried out so far. Thus, it is unclear, which of the events outlined above, apart from natural selection in the sense of the evolutionary theory of Wallace (Wallace, 1858) and Darwin (Darwin, 1858), are the major factors in the adaptation to divergent hosts. To estimate the amount of genes gained or lost in the four genomes, all the orthologous genes were checked and scanned whether they are present or absent in other genomes. Interestingly, the *M. pennsylvanicum* genome has lost 292 genes, which are present in the other three species, while only 99, 17, and 47 genes were not present in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively. In terms of the PSEP-encoding genes, 57 were absent in *M. pennsylvanicum* genome, while 17, 3, 2 were only absent in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively. Of 44, 13, 2, and 1 of these genes, respectively, no distantly similar genes or gene remnants could be found in the respective other genomes. These results suggest that *M. pennsylvanicum* has lost a higher proportion of genes involved in the interaction with the plant host, most likely those that did not match a target after the host jump and were thus no

longer required. In contrast, only 324 genes of *M. pennsylvanicum* did not show a hit in the other three genomes, but 510, 335, and 422 of such genes were observed only in *U. hordei*, *S. reilianum* and *U. maydis*, respectively, highlighting that gene loss, rather than gene gain was the hallmark of adaptation to the dicot host. The genes encoding RNA interference and DNA remodeling components, which have been reported to be absent within *U. maydis* (Laurie et al., 2012) were present in the *M. pennsylvanicum* genome.

Host jump events might show some effects on the genome architecture of the species. Probably as a result of gene losses, genes of *M. pennsylvanicum* are more distinctly apart from each other, when compared to the other three genomes. However, no pronounced genome expansion or genome rearrangements could be found in whole genome alignments, which contrasts studies on other fungal pathogens, were these processes were reported to play important roles (Ma et al., 2010).

Most of the PSEP-encoding genes are reported to be in clusters within the genomes of the Ustilaginaceae (Kämper et al. 2006; Schirawski, et al. 2010). Thus, the clustering of effectors among the four genomes and the effect of the host jumps on the pathogenicity clusters were analyzed in detail. In the genome of *U. maydis*, 12 clusters were already defined according to the continuity of the genes encoding PSEPs (Kämper et al., 2006). A limitation of this method is that it skips clusters that have 3 or more effector genes in very close vicinity, but with non-effector genes interspersed. To overcome this, we used a window-based method with a window size of 3 kb, and clusters were defined as such if at least three effector genes appear in three consecutive 3-kb windows. This method picked 98% of the clusters defined by the previous approach and revealed some more potential pathogenicity clusters. Combining the output of both of these methods, 24, 29, 55 and 46 candidate pathogenicity clusters were defined in the genomes of *M. pennsylvanicum, U. hordei, S. reilianum*, and *U. maydis* respectively. While the other three

genomes showed a conservation of clusters in the range of 43-80%, conservation ranged from 20-26% in *M. pennsylvanicum*. This highlights that the genes in pathogenicity clusters are also strongly affected by the host jump. But the fact that only 12.2% to 31.6% of the genes encoding PSEPs are clustered suggests that clustering might not be a key event for pathogenicity development in Ustilaginaceae in general. In fact, the most well-known effector of *U. maydis*, PEP1, which is also conserved in *M. pennsylvanicum*, is encoded by a gene that is not embedded within a pathogenicity cluster. However, there seems to be some trend towards the clustering of effectors, as at minimum 1.4 times more PSEP-encoding genes were observed to be next to another PSEP-encoding gene than expected by chance in *U. hordei*, while *U. maydis* contained more than 2.7 fold more occurrences of PSEP-encoding genes next to another PSEP than expected by chance. While initially useful for the elucidation of some general pathogenicity features of the Ustilaginaceae (Kämper et al., 2006; Laurie et al., 2012; Schirawski et al., 2010) it thus seems reasonable to focus more on the non-clustered putative secreted effector genes in future studies.

Of the genes encoding PSEPs, 57 (Supplementary Table 7) were found in the three graminicolous species but not in the dicot-infecting *M. pennsylvanicum* For 44 of these, no distantly similar gene remnants were found in intergenic regions *M. pennsylvanicum*. It seems possible that these PSEPs contain pathogenicity effectors that are of particular importance for the colonization of grass hosts, while those 248 orthologous genes (Supplementary Table 8) encoding PSEPs that are present in all four species might contain those that are vital for pathogenicity in general, possibly targeting key hubs in plant defense pathways. Future functional studies that take advantage of the findings presented here might result in a better understanding of the evolution of pathogenicity in the Ustilaginaceae in particular and in biotrophic plant pathogens in general.

Material and Methods

DNA isolation and sample preparation

DNA was isolated from yeasts harvested from PDA medium using a phenol-chloroform extraction method as described in (Ploch et al., 2011).

Data pre-processing

Illumina reads of 76bp read length and 300bp insert size derived from GAII sequencers were used. In the data filtering steps, Illumina adapter and primers were trimmed, reads that were having N's were filtered out along with their pairs. In the final step of data processing, all reads with average base quality score less than 26 were excluded along with their pairs.

Genome assembly and scaffolding

Initially Velvet (Zerbino and Birney, 2008) was run on the two lanes for *k*-mers 21, 31, 41, 45, 49, 51, 55 and 61 for calculating the average insert size and insert size standard deviation of the two Illumina lanes. Reads from both of the lanes were mapped back on the assemblies using Bowtie2 (Langmead and Salzberg, 2012) with input insert size within the limit of 100-600 bps. The resulting SAM file from Bowtie2 was used to calculate the average and standard deviations in insert sizes of the mapped reads, which were 220 and 20, respectively. After calculating the average insert size and standard deviation, Velveth was run using these values for all odd *k*-mers from 21 to 67. The *k*-mer coverage cutoff for different *k*-mers were calculated using the R statistical package (team, 2008), according to the manual of the Velvet software. These values were in the range of 2 to 15, for all tested *k*-mers. All generated assemblies from different *k*-mers were compared using the following assembly quality parameters: N50 contig size, largest contig size, number of contigs, number of reads used in the assembly, number of reads mapped back to the assembly, assembly completeness as assessed by CEGMA pipeline (Parra et al., 2007) and length of the assembled genome. Assemblies with a *k*-mer length 63 and

k-mer coverage cutoff of 3 generated the best assembly. Scaffolding of the Velvet contigs was performed by SSPACE (Boetzer et al., 2011) scaffolding package. The optimal scaffolded assembly was picked up again considering the above discussed assembly quality parameters.

CEGMA analyses to check the genome completeness

CEGMA (Core Eukaryotic Genes Mapping Approach) is a pipeline to detect the core housekeeping genes of eukaryotes. CEGMA uses the KOGs database (clusters of euKaryotic Orthologous Groups) (Tatusov et al., 2003) to build a set of 458 highly conserved ubiquitous proteins. The CEGMA pipeline was run to compare the completeness and continuity of the four smut genomes on the basis of these proteins according to the manual for all the four smut fungi genomes, with their respective average intron lengths that were calculated before starting the analyses.

Genome comparison

To align the genome of Mp4 to other three smut genomes, the Mummer3 (Kurtz et al., 2004) whole genome alignment tools were used. Mummer3 is a robust tool for comparing whole genomes and to graphically visualize the alignments in the form of dot-plots and maps. To check the similarity of Mp4 genome with the other three genomes at nucleotide level, the nucmer module was used with default arguments. Plots were produced using mummerplots with the delta file generated by promer. The promer module was used to generate alignments at protein level by translating the genome in all six reading frames prior to the alignments. Again plots were produced using mummerplots and the delta file generated by promer.

Data sources

Genomic sequences, general feature format **(**Gff) for gene coordinates, protein and gene/transcript sequence files of *Ustilago maydis, U. hordei, Sporisorium reilianum* and *Malassezia globosa* were downloaded from the MIPS (Munich Information Center for Protein

Sequences) and JGI (The Genome Portal of the Department of Energy Joint Genome Institute) (Grigoriev et al., 2012) databases. The upload dates, ftp sources and references for these genomes are given in the Supplementary Table 9.

Repeat element prediction

Repeat elements were predicted using the package RepeatScout (Price et al., 2005). RepeatScout uses five steps to investigate and mask the repeat elements within the genome. The program '*build_lmer_table'* of RepeatScout was run with the *l*-mer length of 14 bps, which generated the hash table. Low-complexity regions were removed using the '*filter-stage-1.prl*' program the and tandem repeats using the TRF software (Benson, 1999), and the program '*filter-stage-2.prl'* was used to filter out the repeat elements which were not present more than 3 times in different genomic locations. RepeatMasker (A.F.A. Smit, R. Hubley & P. Green RepeatMasker at http://repeatmasker.org) was used to mask the predicted repeats from the above steps. In the final step, the *U. maydis* repeat libraries from the RepBase library (version 20110920) (Jurka et al., 2005; Kohany et al., 2006) were taken for the repeat masking using RepeatMasker.

Gene prediction

Both *ab-initio* and homology based prediction were used for defining the genes within the repeat-masked *M. pennsylvanicum* genome (Supplementary Figure 5). An Exonerate (Slater and Birney, 2005) hint file was generated by mapping the *U. maydis* protein sequences on the assembled Mp4 genome. Augustus (Stanke, et al. 2006) was run by using the generated hint file and input parameters: --strand=both; --genemodel=partial; --extrinsicCfgFile =extrinsic.E.XNT.cfg. Another set of genes was generated using GlimmerHMM (Majoros et al., 2004) according to the manual. The TrainGlimmerHMM module was used to train the GlimmerHMM with the *U. maydis* gene set. In the final step of GlimmerHMM annotations,

GlimmerHMM was run on the generated *U. maydis* training set. A third gene model was generated by GeneMark-hmm-ES (Ter-Hovhannisyan, et al. 2008). For this, the Perl script *gm_es.pl* was used according to the manual.

The three gene models were then fed into the Evigan (Liu et al., 2008) package to predict the consensus gene models. Transfer-RNA genes were predicted using the tRNA-Scan (Lowe and Eddy, 1997; Schattner et al., 2005) according to the user manual.

In the later annotation steps all the intergenic sequences were extracted and aligned against all the protein sequences from three Ustilaginales genomes including the protein sequences of Mp4 generated from above annotations. These alignments were performed by tBlastn (Altschul et al., 1990) and the Exonerate software (Slater and Birney, 2005). Genes found were then added to initially predicted gene models.

Gene annotation

Gene annotations were added on the basis of orthology information from other three annotated genomes. InterProScan (Zdobnov and Apweiler, 2001) was used to assign biological functions, gene ontology and biological pathway information of the predicted genes of Mp4. The InterProScan program searches Interpro (Hunter et al., 2009) database which integrates several other databases: PROSITE (Sigrist et al., 2002), PRINTS (Attwood et al., 1994), Pfam (Sonnhammer et al., 1997), ProDom (Corpet et al., 1998), SMART (Schultz et al., 1998), TIGRFAMs (Haft et al., 2003), PIR superfamily (Barker et al., 1996), SUPERFAMILY (de Lima Morais et al., 2011), Gene3D (Buchan et al., 2002), PANTHER (Mi et al., 2005) and HAMAP (Lima et al., 2009) databases. These searches provide the information regarding the Gene Ontology (GO) (Harris et al., 2004) and KEGG (Kanehisa, 2002) pathways of the predicted genes.

Prediction of putative secreted effectors proteins (PSEPs)

The Expassy toolkit (Gasteiger et al., 2003) was used to generate protein sequences from the predicted genes. The SignalP v4.0 package (Petersen et al., 2011) was used to investigate the proteins that are having extracellular secretion signal. SignalP v4.0 can discriminate signal peptides from transmembrane regions, which makes it highly accurate for secreted protein predictions (Petersen et al., 2011). PSEPs within all the four Ustilaginales genomes were investigated and were compared with the available published data. Another set of candidate secreted proteins was generated by using TargetP v1 (Emanuelsson et al., 2007) for all the four genomes.

Pathogenicity cluster prediction

PSEPs from all the four genomes were defined as organized in pathogenicity clusters, if at least three PSEPs were present in a row. Pathogenicity clusters were further extended if the initially defined cluster has PSEP-encoding genes downstream or upstream to it and two interruptions by non-secreted protein-coding genes were allowed if the next gene was again a PSEP-encoding gene. It is referred to as "three direct neighbor" (TDN) method here. This method had also been used previously for *U. maydis* pathogenicity cluster determination (Kämper et al., 2006). Both TargetP v1.1 (Emanuelsson et al., 2007) and SignalP v4.0 predictions for secreted proteins were used for pathogenicity cluster definition. In another approach of defining pathogenicity clusters, windows of size 3kb were searched for secreted effectors. Regions were defined as pathogenicity clusters, if in three such windows in a row effectors were present. The output of this method was compared to the previous method by estimating the percentage of pathogenicity clusters picked by both of these methods (Supplementary Figure 6).

To check the conservation of the pathogenicity clusters, orthologs were investigated in all four species. The pathogenicity clusters was defined as conserved in all of the four genomes, if at

least one of the secreted proteins of the respective pathogenicity cluster had an ortholog in all four genomes and was observed in a pathogenicity cluster of that particular species.

Prediction of subcellular localization

Proteins subcellular localizations were identified using the ProtComp v9 package (www.softberry.com). ProtComp was locally installed and ran for all four genomes and percentages of proteins localized to certain subcellular components were calculated. Transmembrane domains within the protein sequences were investigated using TMHMM2.0c (Krogh et al., 2001).

Ortholog prediction

For orthologs predictions both orthoMCL (Li et al., 2003) and Inparanoid (Ostlund et al., 2010) were run on the protein sequences of all four genomes. Both of these tools generated orthologs and paralogs within and among the four genomes. After generating the list of orthologs and paralogs, perl and shell scripts were used for further downstream analysis of the generated orthologs information.

The percentage of identity of the 1:1 orthologous within the four smut genomes were calculated using the BlastP searches. A circular plot (Figure 4) of the aligned sequences was generated using the Circos package (Krzywinski et al., 2009).

Gene gain and gene loss

To investigate the genes lost or gained in the smut genomes during evolution, the ortholog information generated by the methods described above was used. Orthologs were considered to be absent in one genome, if the orthologous were present in all of the three genomes and not predicted in the genome under consideration. Similarly a gene was said to be a species-specific gene, if it was only found in the species under consideration, but no orthologs were found in the

other three genomes. Gene presence and absence was further tested with tBlastn and Exonerate searches, by compiling a database of all proteins from the four genomes and performing tBlastn searches using a percentage of identity cutoff of 55% against the intergenic regions of all the four genomes separately. More relaxed searches were further done by using 35% identity cutoff and if alignment length was at least 30% of the query protein.

To perform more stringent search of gene gain, BlastP searches were done on the proteins which did not show any orthologous in other three species. For these BlastP searches only those hits were considered which showed alignment length more than 35% of the query protein length, e-value less than e^{-2} and more than 35% percentage identity. To further confirm gene losses, local BlastP searches of the protein sequences that were present in all three genomes but absent in one were performed. For these BlastP searches, a very relaxed search string was used, with e-value cutoff as e^{-1} and a minimum percentage identity of 35%. Genes were only considered to be fully absent (lost), if they failed to return hits with this search strategy.

Genome architecture comparison

To compare the genome architecture of all four species, 5' and 3' distances until the next gene were calculated for all genes for all four genomes. Heat-maps for all the four genomes were generated using the ggplot2 (http://ggplot2.org/) R-package after calculating the distances. Heat-maps of the four genomes were then further analyzed to infer the compactness of gene coding regions within the genomes.

In another approach, the average 5' and 3' flanking distances of all genes were computed and compared with the average 5' and 3' flanking distances of PSEP-encoding genes. The same analysis was done by using the smaller and greater lengths among 5' and 3' flanking distances of all genes and PSEP-encoding genes of all the four genomes.

Phylogeny

To perform phylogenetic analysis on all orthologous genes in the four genomes and to produce an unrooted tree of the four smut species for positive selection studies, the predicted 1:1 orthologs inferred by OrthoMCL were used. Mafft (Katoh et al., 2002) with the *G-INS-i* algorithm (global alignments) was used to generate the alignments of all 5200 1:1 orthologous genes of four species. Alignments were used as input for RAxML (Stamatakis, 2006; Stamatakis et al., 2005) for Maximum Likelihood phylogenetic inference. RAxML was run with 1000 bootstrap replicates using the GTRGAMMA model.

In another analysis, to produce a rooted tree, the genome of *Malassezia globosa*, which is a human skin pathogen was used as outgroup, while all other steps were done as described above.

Pseudogene discovery

To investigate both processed and unprocessed pseudogenes, first a database of all proteins from all four genomes was created. To align the protein sequences on the intergenic positions, all the intergenic sequences from four repeat-masked genomes were extracted. For this, two approaches were implemented, in the first approach tBlastn was used with the '-*max_intron_length*=5000' option. After obtaining the alignment from the standalone tBlastn, those alignments were kept for further analysis, which were having a percentage of identity greater than 65%, an alignment length greater than 70% of the parent protein length and an e-value less than e^{-10}. From these blast hits only those were further analyzed that were starting from the first position of the query protein and were having at least one premature stop codon inside the alignment compared to the parent protein.

In another approach Exonerate was used to map the proteins from all of the four genomes on the intergenic sequences. Exonerate was run by using the following input parameters: --model=protein2genome; maxintron=5000; bestn=20; percent=55; score=100. The output from Exonerate for all four genomes was again interrogated for the start codon of any predicted gene

structure. Pseudogenes were defined if the predicted gene structure was starting from the 1st position of the query protein and had at least one premature stop codon. These methods also predicted some new genes, which were overlooked by the previous annotation methods.

Positive selection inference

The Prank-codon alignment module of Prank (Loytynoja and Goldman, 2010) was used for multiple sequence alignments of all 1:1 orthologs within the four genomes. Prank-codon has performed best in comparison to other multiple alignment tools (Fletcher and Yang, 2010), like Mafft (Katoh et al., 2002), Muscle (Edgar, 2004), ClustalW (Thompson et al., 2002), and prank-aminoacid was used for obtaining input files for downstream positive selection studies. The CodeML module of the PAML package V4.6 (Yang, 2007) was used for predicting positively selected genes within the four genomes. According to the PAML manual, test2 is the highly recommended test for a branch-site model and was used accordingly. Parameters for the branch-site model were as follows – H_0 control files were generated by assigning the parameters model=2, NSsite=2, fix_omega=1 and omega=1. In the alternate hypothesis H_1 files were generated using the parameters model=2, NSsite=2, fix_omega=0 and omega=1. A RAxML tree for the four genomes was used as the input tree file for CodeML. After obtaining the output for the H_0 and H_1 hypotheses, a Likelihood Ratio Test (LRT) (Anisimova et al., 2001; Yang and Nielsen, 2002; Zhang et al., 2005) was performed to compare both null and alternate hypotheses. The testing of the hypotheses was done with a χ^2 distribution at 5%, 1% and 0.1% level of significance. P-values were calculated by using the Statistics::Distributions (http://search.cpan.org/~mikek/Statistics-Distributions-1.02/Distributions.pm) perl module.

To perform multiple hypothesis testing, Bonferroni Corrections (BC) (Anisimova and Yang, 2007) and false discovery rate (FDR) tests were used. FDR inference was performed by using the

Q-value R package (Storey, 2002). Both of these tests were performed at 5%, 1% and 0.1% levels of significance.

Positively selected sites were detected by using Naïve Empirical Bayes (NEB) (Nielsen and Yang, 1998; Yang, 2000; Yang and Bielawski, 2000) and the Bayes Empirical Bayes (BEB) (Yang et al., 2005) information from the CodeML output files. Only those genes were considered under positive selection, which had at least one site under selection with >95% BEB confidence at <1% FDR.

Data Access

All the Illumina short reads used in this study have been submitted to the European Nucleotide Archive (ENA) database (Study accession number: PRJEB4565). The assembled genome and annotations of the *Melanopsichium pennsylvanicum* genome have been submitted to the European Nucleotide Archive (ENA) and can be accessed from accession ids HG529494 to HG529928. Mp4 Genome scaffolds, protein sequences and gff file are available at http://dx.doi.org/10.12761/SGN.2014.3.

Acknowledgements

We thank Claus Weiland for support with respect to cluster access. We would also like to thank Jodie Pike for handling the illumina sequencing and for creating libraries. This work was supported by the Max-Planck Society through a fellowship awarded to MT, and the research funding program LOEWE "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts.

Disclosure declaration

The authors declare that no competing interest exists.

Author contributions

MT designed the study. BM, MT, and RS conceived analyses and provided ideas. RS assembled and annotated the genome and carried out all computational analyses on the data. FR handled the Mp4 strain and isolated genomic DNA for sequencing. RS and MT wrote the manuscript, with contributions from BM and FR.

Figure Legends

Figure 1. Genome assembly quality and completeness estimation. (A) Genome assembly quality as assessed by calculating the number of scaffolds and minimum scaffold length in the respective N-class, where N is the percentage of the genome covered after sorting the scaffolds from largest to smallest. **(B)** Genome completeness assessed by CEGMA analysis on the basis of 458 euKaryotic clusters of Orthologous Groups (KOGs). The CEGMA pipeline categorizes these 458 core proteins in four groups on the basis of their conservation in the eukaryotic genomes. Dotted and solid lines are representing figures when considering partial mapping and complete mapping of the KOGs, respectively.

Figure 2. Orthologous genes within the four smut genomes. (A) Venn diagram representing the orthologous genes within the four genomes. **(B)** Orthology of the putative secreted effector proteins from all four genomes.

Figure 3. Phylogenetic distance estimation using all nuclear genes 1:1 orthologs from five species. Maximum likelihood inference based on MAFFT alignments using RAxML with 1000 bootstrap replicates. Numbers at branches indicate bootstrap support percentages for the respective branches. Red and green arrows represent the number of gene lost and gained, respectively, number in brackets represents the number of gene lost or gained in terms of PSEP-encoding genes. Numbers in bold represent gene losses/gains in the four genomes without including gene remnants in intergenic regions. Genes losses/gains were further tested for gene remnants or distantly similar genes using lower cutoffs, the corresponding figures are given in italics. Genome sizes refer to assembled genome sizes.

Figure 4. Conservation of proteins in smut genomes and their dN/dS ratios according to their subcellular localizations. The outmost ring 'A' represents *M. pennsylvanicum* protein sequences according to their subcellular localization. Ring 'B' represents the dN/dS ratios of the proteins of *M. pennsylvanicum* shown in Ring 'A'. Red and green bars represent the dN/dS ratios of the positively selected (1% FDR, >95% BEB confidence) and non-selected (Non-significant considering 1% FDR) genes, respectively. Similarly the rings 'D', 'F' and 'H' represent the dN/dS ratios of *S. reilianum*, *U. maydis* and *U. hordei*, respectively. Rings 'C', 'E' and 'G' represent the BlastP percentage identity of *M. pennsylvanicum* proteins with the *S. reilianum*, *U. maydis* and *U. hordei* proteins, respectively. Green dots highlight a BlastP identity greater than 85%, blue dots represent an identity in the range of 65-85%, red dots in the range of 50-65% and black dots are highlighting a BlastP percentage identity less than 50%.

Figure 5. Percentage of positive selection among the four genomes and comparisons of positively selected genes encoding PSEPs and non-secreted proteins. (A) Percentage of positively selected genes among the four species. **(B)** Percentage of positively selected PSEP-encoding and non-secreted protein encoding genes. **(C)** Percentage of positively selected sites within the four genomes.

Figure 6. *Ustilago maydis* 19A pathogenicity cluster synteny in the other three genomes. Grey shading represents orthologous genes. Yellow arrows represent PSEPs and green arrows represent the non-secreted protein encoding genes. The orientation of the arrows represents the orientation of the genes and length of arrows is proportional to the length of the gene.

Figure 7. Heat-maps representing the 3' and 5' end flanking region lengths. (A-D) Heat-maps of 3' and 5' end flanking regions of all genes of *M. pennsylvanicum, U. maydis, S. reilianum,* and *U. hordei,* respectively. **(E-H)** Heat-maps of 3' and 5' end flanking regions of PSEP-encoding genes of *M. pennsylvanicum, U. maydis, S. reilianum,* and *U. hordei,* respectively. The 3' and 5' distance values less than 3kb are represented in the portion of the plot shaded grey.

Table Legends

Table 1. Number of candidate pathogenicity clusters predicted by the ‘three direct neighbors’ and the ‘3kb Distance’ approaches.

Table 2. Conservation of candidate pathogenicity clusters within the four genomes

References

Altschul, S. F., W. Gish, W. Miller, E. W. Myers and D. J. Lipman 1990. Basic local alignment search tool. J Mol Biol 215: 403-410. doi: 10.1016/S0022-2836(05)80360-2

Anisimova, M., J. P. Bielawski and Z. Yang 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. Mol Biol Evol 18: 1585-1592.

Anisimova, M. and Z. Yang 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol 24: 1219-1228. doi: 10.1093/molbev/msm042

Attwood, T. K., M. E. Beck, A. J. Bleasby and D. J. Parry-Smith 1994. PRINTS--a database of protein motif fingerprints. Nucleic Acids Res 22: 3590-3596.

Barbe, P. M. Halisky and G. D. 1962. A Study of Melanopsichium pennsylvanicum Causing Gall Smut on Polygonum. Bulletin of the Torrey Botanical Club, Torrey Botanical Society 89: 181-186.

Barker, W. C., F. Pfeiffer and D. G. George 1996. Superfamily classification in PIR-International Protein Sequence Database. Methods Enzymol 266: 59-71.

Begerow, D., R. Bauer and T. Boekhout 2000. Phylogenetic placements of ustilaginomycetous anamorphs as deduced from nuclear LSU rDNA sequences. Mycol Res 104: 53-60. doi: Doi 10.1017/S0953756299001161

Begerow, D., M. Göker, M. Lutz and M. Stoll 2004. *On the evolution of smut fungi on their hosts.*: Frontiers in Basidiomycote mycology

Begerow, D., M. Stoll and R. Bauer 2006. A phylogenetic hypothesis of Ustilaginomycotina based on multiple gene analyses and morphological data. Mycologia 98: 906-916.

Benson, G. 1999. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27: 573-580.

Birch, P. R., P. C. Boevink, E. M. Gilroy, I. Hein, L. Pritchard and S. C. Whisson 2008. Oomycete RXLR effectors: delivery, functional redundancy and durable disease resistance. Curr Opin Plant Biol 11: 373-379. doi: 10.1016/j.pbi.2008.04.005

Boetzer, M., C. V. Henkel, H. J. Jansen, D. Butler and W. Pirovano 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27: 578-579. doi: 10.1093/bioinformatics/btq683

Buchan, D. W., A. J. Shepherd, D. Lee, F. M. Pearl, S. C. Rison, J. M. Thornton and C. A. Orengo 2002. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res 12: 503-514. doi: 10.1101/gr.213802

Corpet, F., J. Gouzy and D. Kahn 1998. The ProDom database of protein domain families. Nucleic Acids Res 26: 323-326.

Darwin, C. R., A. R. Wallace. 1858. On the tendency of species to form varieties; and on the perpetuation of varieties and species by natural means of selection. Journal of the Proceedings of the Linnean Society of London. : 45-50.

de Lima Morais, D. A., H. Fang, O. J. Rackham, D. Wilson, R. Pethica, C. Chothia and J. Gough 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res 39: D427-434. doi: 10.1093/nar/gkq1130

Djamei, A., K. Schipper, F. Rabe, A. Ghosh, V. Vincon, J. Kahnt, S. Osorio, T. Tohge, A. R. Fernie, I. Feussner, K. Feussner, P. Meinicke, Y. D. Stierhof, H. Schwarz, B. Macek, M. Mann and R. Kahmann 2011. Metabolic priming by a secreted fungal effector. Nature 478: 395-398. doi: 10.1038/nature10454

Doehlemann, G., K. van der Linde, D. Assmann, D. Schwammbach, A. Hof, A. Mohanty, D. Jackson and R. Kahmann 2009. Pep1, a secreted effector protein of Ustilago maydis, is required for successful invasion of plant cells. PLoS Pathog 5: e1000290. doi: 10.1371/journal.ppat.1000290

Edgar, R. C. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32: 1792-1797. doi: 10.1093/nar/gkh340

Emanuelsson, O., S. Brunak, G. von Heijne and H. Nielsen 2007. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2: 953-971. doi: 10.1038/nprot.2007.131

Fischer, George William 1953. *Manual of the North American smut fungi.* New York, N. Y.: Ronald Press Co. .

Fletcher, W. and Z. Yang 2010. The effect of insertions, deletions, and alignment errors on the branch-site test of positive selection. Mol Biol Evol 27: 2257-2267. doi: 10.1093/molbev/msq115

Gasteiger, E., A. Gattiker, C. Hoogland, I. Ivanyi, R. D. Appel and A. Bairoch 2003. ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784-3788.

Grigoriev, I. V., H. Nordberg, I. Shabalov, A. Aerts, M. Cantor, D. Goodstein, A. Kuo, S. Minovitsky, R. Nikitin, R. A. Ohm, R. Otillar, A. Poliakov, I. Ratnere, R. Riley, T. Smirnova, D. Rokhsar and I. Dubchak 2012. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res 40: D26-32. doi: 10.1093/nar/gkr947

Haas, B. J., S. Kamoun, M. C. Zody, R. H. Jiang, R. E. Handsaker, L. M. Cano, M. Grabherr, C. D. Kodira, S. Raffaele, T. Torto-Alalibo, T. O. Bozkurt, A. M. Ah-Fong, L. Alvarado, V. L. Anderson, M. R. Armstrong, A. Avrova, L. Baxter, J. Beynon, P. C. Boevink, S. R. Bollmann, J. I. Bos, V. Bulone, G. Cai, C. Cakir, J. C. Carrington, M. Chawner, L. Conti, S. Costanzo, R. Ewan, N. Fahlgren, M. A. Fischbach, J. Fugelstad, E. M. Gilroy, S. Gnerre, P. J. Green, L. J. Grenville-Briggs, J. Griffith, N. J. Grunwald, K. Horn, N. R. Horner, C. H. Hu, E. Huitema, D. H. Jeong, A. M. Jones, J. D. Jones, R. W. Jones, E. K. Karlsson, S. G. Kunjeti, K. Lamour, Z. Liu, L. Ma, D. Maclean, M. C. Chibucos, H. McDonald, J. McWalters, H. J. Meijer, W. Morgan, P. F. Morris, C. A. Munro, K. O'Neill, M. Ospina-Giraldo, A. Pinzon, L. Pritchard, B. Ramsahoye, Q. Ren, S. Restrepo, S. Roy, A. Sadanandom, A. Savidor, S. Schornack, D. C. Schwartz, U. D. Schumann, B. Schwessinger, L. Seyer, T. Sharpe, C. Silvar, J. Song, D. J. Studholme, S. Sykes, M. Thines, P. J. van de Vondervoort, V. Phuntumart, S. Wawra, R. Weide, J. Win, C. Young, S. Zhou, W. Fry, B. C. Meyers, P. van West, J. Ristaino, F. Govers, P. R. Birch, S. C. Whisson, H. S. Judelson and C. Nusbaum 2009. Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 461: 393-398. doi: 10.1038/nature08358

Haft, D. H., J. D. Selengut and O. White 2003. The TIGRFAMs database of protein families. Nucleic Acids Res 31: 371-373.

Harris, M. A., J. Clark, A. Ireland, J. Lomax, M. Ashburner, R. Foulger, K. Eilbeck, S. Lewis, B. Marshall, C. Mungall, J. Richter, G. M. Rubin, J. A. Blake, C. Bult, M. Dolan, H. Drabkin, J. T. Eppig, D. P. Hill, L. Ni, M. Ringwald, R. Balakrishnan, J. M. Cherry, K. R. Christie, M. C. Costanzo, S. S. Dwight, S. Engel, D. G. Fisk, J. E. Hirschman, E. L. Hong, R. S. Nash, A. Sethuraman, C. L. Theesfeld, D. Botstein, K. Dolinski, B. Feierbach, T. Berardini, S. Mundodi, S. Y. Rhee, R. Apweiler, D. Barrell, E. Camon, E. Dimmer, V. Lee, R. Chisholm, P. Gaudet, W. Kibbe, R. Kishore, E. M. Schwarz, P. Sternberg, M. Gwinn, L. Hannick, J. Wortman, M. Berriman, V. Wood, N. de la Cruz, P. Tonellato, P. Jaiswal, T. Seigfried, R. White and Consortium Gene Ontology 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: D258-261. doi: 10.1093/nar/gkh036

Hirschhorn, Elisa 1941. Una nueva especie de Melanopsichium. Notas des Museo de la Plata VI: 147-151.

Hunter, S., R. Apweiler, T. K. Attwood, A. Bairoch, A. Bateman, D. Binns, P. Bork, U. Das, L. Daugherty, L. Duquenne, R. D. Finn, J. Gough, D. Haft, N. Hulo, D. Kahn, E. Kelly, A. Laugraud, I. Letunic, D. Lonsdale, R. Lopez, M. Madera, J. Maslen, C. McAnulla, J. McDowall, J. Mistry, A. Mitchell, N. Mulder, D. Natale, C. Orengo, A. F. Quinn, J. D. Selengut, C. J. Sigrist, M. Thimma, P. D. Thomas, F. Valentin, D. Wilson, C. H. Wu and C. Yeats 2009. InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211-215. doi: 10.1093/nar/gkn785

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res 110: 462-467. doi: 10.1159/000084979

Kamoun, S. 2006. A catalogue of the effector secretome of plant pathogenic oomycetes. Annu Rev Phytopathol 44: 41-60. doi: 10.1146/annurev.phyto.44.070505.143436

Kämper, J., R. Kahmann, M. Bolker, L. J. Ma, T. Brefort, B. J. Saville, F. Banuett, J. W. Kronstad, S. E. Gold, O. Muller, M. H. Perlin, H. A. Wosten, R. de Vries, J. Ruiz-Herrera, C. G. Reynaga-Pena, K. Snetselaar, M. McCann, J. Perez-Martin, M. Feldbrugge, C. W. Basse, G. Steinberg, J. I. Ibeas, W. Holloman, P. Guzman, M. Farman, J. E. Stajich, R. Sentandreu, J. M. Gonzalez-Prieto, J. C. Kennell, L. Molina, J. Schirawski, A. Mendoza-Mendoza, D. Greilinger, K. Munch, N. Rossel, M. Scherer, M. Vranes, O. Ladendorf, V. Vincon, U. Fuchs, B. Sandrock, S. Meng, E. C. Ho, M. J. Cahill, K. J. Boyce, J. Klose, S. J. Klosterman, H. J. Deelstra, L. Ortiz-Castellanos, W. Li, P. Sanchez-Alonso, P. H. Schreier, I. Hauser-Hahn, M. Vaupel, E. Koopmann, G. Friedrich, H. Voss, T. Schluter, J. Margolis, D. Platt, C. Swimmer, A. Gnirke, F. Chen, V. Vysotskaia, G. Mannhaupt, U. Guldener, M. Munsterkotter, D. Haase, M. Oesterheld, H. W. Mewes, E. W. Mauceli, D. DeCaprio, C. M. Wade, J. Butler, S. Young, D. B. Jaffe, S. Calvo, C. Nusbaum, J. Galagan and B. W. Birren 2006. Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis. Nature 444: 97-101. doi: 10.1038/nature05248

Kanehisa, M. 2002. The KEGG database. Novartis Found Symp 247: 91-101; discussion 101-103, 119-128, 244-152.

Katoh, K., K. Misawa, K. Kuma and T. Miyata 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30: 3059-3066.

Kemen, E., A. Gardiner, T. Schultz-Larsen, A. C. Kemen, A. L. Balmuth, A. Robert-Seilaniantz, K. Bailey, E. Holub, D. J. Studholme, D. Maclean and J. D. Jones 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of Arabidopsis thaliana. PLoS Biol 9: e1001094. doi: 10.1371/journal.pbio.1001094

Kemen, E. and J. D. Jones 2012. Obligate biotroph parasitism: can we link genomes to lifestyles? Trends Plant Sci 17: 448-457. doi: 10.1016/j.tplants.2012.04.005

Kohany, O., A. J. Gentles, L. Hankus and J. Jurka 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7: 474. doi: 10.1186/1471-2105-7-474

Krogh, A., B. Larsson, G. von Heijne and E. L. Sonnhammer 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305: 567-580. doi: 10.1006/jmbi.2000.4315

Krzywinski, M., J. Schein, I. Birol, J. Connors, R. Gascoyne, D. Horsman, S. J. Jones and M. A. Marra 2009. Circos: an information aesthetic for comparative genomics. Genome Res 19: 1639-1645. doi: 10.1101/gr.092759.109

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S. L. Salzberg 2004. Versatile and open software for comparing large genomes. Genome Biol 5: R12. doi: 10.1186/gb-2004-5-2-r12

Langmead, B. and S. L. Salzberg 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9: 357-359. doi: 10.1038/nmeth.1923

Laura Baxter, Sucheta Tripathy, Naveed Ishaque, Nico Boot, Adriana Cabral, Eric Kemen, Marco Thines, Audrey Ah-Fong, Ryan Anderson, Wole Badejoko, Peter Bittner-Eddy, Jeffrey L. Boore, Marcus C. Chibucos, Mary Coates, Paramvir Dehal, Kim Delehaunty, Suomeng Dong, Polly Downton, Bernard Dumas, Georgina Fabro, Catrina Fronick, Susan I. Fuerstenberg, Lucinda Fulton, Elodie Gaulin, Francine Govers, Linda Hughes, Sean Humphray, Rays H. Y. Jiang, Howard Judelson, Sophien Kamoun, Kim Kyung, Harold Meijer, Patrick Minx, Paul Morris, Joanne Nelson, Vipa Phuntumart, Dinah Qutob, Anne Rehmany, Alejandra Rougon-Cardoso, Peter Ryden, Trudy Torto-Alalibo, David Studholme, Yuanchao Wang, Joe Win, Jo Wood, Sandra W. Clifton, Jane Rogers, Guido Van den Ackerveken, Jonathan D. G. Jones, John M. McDowell, Jim Beynon, Brett M. Tyler 2010. Signatures of Adaptation to Obligate Biotrophy in the Hyaloperonospora arabidopsidis Genome. Science 330.

Laurie, J. D., S. Ali, R. Linning, G. Mannhaupt, P. Wong, U. Guldener, M. Munsterkotter, R. Moore, R. Kahmann, G. Bakkeren and J. Schirawski 2012. Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. Plant Cell 24: 1733-1745. doi: 10.1105/tpc.112.097261

Li, L., C. J. Stoeckert, Jr. and D. S. Roos 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res 13: 2178-2189. doi: 10.1101/gr.1224503

Lima, T., A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. de Castro, C. Lachaize, D. Baratin, I. Phan, L. Bougueleret and A. Bairoch 2009. HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot. Nucleic Acids Res 37: D471-478. doi: 10.1093/nar/gkn661

Liu, Q., A. J. Mackey, D. S. Roos and F. C. Pereira 2008. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. Bioinformatics 24: 597-605. doi: 10.1093/bioinformatics/btn004

Lowe, T. M. and S. R. Eddy 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25: 955-964.

Loytynoja, A. and N. Goldman 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11: 579. doi: 10.1186/1471-2105-11-579

Ma, L. J., H. C. van der Does, K. A. Borkovich, J. J. Coleman, M. J. Daboussi, A. Di Pietro, M. Dufresne, M. Freitag, M. Grabherr, B. Henrissat, P. M. Houterman, S. Kang, W. B. Shim, C. Woloshuk, X. Xie, J. R. Xu, J. Antoniw, S. E. Baker, B. H. Bluhm, A. Breakspear, D. W. Brown, R. A. Butchko, S. Chapman, R. Coulson, P. M. Coutinho, E. G. Danchin, A. Diener, L. R. Gale, D. M. Gardiner, S. Goff, K. E. Hammond-Kosack, K. Hilburn, A. Hua-Van, W. Jonkers, K. Kazan, C. D.

Kodira, M. Koehrsen, L. Kumar, Y. H. Lee, L. Li, J. M. Manners, D. Miranda-Saavedra, M. Mukherjee, G. Park, J. Park, S. Y. Park, R. H. Proctor, A. Regev, M. C. Ruiz-Roldan, D. Sain, S. Sakthikumar, S. Sykes, D. C. Schwartz, B. G. Turgeon, I. Wapinski, O. Yoder, S. Young, Q. Zeng, S. Zhou, J. Galagan, C. A. Cuomo, H. C. Kistler and M. Rep 2010. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464: 367-373. doi: 10.1038/nature08850

Majoros, W. H., M. Pertea and S. L. Salzberg 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20: 2878-2879. doi: 10.1093/bioinformatics/bth315

McAlpine, D. (Daniel); Victoria. Dept. of Agriculture 1910. *The smuts of Australia, their structure, life history, treatment, and classification* Melbourne, J. Kemp, government printer.

McTaggart, A. R., R. G. Shivas, A. D. Geering, K. Vanky and T. Scharaschkin 2012. A review of the complex. Persoonia 29: 55-62. doi: 10.3767/003158512X660283

Mi, H., B. Lazareva-Ulitsky, R. Loo, A. Kejariwal, J. Vandergriff, S. Rabkin, N. Guo, A. Muruganujan, O. Doremieux, M. J. Campbell, H. Kitano and P. D. Thomas 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33: D284-288. doi: 10.1093/nar/gki078

Nielsen, R. and Z. Yang 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148: 929-936.

Ostlund, G., T. Schmitt, K. Forslund, T. Kostler, D. N. Messina, S. Roopra, O. Frings and E. L. Sonnhammer 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res 38: D196-203. doi: 10.1093/nar/gkp931

Parra, G., K. Bradnam and I. Korf 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061-1067. doi: 10.1093/bioinformatics/btm071

Petersen, T. N., S. Brunak, G. von Heijne and H. Nielsen 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods 8: 785-786. doi: 10.1038/nmeth.1701

Ploch, S., S. Telle, Y. J. Choi, J. H. Cunnington, M. Priest, C. Rost, H. D. Shin and M. Thines 2011. The molecular phylogeny of the white blister rust genus Pustula reveals a case of underestimated biodiversity with several undescribed species on ornamentals and crop plants. Fungal Biol 115: 214-219. doi: 10.1016/j.funbio.2010.12.004

Price, A. L., N. C. Jones and P. A. Pevzner 2005. De novo identification of repeat families in large genomes. Bioinformatics 21 Suppl 1: i351-358. doi: 10.1093/bioinformatics/bti1018

Raffaele, S., R. A. Farrer, L. M. Cano, D. J. Studholme, D. MacLean, M. Thines, R. H. Jiang, M. C. Zody, S. G. Kunjeti, N. M. Donofrio, B. C. Meyers, C. Nusbaum and S. Kamoun 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 330: 1540-1543. doi: 10.1126/science.1193070

Schattner, P., A. N. Brooks and T. M. Lowe 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res 33: W686-689. doi: 10.1093/nar/gki366

Schirawski, J., G. Mannhaupt, K. Munch, T. Brefort, K. Schipper, G. Doehlemann, M. Di Stasio, N. Rossel, A. Mendoza-Mendoza, D. Pester, O. Muller, B. Winterberg, E. Meyer, H. Ghareeb, T. Wollenberg, M. Munsterkotter, P. Wong, M. Walter, E. Stukenbrock, U. Guldener and R. Kahmann 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330: 1546-1548. doi: 10.1126/science.1195330

Schultz, J., F. Milpetz, P. Bork and C. P. Ponting 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A 95: 5857-5864.

Sigrist, C. J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch and P. Bucher 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3: 265-274.

Slater, G. S. and E. Birney 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6: 31. doi: 10.1186/1471-2105-6-31

Sonnhammer, E. L., S. R. Eddy and R. Durbin 1997. Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 28: 405-420.

Stamatakis, A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22: 2688-2690. doi: 10.1093/bioinformatics/btl446

Stamatakis, A., T. Ludwig and H. Meier 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21: 456-463. doi: 10.1093/bioinformatics/bti191

Stoll, M., D. Begerow and F. Oberwinkler 2005. Molecular phylogeny of Ustilago, Sporisorium, and related taxa based on combined analyses of rDNA sequences. Mycol Res 109: 342-356.

Storey, John D. 2002. A direct approach to false discovery rates. Journal of the Royal Statistical Society: Series B 64: 479-498.

Suh, S. O. and J. Sugiyama 1993. Phylogeny among the basidiomycetous yeasts inferred from small subunit ribosomal DNA sequence. J Gen Microbiol 139: 1595-1598.

Tatusov, R. L., N. D. Fedorova, J. D. Jackson, A. R. Jacobs, B. Kiryutin, E. V. Koonin, D. M. Krylov, R. Mazumder, S. L. Mekhedov, A. N. Nikolskaya, B. S. Rao, S. Smirnov, A. V. Sverdlov, S. Vasudevan, Y. I. Wolf, J. J. Yin and D. A. Natale 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41. doi: 10.1186/1471-2105-4-41

team, R Development core 2008. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Thompson, J. D., T. J. Gibson and D. G. Higgins 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics Chapter 2: Unit 2 3. doi: 10.1002/0471250953.bi0203s00

Tyler, B. M. 2009. Entering and breaking: virulence effector proteins of oomycete plant pathogens. Cell Microbiol 11: 13-20. doi: 10.1111/j.1462-5822.2008.01240.x

Vánky, Kálmán 1994. *European smut fungi* Stuttgart; Jena; New York: Gustav Fischer Verlag.

Vánky, Kálmán 2002. *Illustrated genera of smut fungi*. St. Paul, Minnesota: American Phytopathological Society.

Wallace, A. R. 1858. On the tendency of varieties to depart indefinitely from the original type. . Journal of the Proceedings of the Linnean Society: Zoology 3: 53-62.

Weiß M, Bauer R, Begerow D 2004. Spotlights on heterobasidiomycetes In Agerer, Blanz, Piepenbring (eds.) Frontiers in Basidiomycete Mycology. IHW-Verlag: 7-48.

Win, J., W. Morgan, J. Bos, K. V. Krasileva, L. M. Cano, A. Chaparro-Garcia, R. Ammar, B. J. Staskawicz and S. Kamoun 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. Plant Cell 19: 2349-2369. doi: 10.1105/tpc.107.051037

Yang, Z. 2000. Maximum likelihood estimation on large phylogenies and analysis of adaptive evolution in human influenza virus A. J Mol Evol 51: 423-432.

Yang, Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586-1591. doi: 10.1093/molbev/msm088

Yang, Z. and J. P. Bielawski 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496-503.

Yang, Z. and R. Nielsen 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol 19: 908-917.

Yang, Z., W. S. Wong and R. Nielsen 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol Biol Evol 22: 1107-1118. doi: 10.1093/molbev/msi097

Zdobnov, E. M. and R. Apweiler 2001. InterProScan--an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847-848.

Zerbino, D. R. and E. Birney 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18: 821-829. doi: 10.1101/gr.074492.107

Zhang, J., R. Nielsen and Z. Yang 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol 22: 2472-2479. doi: 10.1093/molbev/msi237

Table 1

Species	TDN[1] (A)	3kbD[2] (B)	ø 3kbD[3]	(A ∩ B)	(A U B)
Melanopsichium pennsylvanicum	17	22	2	15	24
Ustilago hordei	18	27	1	17	29
Sporisorium reilianum	37	54	3	33	55
Ustilago maydis	35	43	3	32	46

[1] Three direct neighbors approach
[2] 3kb distance approach
[3] Clusters not found by the 3kb distance approach, but predicted by the "three direct neighbors" approach

Table 2

Query cluster [1] / Subject cluster[2]	*M. pennsylvanicum*	*U. maydis*	*S. reilianum*	*U. hordei*
M. pennsylvanicum	-	13	13	8
U. maydis	13	-	35[a]	23[b]
S. reilianum	13	31	-	22
U. hordei	8	21	24[c]	-
Conserved in all	7	10	9	7
Own clusters	8	12	14	3

[1] Genome which pathogenicity clusters were tested for conservation
[2] Genome querried for pathogenicity clusters conservation
[a] 4 of the pathogenicity clusters of *S. reilianum* were fragmented, with respect to *U. maydis* clusters
[b] 2 of the pathogenicity clusters of *U. hordei* were fragmented, with respect to *U. maydis* clusters
[c] 2 of the pathogenicity clusters of *S. reilianum* were fragmented, with respect to *U. hordei* clusters

Genome Biology and Evolution

SMBE

Figure 1

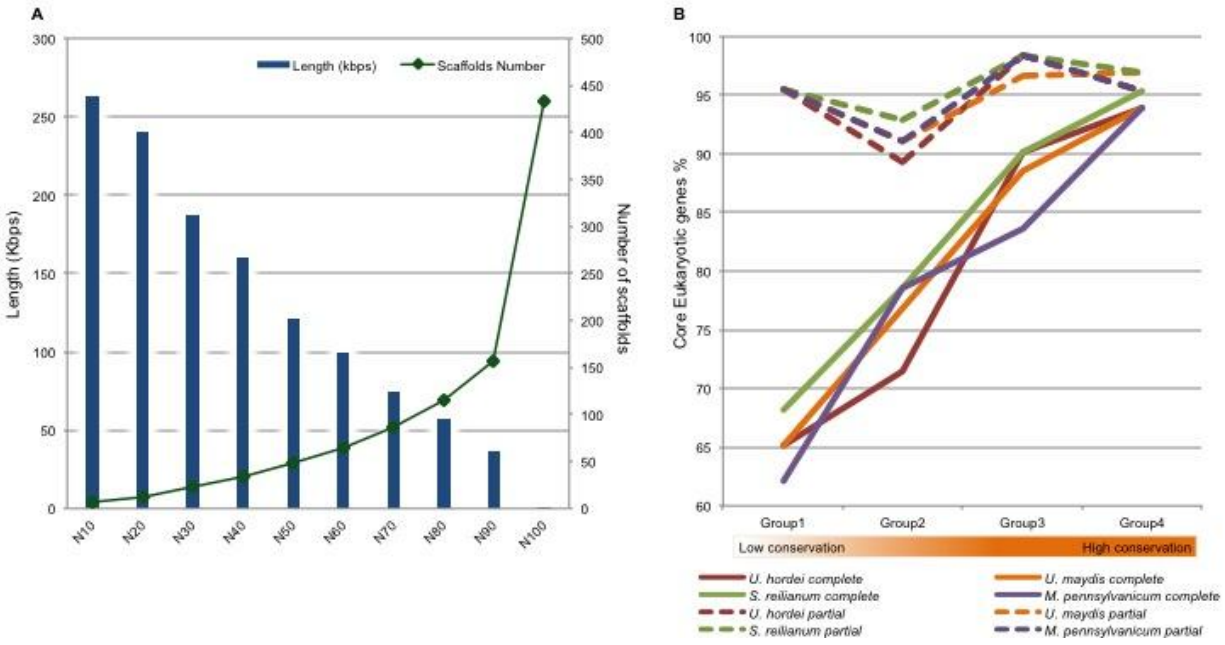

Figure 1. Genome assembly quality and completeness estimation. (A) Genome assembly quality as assessed by calculating the number of scaffolds and minimum scaffold length in the respective N-class, where N is the percentage of the genome covered after sorting the scaffolds from largest to smallest. **(B)** Genome completeness assessed by CEGMA analysis on the basis of 458 euKaryotic clusters of Orthologous Groups (KOGs). The CEGMA pipeline categorizes these 458 core proteins in four groups on the basis of their conservation in the eukaryotic genomes. Dotted and solid lines are representing figures when considering partial mapping and complete mapping of the KOGs, respectively.

Figure 2

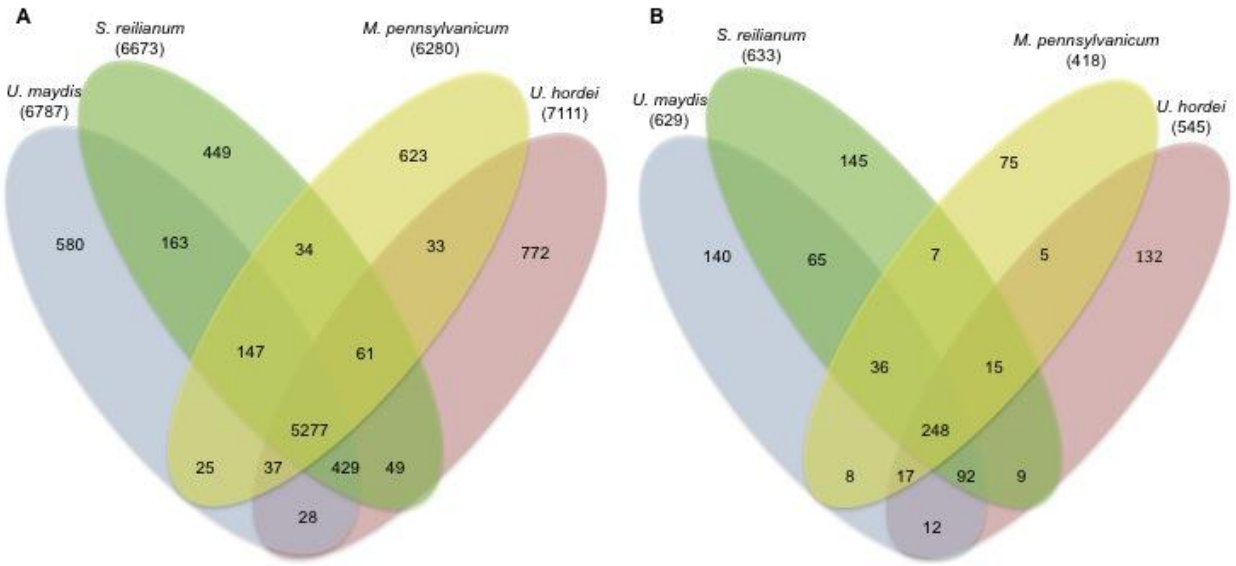

Figure 2. Orthologous genes within the four smut genomes. (A) Venn diagram representing the orthologous genes within the four genomes. **(B)** Orthology of the putative secreted effector proteins from all four genomes.

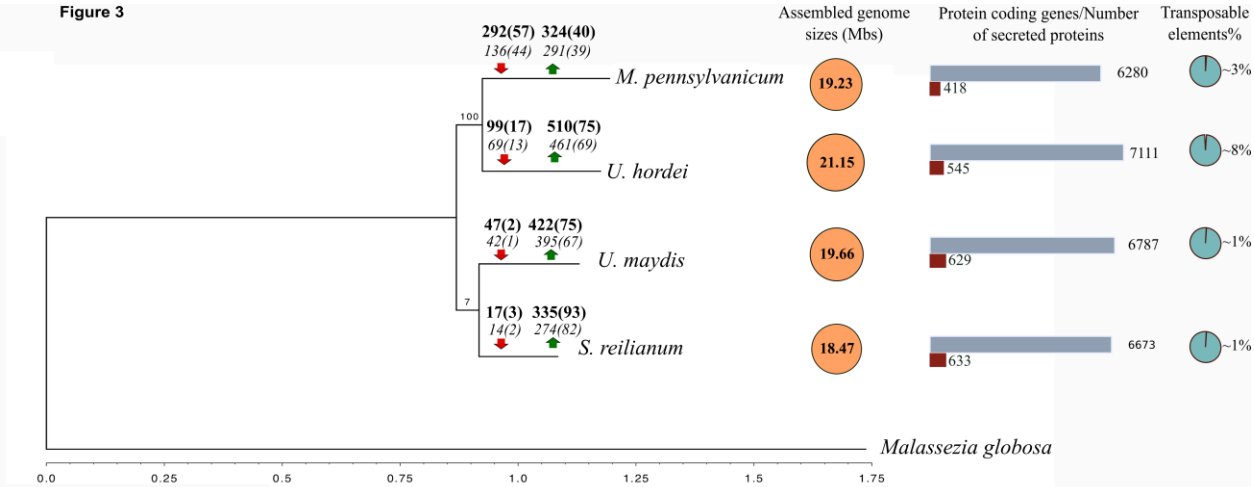

Figure 3. Phylogenetic distance estimation using all nuclear genes 1:1 orthologs from five species. Maximum likelihood inference based on MAFFT alignments using RAxML with 1000 bootstrap replicates. Numbers at branches indicate bootstrap support percentages for the respective branches. Red and green arrows represent the number of gene lost and gained, respectively, number in brackets represents the number of gene lost or gained in terms of PSEP-encoding genes. Numbers in bold represent gene losses/gains in the four genomes without including gene remnants in intergenic regions. Genes losses/gains were further tested for gene remnants or distantly similar genes using lower cutoffs, the corresponding figures are given in italics. Genome sizes refer to assembled genome sizes.

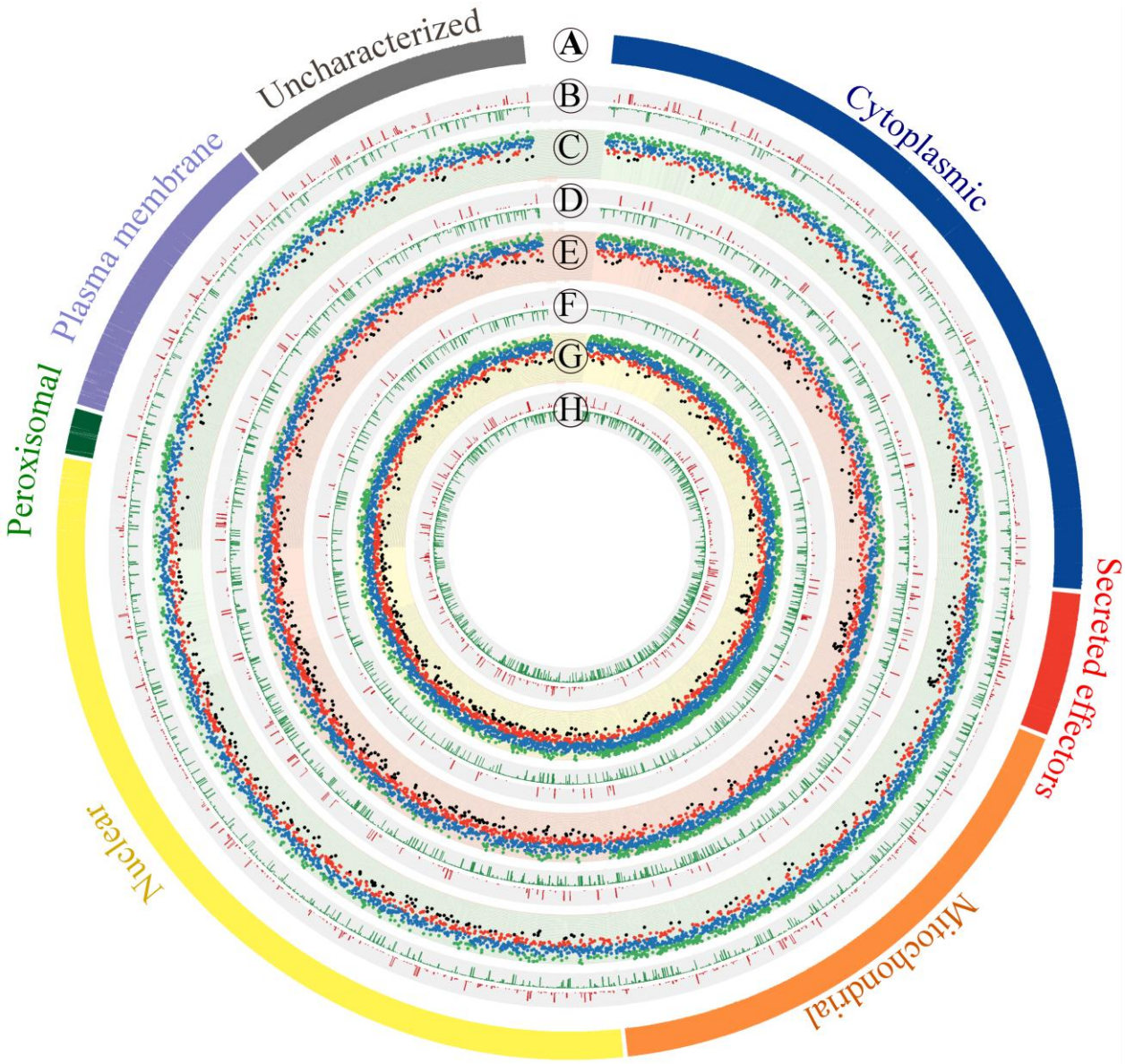

Figure 4. Conservation of proteins in smut genomes and their dN/dS ratios according to their subcellular localizations. The outmost ring 'A' represents *M. pennsylvanicum* protein sequences according to their subcellular localization. Ring 'B' represents the dN/dS ratios of the proteins of *M. pennsylvanicum* shown in Ring 'A'. Red and green bars represent the dN/dS ratios of the positively selected (1% FDR, >95% BEB confidence) and non-selected (Non-significant considering 1% FDR) genes, respectively. Similarly the rings 'D', 'F' and 'H' represent the dN/dS ratios of *S. reilianum*, *U. maydis* and *U. hordei*, respectively. Rings 'C', 'E' and 'G' represent the BlastP percentage identity of *M. pennsylvanicum* proteins with the *S. reilianum*, *U. maydis* and *U. hordei* proteins, respectively. Green dots highlight a BlastP identity greater than 85%, blue dots represent an identity in the range of 65-85%, red dots in the range of 50-65% and black dots are highlighting a BlastP percentage identity less than 50%.

Genome Biology and Evolution
SMBE

Figure 5

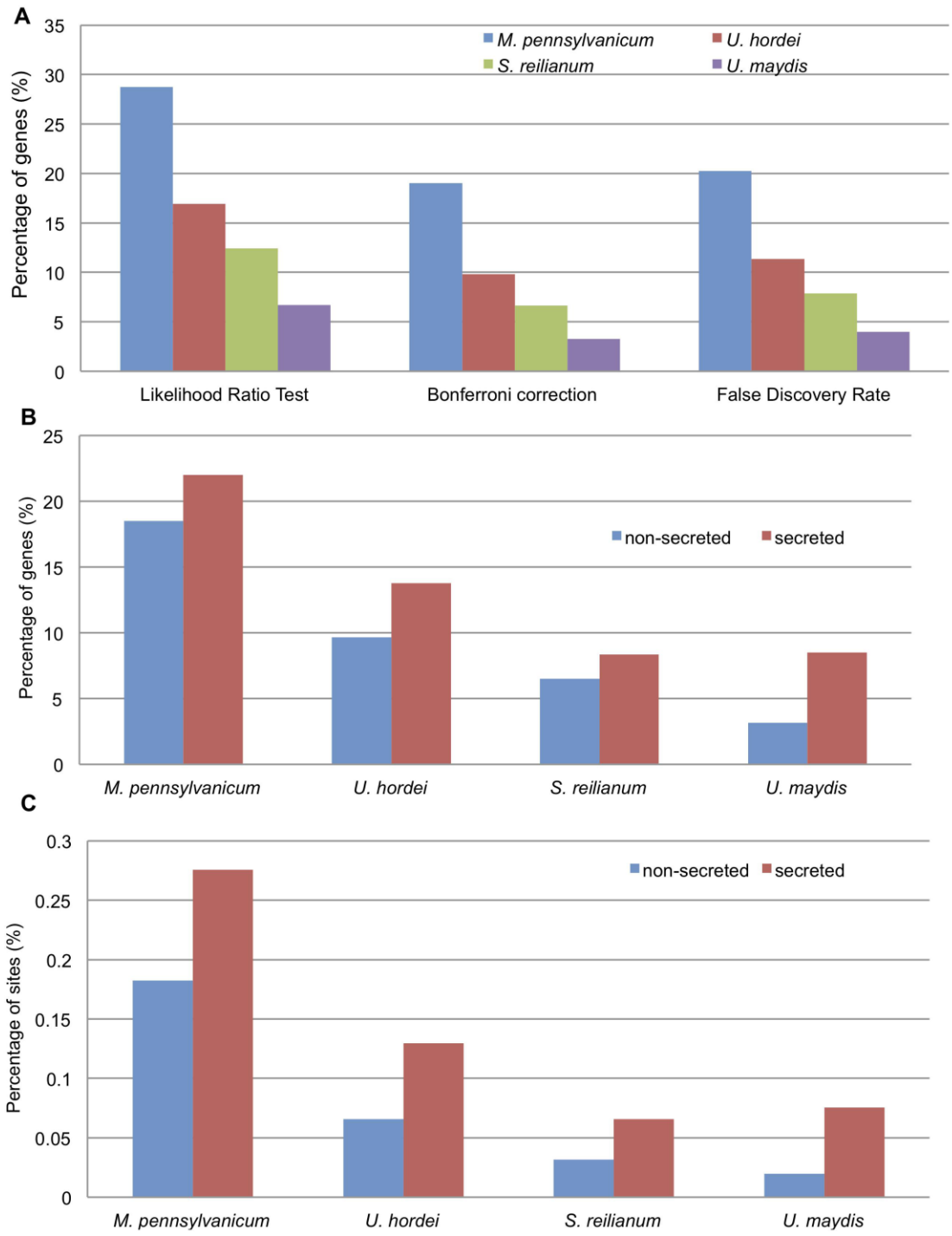

Figure 5. Percentage of positive selection among the four genomes and comparisons of positively selected genes encoding PSEPs and non-secreted proteins. (A) Percentage of positively selected genes among the four species. **(B)** Percentage of positively selected PSEP-encoding and non-secreted protein encoding genes. **(C)** Percentage of positively selected sites within the four genomes.

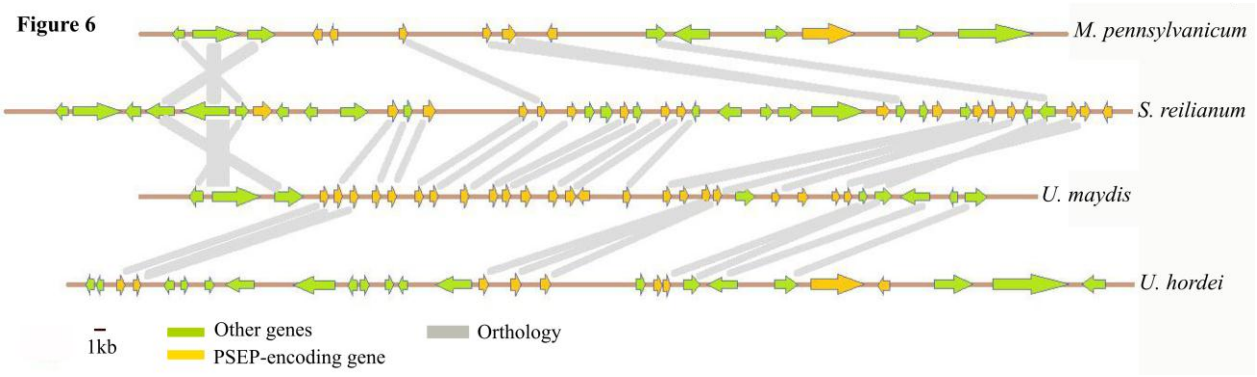

Figure 6. *Ustilago maydis* 19A pathogenicity cluster synteny in the other three genomes. Grey shading represents orthologous genes. Yellow arrows represent PSEPs and green arrows represent the non-secreted protein encoding genes. The orientation of the arrows represents the orientation of the genes and length of arrows is proportional to the length of the gene.

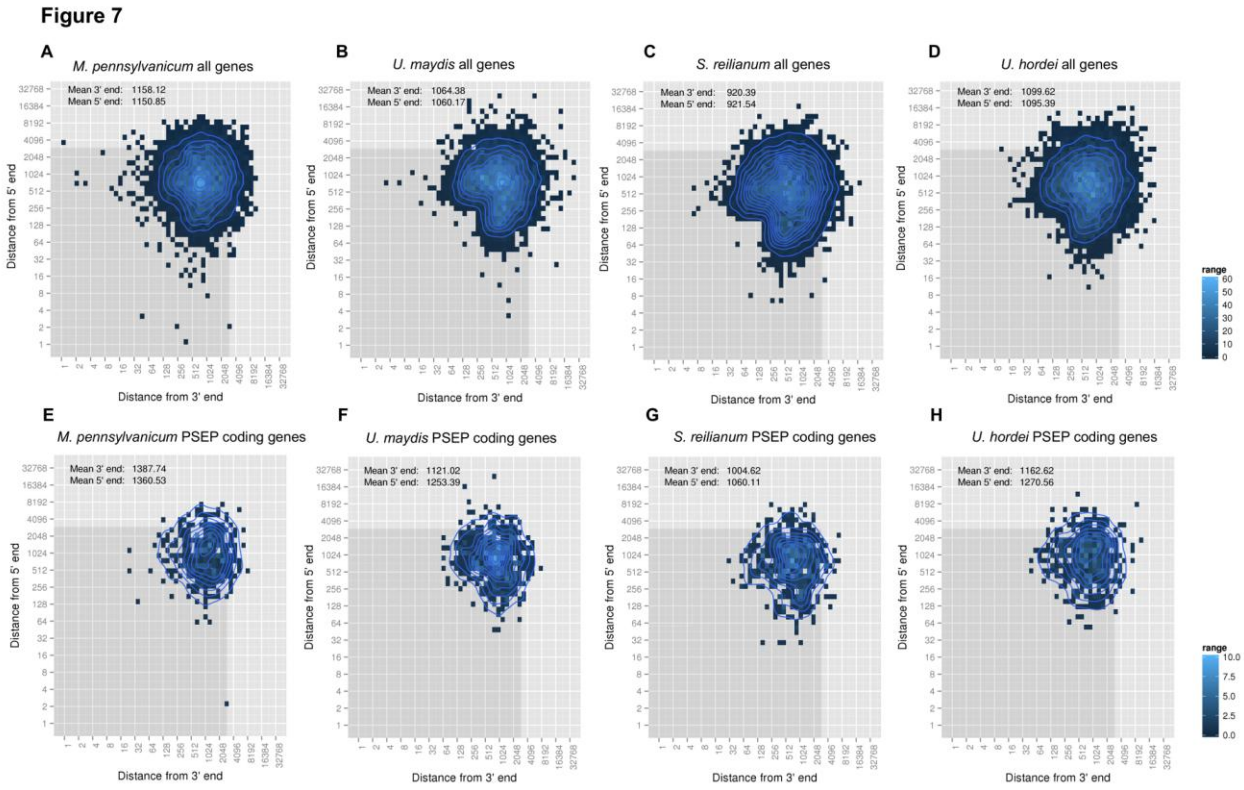

Figure 7. Heat-maps representing the 3' and 5' end flanking region lengths. (A-D) Heat-maps of 3' and 5' end flanking regions of all genes of *M. pennsylvanicum, U. maydis, S. reilianum,* and *U. hordei*, respectively. **(E-H)** Heat-maps of 3' and 5' end flanking regions of PSEP-encoding genes of *M. pennsylvanicum, U. maydis, S. reilianum,* and *U. hordei*, respectively. The 3' and 5' distance values less than 3kb are represented in the portion of the plot shaded grey.

Supplementary material is available upon publication or upon request from the corresponding author.