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], and Marco Thines[1,2,3,4,*]

[1]Biodiversity and Climate Research Centre (BiK-F), Frankfurt am Main, Germany

[2]Institute of Ecology, Evolution and Diversity, Goethe University, Frankfurt am Main, Germany

[3]Senckenberg Gesellschaft für Naturforschung, Frankfurt am Main, Germany

[4]Cluster for Integrative Fungal Research (IPF), Frankfurt am Main, Germany

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

*Corresponding author: E-mail: marco.thines@senckenberg.de.

Accepted: July 5, 2014

Data deposition: This project has been deposited at the European Nucleotide Archive (ENA) under the accession number PRJEB4565.

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 *Me. pennsylvanicum* was sequenced and compared with the genomes of *U. maydis*, *Sporisorium reilianum*, and *U. hordei*. Although all four genomes had a similar completeness in CEGMA (Core Eukaryotic Genes Mapping Approach) analysis, gene absence was highest in *Me. 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 *Me. 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 noneffector candidates. The 248 putative secreted effectors found in all four smut genomes might constitute a core set needed for pathogenicity, whereas those 92 that are found in all grass-parasitic smuts but have no ortholog in *Me. pennsylvanicum* might constitute a set of effectors important for successful colonization of grass hosts.

Key words: comparative genomics, effector genes, evolutionary biology, genome assembly, host jump, positive selection, smut fungi.

Introduction

Melanopsichium pennsylvanicum is a nonobligate 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 (McAlpine 1910; Fischer 1953). Most of the members of Ustilaginaceae are parasitic to Poales and Cyperales, and some are infecting economically important cereal crops such as maize, barley, wheat, and oat (Vánky 1994). As an exception among Ustilaginaceae, *Me. pennsylvanicum* colonizes the dicot genus *Persicaria* (Begerow et al. 2000, 2006). Mature galls are often covered with black material, which hardens upon desiccation (Halisky and Barbe 1962; Vánky 2002), in contrast to most members of the Ustilaginaceae which liberate a powder of dark-colored spores from their galls. Molecular phylogenetic studies have revealed that *Me. pennsylvanicum* is embedded in *Ustilago* s.l. infecting *Poaceae* (Begerow et al. 2004; Weiß et al. 2004; Stoll et al. 2005). Other, more distantly related species of the Ustilaginaceae are parasitic to the monocot Cyperaceae and Juncaceae (Begerow et al. 2000, 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 (Kamoun 2006; Birch et al. 2008; Doehlemann et al. 2009; Tyler 2009; Djamei et al. 2011). 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 *U. maydis* (Kämper et al. 2006). In general, PSEPs show higher sequence divergence and less sequence conservation than noneffector 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 (Schirawski et al. 2010; Laurie et al. 2012).

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 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 noneffector genes affected by the adaptation process?

To address the above-mentioned questions, whole-genome sequencing, assembly, and annotation of the *Me. pennsylvanicum* strain 4 (Mp4) were performed using high-throughput sequencing 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.

This 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.

Materials and Methods

DNA Isolation and Sample Preparation

DNA was isolated from the single yeast strain Mp4, which was grown from the specimen Mycotheca Graecensis number 285 distributed by the herbarium GZU. Yeasts cells were harvested from PDA (Potato Dextrose Agar) medium using a phenol–chloroform extraction method as described in Ploch et al. (2011).

Data Preprocessing

Illumina reads of 76 bp read length and 300 bp 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 reads for *k*-mers 21, 31, 41, 45, 49, 51, 55, and 61 for calculating the average insert size and insert size standard deviation. Reads from both of the lanes used were mapped back on the assemblies using Bowtie2 (Langmead and Salzberg 2012) with an input insert size within the limit of 100–600 bp. The resulting SAM file from Bowtie2 was used to calculate the average and standard deviation of 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 was calculated using the R statistical package (R Development Core Team 2008), according to the manual of the Velvet software. These values were in the range of 2–15 for all tested *k*-mers. All assemblies resulting 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 size of the assembled genome. Assemblies with a *k*-mer length of 63 and a *k*-mer coverage cutoff of 3 generated the best assembly. Scaffolding of the Velvet contigs was performed by the SSPACE (Boetzer et al. 2011) scaffolding package. The optimal scaffolded assembly was selected again considering the above-mentioned assembly quality parameters.

CEGMA Analyses to Check the Genome Completeness

CEGMA 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 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 the other three smut genomes available, the Mummer3 (Kurtz et al. 2004) whole-genome alignment tools were used. Mummer3 is a collection of tools 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 to the other three genomes at the 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 the protein level by translating the genome in all six reading frames prior to 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 *U. maydis*, *U. hordei, S. 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 supplementary table S9, Supplementary Material online.

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. Low-complexity regions were removed using the "filter-stage-1.prl" script and tandem repeats using the TRF software (Benson 1999), and the script "filter-stage-2.prl" was used to filter out the repeat elements not present more than three times in different genomic locations. RepeatMasker (A.F.A. Smit, R. Hubley, and P. Green; RepeatMasker at http://repeatmasker.org, last accessed July 22, 2014) was used to mask the repeats predicted 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 repeat masking using RepeatMasker.

Gene Prediction

Both ab initio and homology-based prediction were used for defining the genes within the repeat-masked *Me. pennsylvanicum* genome (supplementary fig. S5, Supplementary Material online). 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 resulting *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 tRNA-Scan (Lowe and Eddy 1997; Schattner et al. 2005) according to the user manual.

In the later annotation steps, all intergenic sequences were extracted and aligned against all protein sequences from the three Ustilaginales genomes including the protein sequences of Mp4 generated from the above-mentioned annotations. These alignments were performed by tBLASTn (Altschul et al. 1990) and Exonerate (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 (GO), and biological pathway information of the predicted genes of Mp4. The InterProScan program searches the 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). These searches provide information regarding the GO (Harris et al. 2004) and KEGG (Kanehisa 2002) pathways of the predicted genes.

Prediction of 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 identify proteins with an 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 four Ustilaginales genomes were investigated and were compared with published data. Another set of candidate secreted proteins was generated by using TargetP v1 (Emanuelsson et al. 2007) for all 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 nonsecreted protein-coding genes were allowed if the following gene was again a PSEP-encoding gene. This is referred to as the 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 3 kb 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 identified by both of these methods (supplementary fig. S6, Supplementary Material online).

To check the conservation of the pathogenicity clusters, orthologs were investigated in all four species. Pathogenicity clusters were defined as conserved in all 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

Protein subcellular localization was investigated using the ProtComp v9 package (www.softberry.com, last accessed July 22, 2014). ProtComp was locally installed and run for all four genomes and percentages of proteins localized to certain subcellular components were calculated. Transmembrane domains within the protein sequences were predicted using TMHMM2.0c (Krogh et al. 2001).

Ortholog Prediction

For ortholog predictions, both orthoMCL (Li et al. 2003) and Inparanoid (Ostlund et al. 2010) were run on the protein sequences of all four genomes. After generating the list of orthologs and paralogs, perl and shell scripts were used for further analysis of the orthology information obtained.

The percentage of identity of the 1:1 orthologs within the four smut genomes was calculated using BLASTP searches. A circular plot (fig. 1) of the aligned sequences was generated using the Circos package (Krzywinski et al. 2009).

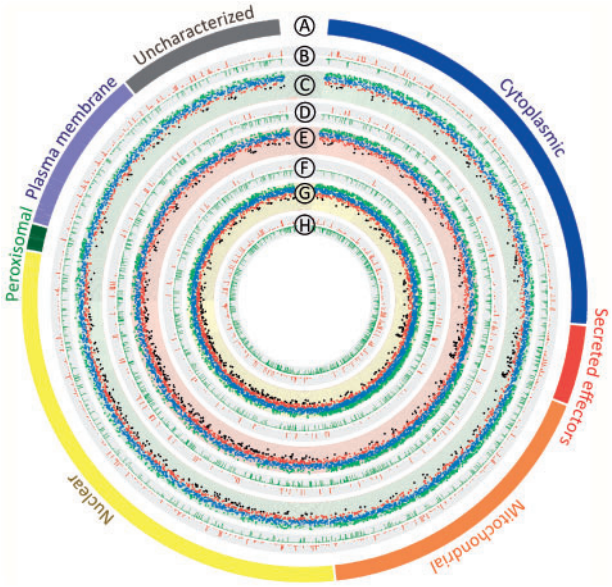

Fig. 1.—Conservation of proteins in smut genomes and their d*N*/d*S* ratios according to their subcellular localization. The outmost ring "A" represents *Melanopsichium pennsylvanicum* protein sequences according to their subcellular localization. Ring "B" represents the d*N*/d*S* ratios of the proteins of *Me. pennsylvanicum* shown in Ring "A." Red and green bars represent the d*N*/d*S* ratios of the positively selected (1% FDR, >95% BEB confidence) and nonselected (nonsignificant considering 1% FDR) genes, respectively. Similarly the rings "D," "F," and "H" represent the d*N*/d*S* ratios of *Sporisorium reilianum*, *Ustilago maydis* and *U. hordei*, respectively. Rings "C," "E" and "G" represent the BLASTP percentage identity of *Me. 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 identity less than 50%.

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 orthologs were present in all other three genomes and not predicted in the genome under consideration. Similarly a gene was considered 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 were further tested with tBLASTn and Exonerate searches, by compiling a database of all proteins from the four genomes and performing tBLASTn searches using an identity cutoff of 55% against the intergenic regions of all the four genomes separately. More relaxed searches were further done by using a 35% identity cutoff and an alignment length of 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 orthologs in other three species. For these BLASTP searches only those hits were considered which showed an alignment length more than 35% of the query protein length, *e* value less than e^{-2}, and more than 35% 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 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 to the next gene were calculated for all genes and for all four genomes. Heat-maps for all four genomes were generated using the ggplot2 (http://ggplot2.org/, last accessed July 22, 2014) 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 performed by using the smaller and greater lengths among 5′- and 3′-flanking distances of all genes and PSEP-encoding genes of all 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 alignments of all 5,200 1:1 orthologous genes of the four species. Alignments were used as input for RAxML (Stamatakis et al. 2005; Stamatakis 2006) for maximum-likelihood phylogenetic inference. RAxML was run with 1,000 bootstrap replicates using the GTRGAMMA model.

In another analysis, to produce a rooted tree, the genome of *Ma. globosa*, 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 with the parent protein.

In another approach, Exonerate was used to map the proteins from all four genomes on 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 presence of a start codon in any predicted gene structure. Pseudogenes were defined if the predicted gene structure was starting from the first 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 among the four genomes. Prank-codon has performed best in comparison to other multiple alignment tools (Fletcher and Yang 2010), such as Mafft (Katoh et al. 2002), Muscle (Edgar 2004) and 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. The test2 algorithm is highly recommended 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 (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, last accessed July 22, 2014) perl module.

To perform multiple hypothesis testing, BC (Anisimova and Yang 2007) and 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 (Nielsen and Yang 1998; Yang 2000; Yang and Bielawski 2000) and the BEB (Yang et al. 2005) information from the CodeML output files. Only those genes were considered to be under positive selection that had at least one site under selection with greater than 95% BEB confidence at less than 1% FDR.

Data Access

All 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 the annotations of the *Me. pennsylvanicum* genome have been submitted to the ENA and can be accessed from accession IDs HG529494 to HG529928. The Mp4 Genome scaffolds, protein sequences, and gff file are available at http://dx.doi.org/10.12761/SGN.2014.3 (last accessed July 22, 2014).

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 1,746 contigs with an N50 contig size of 43.37 kb, and the largest contig was of 218.8 kb. The final scaffolded nuclear genome assembly was of 19,156,659 bp and had an N50 scaffold size of 121.67 kb; the largest scaffold was of 690.5 kb, with 434 scaffolds in total. The mitochondrial genome was of 74,914 bp. Assembly quality was estimated by computing the size and number of scaffolds in respective N-classes (fig. 2*A*). 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.

The completeness and continuity of the assembly were tested using the CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline (Parra et al. 2007) and compared to the completeness and continuity of the other three Ustilaginaceae genomes. The completeness and continuity of the gene space of *Me. pennsylvanicum* was found to be comparable to the three other genomes: 95.16, 94.76, 95.97 and 95.16% of highly conserved genes were found in *Me. pennsylvanicum* (*Mp4*), *U. hordei* (*Uh*), *S. reilianum* (*Sr*), and *U. maydis* (*Uh*), respectively (fig. 2*B*). The genome of *Me. pennsylvanicum* was aligned with the other three smut genomes using Mummer3 (Kurtz et al. 2004) and mapped well on these (See supplementary fig. S1, Supplementary Material online).

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 Materials and Methods). A total of 6,280 genes were identified in the genome of *Me. pennsylvanicum* including 107 transfer RNAs.

To estimate the number of PSEPs, both SignalP v4 (Petersen et al. 2011) and TargetP v1 (Emanuelsson et al. 2007) were used on all four Ustilaginales genomes. These predictions generated 418 PSEP-encoding genes in the genome of *Me. 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); further confirmations were done using BLASTP and tBLASTn searches (see Materials and Methods). In total 5,277 orthologs were found to be present in all four species, out of these 5,200 were 1:1 orthologs (fig. 3*A*). In total, 623 genes present only in *Me. 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 *Me. pennsylvanicum*. In contrast 147, 37 and 61 orthologs 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 all genomes except for *Me. 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 (fig. 3*B*).

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

Gene Gain and Gene Loss

For assuring also the complete absence of similar genes, which did not fulfill the criteria for orthology 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 < 0.1 and percentage identity > 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

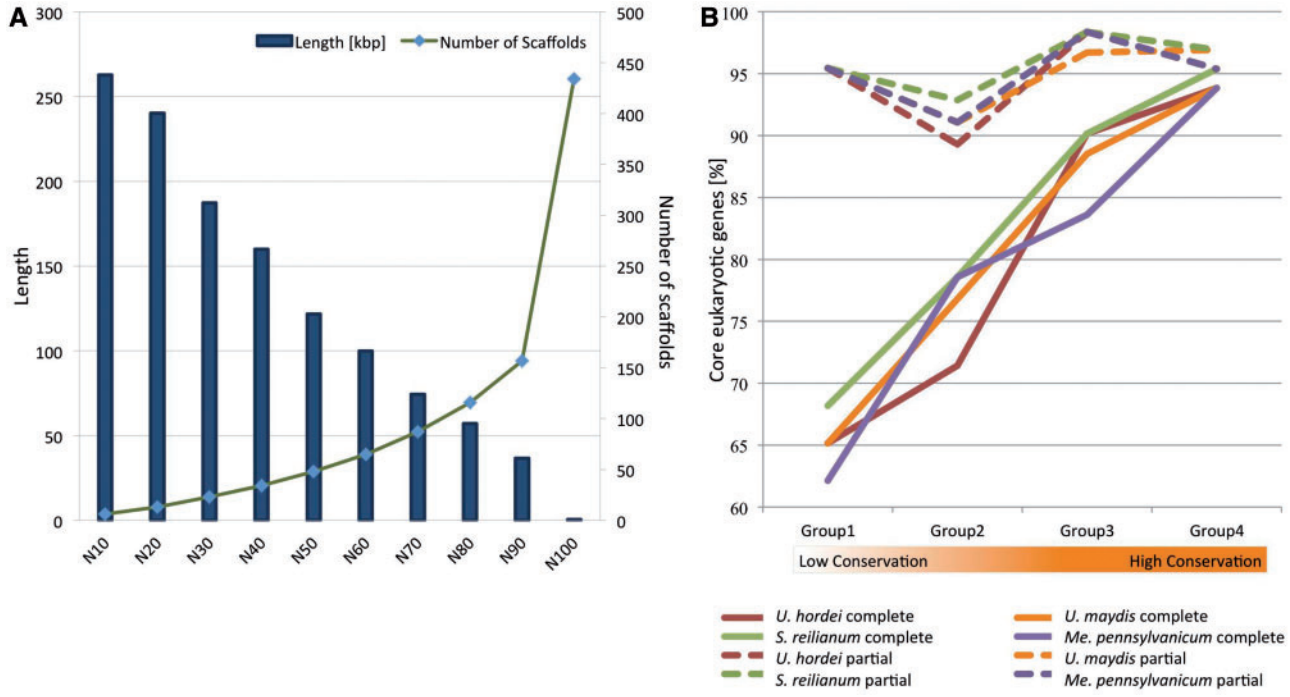

Fig. 2.—Genome assembly quality and completeness estimation. (*A*) Genome assembly quality as assessed by calculating the number of scaffolds and minimum scaffold length in the respective N-class, where N is the percentage of the genome covered after sorting the scaffolds from largest to smallest. (*B*) Genome completeness assessed by CEGMA analysis on the basis of 458 KOGs. The CEGMA pipeline categorizes these 458 core proteins in four groups on the basis of their conservation in eukaryotic genomes. Dotted and solid lines are representing partial mapping and complete mapping of the KOGs, respectively.

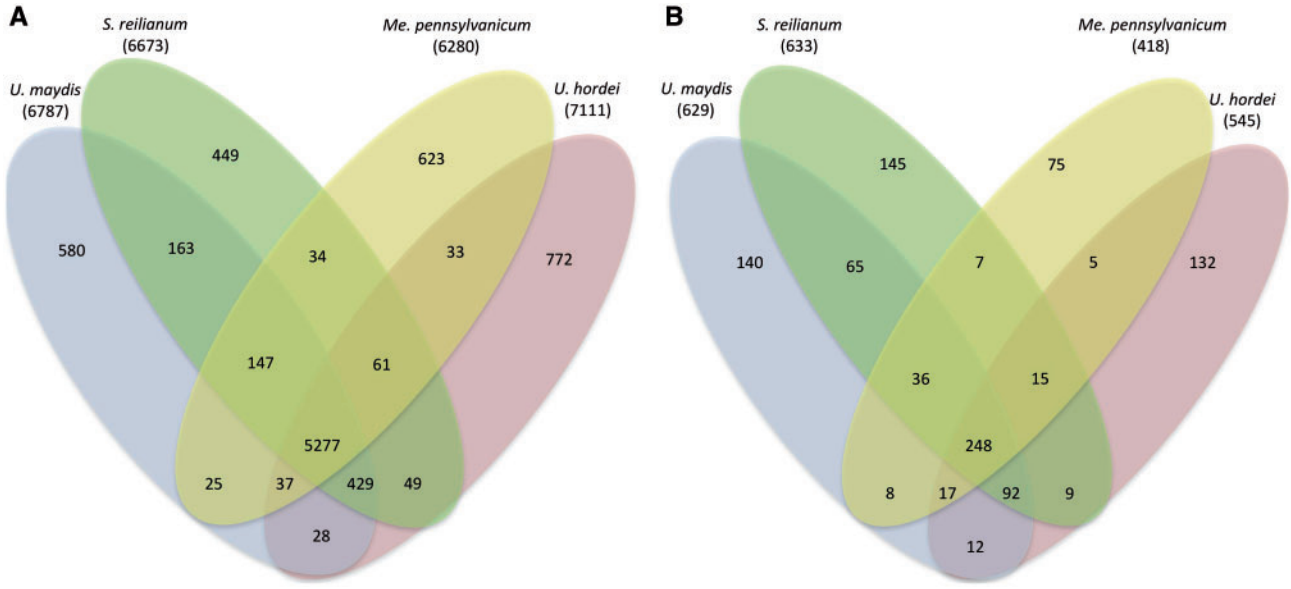

Fig. 3.—Orthologs within the four smut genomes. (*A*) Venn diagram representing the orthologs within the four genomes. (*B*) Orthology of the PSEPs from all four genomes.

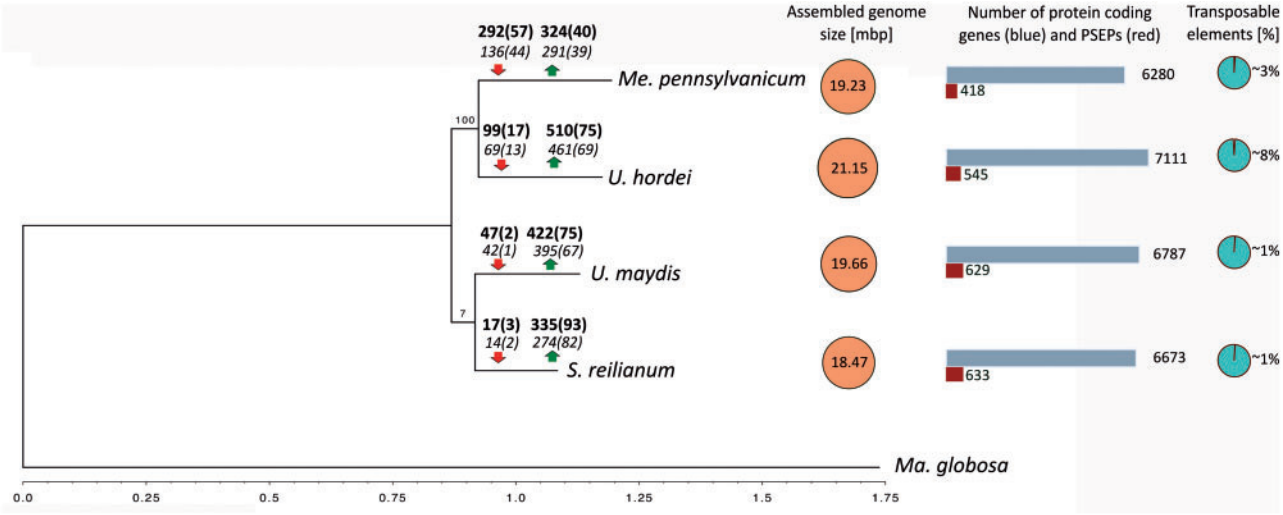

Fig. 4.—Phylogenetic tree reconstruction using all nuclear 1:1 orthologs from five species. Maximum-likelihood inference based on MAFFT alignments using RAxML with 1,000 bootstrap replicates. Numbers at branches indicate bootstrap support percentages for the respective branches. Red and green arrows 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.

292 genes present in all other species but absent in *Me. 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 *Me. pennsylvanicum* but present in the grass-infecting smuts are given in supplementary table S1, Supplementary Material online. 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 *Me. pennsylvanicum*, 69 in *U. hordei*, 14 in *S. reilianum*, and 42 in *U. maydis* (fig. 4). The functional annotations of these lost genes are given in supplementary table S1, Supplementary Material online.

In terms of PSEP-encoding genes, 57 (supplementary table S1, Supplementary Material online) were absent in the genome of *Me. pennsylvanicum*, 17, 3, and 2 were absent in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively (fig. 4). Of these genes 44, 13, 1, and 2 of *Me. 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 *Me. 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 *Me. 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 *Me. pennsylvanicum*, but 75, 93, and 75 were gained in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively (fig. 4). When including searches for gene remnants and distantly similar genes, the respective figures were 291, 461, 274, and 395 for *Me. 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, although *Me. 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 the four species with tBLASTn and Exonerate (Slater and Birney 2005) (see Materials and Methods).

This approach led to the discovery of new genes with intact open reading frames that were not predicted by previous annotations. In total, 14 putative new genes were found in *Me. pennsylvanicum,* 55 in *U. hordei,* 3 in *S. reilianum*, and 23 in *U. maydis*. Using tBLASTn no pseudogene was found in *Me. pennsylvanicum*, 142 pseudogenes were found in *U. hordei,*

6 in *S. reilianum*, and 2 in *U. maydis*. Using Exonerate 3, 160, 2, and 9 pseudogenes were observed in the genomes of *Me. pennsylvanicum*, *U. hordei*, *S. reilianum,* and *U. 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 et al. 2005; Stamatakis 2006) was run with 1,000 bootstrap iterations on multiple sequence alignments of the 1:1 orthologs of the four smut fungi and *Ma. globosa*, which was used as outgroup. A total of 2,979 1:1 orthologs were found among the five genomes and subjected to alignments and phylogenetic analysis. A sister-group relationship of *U. maydis* and *S. reilianum* was found, without significant support. *Ustilago hordei* and *Me. pennsylvanicum* were found to group together, with maximum bootstrap support. The genetic distance between the four smut fungi was similar (fig. 4).

Protein Subcellular Localization

The ProtComp9 package (www.softberry.com, last accessed July 22, 2014) was used for the identification of protein subcellular localization in the four genomes. In total 1,445, 1,455, 1,582, and 1,470 mitochondria-targeted proteins were found in the genomes of *Me. pennsylvanicum*, *U. hordei, S. reilianum,* and *U. maydis*, respectively (supplementary fig. S2, Supplementary Material online). Further, 1,888 nuclear proteins were predicted in *Me. pennsylvanicum*, 2,139 in *U. hordei*, 1,910 in *S. reilianum,* and 1,992 in *U. maydis*. In addition, 1,351, 1,546, 1,271, and 1,323 proteins in the genome of *Me. 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 d*N*/d*S* ratio information were inferred using the codeml Branch-Site model of PAML at a 1% level of significance with respect to false discovery rate (FDR) and with Bayes Empirical Bayes (BEB) support greater than 95%. These analyses revealed the lowest sequence conservation and highest proportion of genes under positive selection for PSEP-encoding genes, whereas peroxisome-targeted genes were most conserved (fig. 1).

Patterns of Positive Selection among Four Smut Species

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

Positively Selected Putative Secreted Effectors and Nonsecreted Protein-Encoding Genes

To compare the percentage of PSEPs and nonsecreted protein-encoding genes positively selected with a high level of confidence among the four smut species, only those genes were used that had greater than 95% BEB support and an FDR at 1% level of significance.

It was revealed that 18.47% of the nonsecreted protein-encoding genes and 22.00% secreted protein-encoding genes of *Me. 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 nonsecreted and secreted protein-encoding genes, respectively.

We further assessed the percentage of positively selected sites among the tested genes that passed the BEB greater than 95% confidence threshold at an FDR less than 1%. A higher percentage of selected sites was found in the PSEP-encoding genes (fig. 5*C*), compared with the noneffector genes in all four genomes.

Candidate Pathogenicity Clusters

Using the "three direct neighbor" (TDN) approach (see Materials and Methods) 23 new candidate pathogenicity clusters were defined in *U. 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 *Me. pennsylvanicum*.

A sliding window of 3 kb (supplementary fig. S3, Supplementary Material online) was found to generate a similar amount of candidate pathogenicity clusters when compared to the TDN method. Using this method, a total of 22, 27, 53, and 43 candidate clusters were found in *Me. pennsylvanicum*, *U. hordei*, *S. reilianum,* and *U. maydis,* respectively. By combining the output of both methods, a total of 24, 29, 55, and 46 candidate clusters were found in *Me. pennsylvanicum*, *U. hordei*, *S. reilianum,* and *U. maydis*, respectively (table 1). Supplementary table S2, Supplementary Material online, 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 can be found in supplementary table S3, Supplementary Material online;

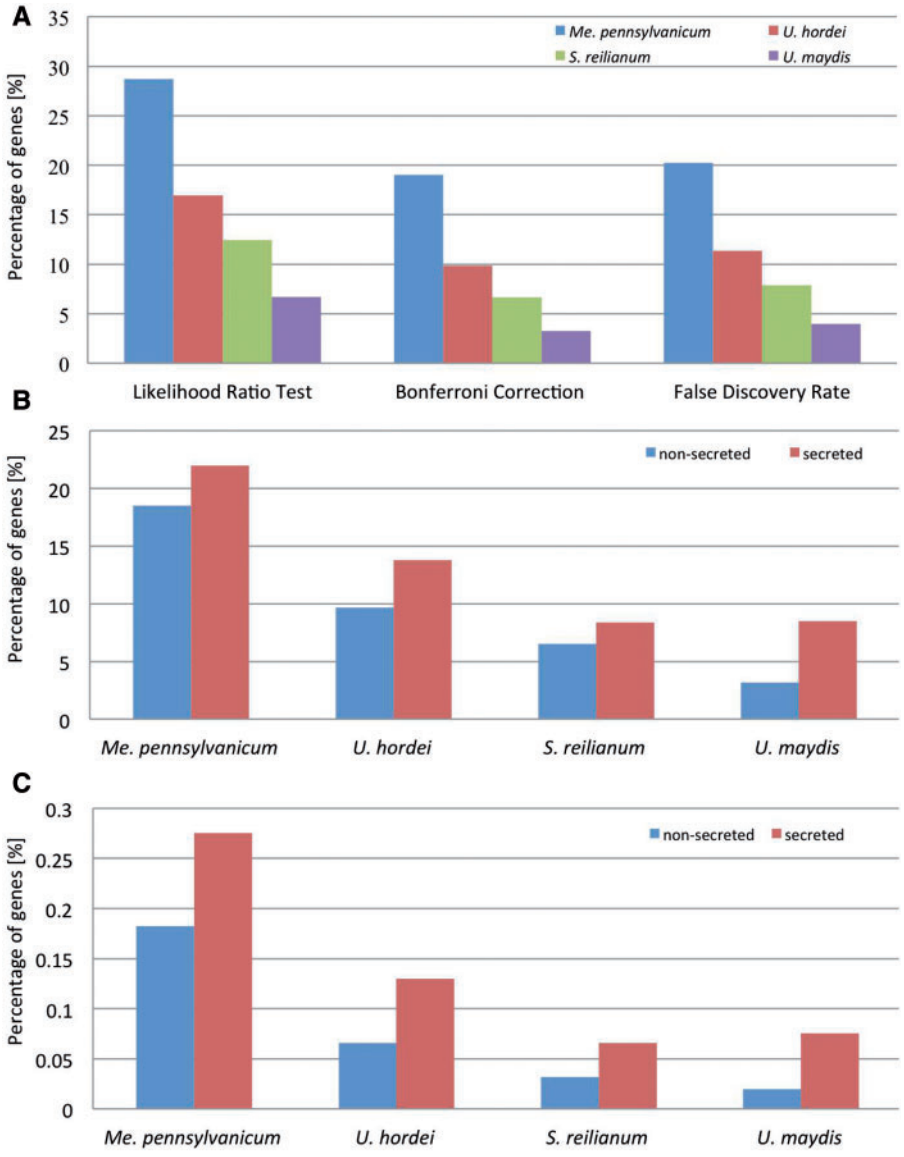

Fig. 5.—Percentage of positive selection among the four genomes and comparisons of positively selected genes encoding PSEPs and nonsecreted proteins. (*A*) Percentage of positively selected genes among the four species. (*B*) Percentage of positively selected PSEP-encoding and nonsecreted protein-encoding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

orthologs of the PSEP-encoding genes in the candidate pathogenicity clusters of *S. reilianum*, *U. hordei,* and *Me. pennsylvanicum* are listed in the supplementary tables S4–S6, Supplementary Material online, respectively.

Pathogenicity Cluster Conservation

While investigating 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 *Me. pennsylvanicum*. This is also apparent for the cluster 19A of *U. maydis*, in which the absence of several PSEPs in *Me. pennsylvanicum* could be observed, as well as the proliferation of genes encoding PSEPs in *U. maydis* and *S. reilianum* (fig. 6).

Genome Architecture Comparisons

To investigate the compactness of the four genomes, the length of the 5′- and 3′-gene flanking regions was 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 and 920.38 bp for

Table 1
Number of Candidate Pathogenicity Clusters Predicted by the "TDN" and the "3-kb Distance" Approaches

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

[a]TDN approach.
[b]3-kb distance approach.
[c]Clusters not found by the 3-kb distance approach, but predicted by the "TDN" approach.

Table 2
Conservation of Candidate Pathogenicity Clusters within the Four Genomes

	Query Cluster[a]			
Subject Cluster[b]	*Melanopsichium pennsylvanicum*	*Ustilago maydis*	*Sporisorium reilianum*	*U. hordei*
Me. pennsylvanicum	—	13	13	8
U. maydis	13	—	35[c]	23[d]
S. reilianum	13	31	—	22
U. hordei	8	21	24[e]	—
Conserved in all	7	10	9	7
Own clusters	8	12	14	3

[a]Genome which pathogenicity clusters were tested for conservation.
[b]Genome queried for pathogenicity clusters conservation.
[c]Four of the pathogenicity clusters of *S. reilianum* were fragmented, with respect to *U. maydis* clusters.
[d]Two of the pathogenicity clusters of *U. hordei* were fragmented, with respect to *U. maydis* clusters.
[e]Two of the pathogenicity clusters of *S. reilianum* were fragmented, with respect to *U. hordei* clusters.

the 5′- and 3′-distance, respectively, in *S. reilianum* to 1,060.17 and 1,064.38 bp for the 5′- and 3′-distance, respectively, in *Me. pennsylvanicum* (fig. 7*A–D*).

Distances to neighboring genes for the genes encoding PSEPs were also assessed (fig. 7*E–H*). These results showed that the mean of 5′- and 3′-distances of PSEP-encoding genes ranged from 1,004.62 and 1,060.11 bp, respectively, in *S. reilianum* to 1,387.74 and 1,360.53 bp, respectively, in *Me. 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 fig. S4*A–D*, Supplementary Material online) and PSEP-encoding genes (supplementary fig. S4*E–H*, Supplementary Material online) 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 *Me. 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, although most PSEP-encoding genes are not residing in clusters, there is a trend toward 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 unravel the genome sequence of *Me. pennsylvanicum,* a nonobligate biotrophic pathogen of the family Ustilaginaceae. *Melanopsichium pennsylvanicum* is responsible for gall smut on *Polygonum pennsylvanicum* (Halisky and 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; Weiß et al. 2004; Stoll et al. 2005; McTaggart et al. 2012). Extensive comparative studies have been carried out previously on the genomes of *U. maydis*, *S. reilianum*, and *U. hordei* to shed light on their pathogenic behavior, evolution, and genomic makeup (Kämper et al. 2006; Schirawski et al. 2010; Laurie et al. 2012). However, comparative studies with the aim to identify genes that might be required for pathogenicity on grasses or required for pathogenicity in general were 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 *Me. pennsylvanicum*.

Genome assemblies of *Me. 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 *Me. 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 *Me. pennsylvanicum*, the gene space is equally covered in all four genomes.

The genome of *Me. pennsylvanicum* encodes 6,280 genes, which is substantially less than in the other Ustilaginaceae sequenced to date, as *U. hordei* has 7,111 protein-coding genes, whereas *S. reilianum* has 6,673 and *U. maydis* 6,787 protein-coding genes. To cross-check this noticeable result, despite the evidence for good gene-calling in all four genomes

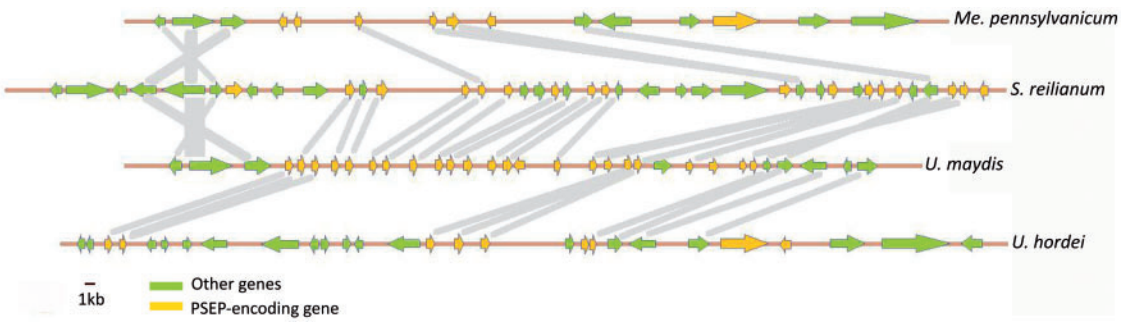

Fig. 6.—*Ustilago maydis* 19A pathogenicity cluster synteny in the other three genomes. Gray lines indicate orthologous genes. Yellow arrows represent PSEPs and green arrows represent nonsecreted protein-encoding genes. The orientation of the arrows represents the orientation of the genes and length of arrows is proportional to the length of the genes.

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. Although a few additional potential genes were found in all four genomes, their amount was comparable and generally low. The lower amount of genes in *Me. 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. Although *Me. pennsylvanicum* contains 7.5% less nonsecreted 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*, *Me. pennsylvanicum* has only 8.4% less nonsecreted proteins, whereas it contains 30.6% fewer PSEPs.

Nonsecreted 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 *Me. pennsylvanicum* is embedded within the *Ustilago* s.l. clades (Stoll et al. 2005; Begerow et al. 2006; McTaggart et al. 2012). As the regulation of the morphology and disease symptoms of smut fungi is still poorly understood, it remains uncertain whether 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 and RNA sequences, have already been reported (Suh and Sugiyama 1993; Begerow et al. 2006), 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 *Ma. globosa*. Even though a sister-group relationship of *U. maydis* and *S. reilianum* was not supported, *U. hordei* and *Me. 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; Schirawski et al. 2010; Kemen and Jones 2012). 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 noneffector proteins (Schirawski et al. 2010; Kemen et al. 2011). 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, *Me. pennsylvanicum* showed the highest percentage of PSEPs under positive selection, 59.5% to more than 2-fold higher than in the other genomes at BEB support greater than 95% and a FDR lower than 1%.

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, such as genome

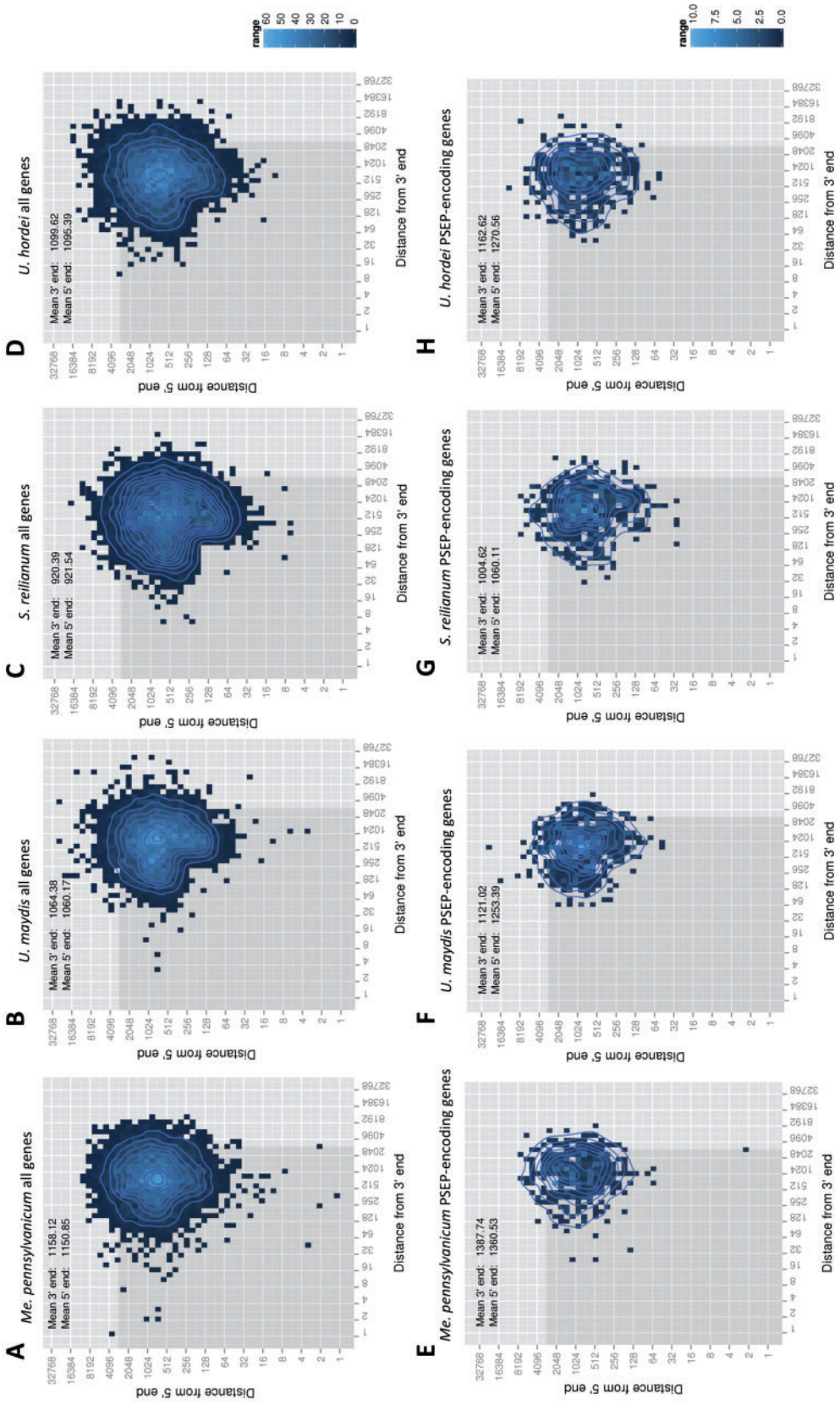

Fig. 7.—Heat-maps representing 3′- and 5′-end gene flanking region lengths. (*A–D*) Heat-maps of 3′- and 5′-end flanking regions of all genes of *Melanopsichium pennsylvanicum*, *Ustilago maydis*, *Sporisorium reilianum*, and *U. hordei*, in the respective order. (*E–H*) Heat-maps of 3′- and 5′-end flanking regions of PSEP-encoding genes of *Me. pennsylvanicum*, *U. maydis*, *S. reilianum*, and *U. hordei*, in the respective order. The 3′- and 5′-distance values less than 3 kb are in the portion of the plot shaded in darker gray.

rearrangements, positive selection, gene losses, and gene gains. Although some comparative genomics studies (Baxter et al. 2010; Raffaele et al. 2010; Kemen et al. 2011) and a few experimental works (Baxter et al. 2010; Dong et al. 2014) have previously been carried out 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 (1858) and 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 *Me. pennsylvanicum* genome has lost 292 genes, which are present in the other three species, whereas 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 *Me. pennsylvanicum* genome, whereas 17, 3, and 2 were absent only in *U. hordei*, *S. reilianum*, and *U. maydis*, respectively. Of 44, 13, 2, and 1 of these genes, no distantly similar genes or gene remnants could be found in the respective other genomes. These results suggest that *Me. 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 *Me. 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 *Me. pennsylvanicum* genome.

Host jump events might also have some effects on the genome architecture of a pathogen. Probably as a result of gene losses, genes of *Me. pennsylvanicum* are farther apart from each other, when compared with 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, where these processes were reported to play important roles (Ma et al. 2010).

Many 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 will not identify clusters that have three or more effector genes in very close vicinity, but with noneffector 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 led to the identification of 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 *Me. pennsylvanicum, U. hordei, S. reilianum*, and *U. maydis,* respectively. Although the other three genomes showed a conservation of clusters in the range of 43–80%, conservation ranged from 20% to 26% in *Me. pennsylvanicum*. This highlights that the genes in pathogenicity clusters are also strongly affected by the host jump. But the fact that only 12.2–31.6% 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 *Me. pennsylvanicum*, is encoded by a gene that is not embedded within a pathogenicity cluster. However, there seems to be some trend toward 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*, whereas *U. maydis* contained more than 2.7-fold more of such occurrences. While initially useful for the elucidation of some general pathogenicity features of the Ustilaginaceae (Kämper et al. 2006; Schirawski et al. 2010; Laurie et al. 2012) it thus seems reasonable to focus more on the nonclustered putative-secreted effector genes in future studies.

Of the genes encoding PSEPs, 57 (supplementary table S7, Supplementary Material online) were found in the three graminicolous species but not in the dicot-infecting *Me. pennsylvanicum.* For 44 of these, no distantly similar gene remnants were found in intergenic regions of *Me. pennsylvanicum*. It seems possible that these PSEPs contain pathogenicity effectors that are of particular importance for the colonization of grass hosts, whereas those 248 orthologous genes (supplementary table S8, Supplementary Material online) 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.

Supplementary Material

Supplementary figures S1–S6 and tables S1–S9 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

The authors thank Claus Weiland for support with respect to cluster access and 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 M.T., and the research funding program LOEWE "Landes-Offensive zur Entwicklung Wissenschaftlich-ökonomischer Exzellenz" of Hesse's Ministry of Higher Education, Research, and the Arts. The authors declare that no competing interest exists. M.T. designed the study. B.M., M.T., and R.S. conceived analyses and provided ideas. R.S. assembled and annotated the genome and carried out all computational analyses on the data. F.R. handled the Mp4 strain and isolated genomic DNA for sequencing. R.S. and M.T. wrote the manuscript, with contributions from B.M. and F.R.

Literature Cited

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

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

Anisimova M, Yang Z. 2007. Multiple hypothesis testing to detect lineages under positive selection that affects only a few sites. Mol Biol Evol. 24: 1219–1228.

Attwood TK, Beck ME, Bleasby AJ, Parry-Smith DJ. 1994. PRINTS—a database of protein motif fingerprints. Nucleic Acids Res. 22:3590–3596.

Barker WC, Pfeiffer F, George DG. 1996. Superfamily classification in PIR-International Protein Sequence Database. Methods Enzymol. 266: 59–71.

Baxter L, et al. 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. Science 330:1549–1551.

Begerow D, Bauer R, Boekhout T. 2000. Phylogenetic placements of ustilaginomycetous anamorphs as deduced from nuclear LSU rDNA sequences. Mycol Res. 104:53–60.

Begerow D, Göker M, Lutz M, Stoll M. 2004. On the evolution of smut fungi on their hosts. In: Agerer R, Blanz P, Piepenbring M, editors. Frontiers in Basidiomycote mycology. München (Germany): IHW-Verlag. p. 81–98.

Begerow D, Stoll M, Bauer R. 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 PR, et al. 2008. Oomycete RXLR effectors: delivery, functional redundancy and durable disease resistance. Curr Opin Plant Biol. 11: 373–379.

Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. 2011. Scaffolding pre-assembled contigs using SSPACE. Bioinformatics 27:578–579.

Buchan DW, et al. 2002. Gene3D: structural assignment for whole genes and genomes using the CATH domain structure database. Genome Res. 12:503–514.

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

Dong S, et al. 2014. Effector specialization in a lineage of the Irish potato famine pathogen. Science 343:552–555.

Darwin C. 1858. On the variation of organic beings in a state of nature; on the natural means of selection; on the comparison of domestic races and true species. J Proc Linnean Soc (Zool) 3:46–50.

de Lima Morais DA, et al. 2011. SUPERFAMILY 1.75 including a domain-centric gene ontology method. Nucleic Acids Res. 39:D427–D434.

Djamei A, et al. 2011. Metabolic priming by a secreted fungal effector. Nature 478:395–398.

Doehlemann G, et al. 2009. Pep1, a secreted effector protein of *Ustilago maydis*, is required for successful invasion of plant cells. PLoS Pathog. 5:e1000290.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32:1792–1797.

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

Fischer GW. 1953. Manual of the North American smut fungi. New York: Ronald Press Co.

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

Gasteiger E, et al. 2003. ExPASy: the proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res. 31:3784–3788.

Grigoriev IV, et al. 2012. The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res. 40:D26–D32.

Haas BJ, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461:393–398.

Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. Nucleic Acids Res. 31:371–373.

Halisky PM, Barbe GD. 1962. A Study of *Melanopsichium pennsylvanicum* causing gall smut on *Polygonum*. Bull Torrey Bot Club. 89:181–186.

Harris MA, et al. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res. 32:D258–D261.

Hirschhorn E. 1941. Una nueva especie de *Melanopsichium*. Notas del Museo de la Plata. Séccion Botánica 6:147–151.

Hunter S, et al. 2009. InterPro: the integrative protein signature database. Nucleic Acids Res. 37:D211–D215.

Jurka J, et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Kamoun S. 2006. A catalogue of the effector secretome of plant pathogenic oomycetes. Annu Rev Phytopathol. 44:41–60.

Kämper J, et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444:97–101.

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

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

Kemen E, et al. 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. PLoS Biol. 9:e1001094.

Kemen E, Jones JD. 2012. Obligate biotroph parasitism: can we link genomes to lifestyles? Trends Plant Sci. 17 448–457.

Kohany O, Gentles AJ, Hankus L, Jurka J. 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics 7:474.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 305:567–580.

Krzywinski M, et al. 2009. Circos: an information aesthetic for comparative genomics. Genome Res. 19:1639–1645.

Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. Genome Biol. 5:R12.

Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods. 9:357–359.

Laurie JD, et al. 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.

Li L, Stoeckert CJ Jr, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189.
Lima T, et al. 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–D478.
Liu Q, Mackey AJ, Roos DS, Pereira FC. 2008. Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. Bioinformatics 24:597–605.
Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 25: 955–964.
Loytynoja A, Goldman N. 2010. webPRANK: a phylogeny-aware multiple sequence aligner with interactive alignment browser. BMC Bioinformatics 11:579.
Ma LJ, et al. 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464:367–373.
Majoros WH, Pertea M, Salzberg SL. 2004. TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics 20: 2878–2879.
McAlpine D. 1910. The smuts of Australia, their structure, life history, treatment, and classification. Melbourne (Vic.): J. Kemp, government printer.
McTaggart AR, Shivas RG, Geering AD, Vanky K, Scharaschkin T. 2012. A review of the complex. Persoonia 29:55–62.
Mi H, et al. 2005. The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res. 33:D284–D288.
Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics 148:929–936.
Ostlund G, et al. 2010. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucleic Acids Res. 38:D196–D203.
Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 23: 1061–1067.
Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8:785–786.
Ploch S, et al. 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.
Price AL, Jones NC, Pevzner PA. 2005. De novo identification of repeat families in large genomes. Bioinformatics 21(Suppl 1), i351–i358.
R Development Core Team. 2008. R: a language and environment for statistical computing. Vienna (Austria): R Foundation for Statistical Computing.
Raffaele S, et al. 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 330:1540–1543.
Schattner P, Brooks AN, Lowe TM. 2005. The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. Nucleic Acids Res. 33:W686–W689.
Schirawski J, et al. 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330:1546–1548.
Schultz J, Milpetz F, Bork P, Ponting CP. 1998. SMART, a simple modular architecture research tool: identification of signaling domains. Proc Natl Acad Sci U S A. 95:5857–5864.
Sigrist CJ, et al. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. 3:265–274.
Slater GS, Birney E. 2005. Automated generation of heuristics for biological sequence comparison. BMC Bioinformatics 6:31.
Sonnhammer EL, Eddy SR, Durbin R. 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.
Stamatakis A, Ludwig T, Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics 21:456–463.
Stanke M, Schoffmann O, Morgenstern B, Waack S. 2006. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 7:62.
Stoll M, Begerow D, Oberwinkler F. 2005. Molecular phylogeny of *Ustilago*, *Sporisorium*, and related taxa based on combined analyses of rDNA sequences. Mycol Res. 109:342–356.
Storey JD. 2002. A direct approach to false discovery rates. J R Stat Soc Series B Stat Methodol. 64:479–498.
Suh SO, Sugiyama J. 1993. Phylogeny among the basidiomycetous yeasts inferred from small subunit ribosomal DNA sequence. J Gen Microbiol. 139:1595–1598.
Tatusov RL, et al. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.
Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. 2008. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 18:1979–1990.
Thompson JD, Gibson TJ, Higgins DG. 2002. Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics, Chapter 2: Unit 2 3.
Tyler BM. 2009. Entering and breaking: virulence effector proteins of oomycete plant pathogens. Cell Microbiol. 11:13–20.
Vánky K. 1994. European smut fungi. Stuttgart: Gustav Fischer Verlag.
Vánky K. 2002. Illustrated genera of smut fungi. St. Paul (MN): American Phytopathological Society.
Wallace AR. 1858. On the tendency of varieties to depart indefinitely from the original type. J Proc Linn Soc Zool. 3:53–62.
Weiß M, Bauer R, Begerow D. 2004. Spotlights on heterobasidiomycetes. In: Agerer R, Blanz P, Piepenbring M, editors. Frontiers in Basidiomycete mycology. München (Germany): IHW-Verlag. p. 7–48.
Win J, et al. 2007. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. Plant Cell 19: 2349–2369.
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.
Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.
Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. Mol Biol Evol. 19:908–917.
Yang Z, Wong WS, Nielsen R. 2005. Bayes empirical Bayes inference of amino acid sites under positive selection. Mol Biol Evol. 22: 1107–1118.
Zdobnov EM, Apweiler R. 2001. InterProScan—an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.
Zerbino DR, Birney E. 2008. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 18:821–829.
Zhang J, Nielsen R, Yang Z. 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Mol Biol Evol. 22:2472–2479.

Associate editor: Cécile Ané