Positively Selected Effector Genes and Their Contribution to Virulence in the Smut Fungus *Sporisorium reilianum*

Gabriel Schweizer¹,⁴, Karin Münch¹, Gertrud Mannhaupt¹,², Jan Schirawski¹,⁵, Regine Kahmann¹,*, and Julien Y. Dutheil¹,³,⁶,*

¹Department of Organismic Interactions, Max Planck Institute for Terrestrial Microbiology, Marburg, Germany
²Institute for Bioinformatics and Systems Biology, Helmholtz Zentrum München, Neuherberg, Germany
³Institute of Evolutionary Sciences of Montpellier, “Genome” Department, CNRS, University of Montpellier 2, France
⁴Present address: Department of Evolutionary Biology and Environmental Studies, University of Zürich, Winterthurerstrasse 190, 8057 Zürich, Switzerland.
⁵Present address: Microbial Genetics, Institute of Applied Microbiology, RWTH Aachen, Aachen, Germany
⁶Present address: Research Group Molecular Systems Evolution, Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Biology, Plön, Germany

*Corresponding authors: E-mails: dutheil@evolbio.mpg.de; kahmann@mpi-marburg.mpg.de.

Accepted: January 29, 2018

Data deposition: Genome sequences generated in this project have been deposited at the European Nucleotide Archive under the accessions LT795054 to LT795076. Supplementary table S1, Supplementary Material online, contains the links toward all publicly available genome sequences used.

Abstract

Plants and fungi display a broad range of interactions in natural and agricultural ecosystems ranging from symbiosis to parasitism. These ecological interactions result in coevolution between genes belonging to different partners. A well-understood example is secreted fungal effector proteins and their host targets, which play an important role in pathogenic interactions. Biotrophic smut fungi (Basidiomycota) are well-suited to investigate the evolution of plant pathogens, because several reference genomes and genetic tools are available for these species. Here, we used the genomes of *Sporisorium reilianum* f. sp. *zeae* and *S. reilianum* f. sp. *reilianum*, two closely related formae speciales infecting maize and sorghum, respectively, together with the genomes of *Ustilago hordei*, *Ustilago maydis*, and *Sporisorium scitamineum* to identify and characterize genes displaying signatures of positive selection. We identified 154 gene families having undergone positive selection during species divergence in at least one lineage, among which 77% were identified in the two investigated formae speciales of *S. reilianum*. Remarkably, only 29% of positively selected genes encode predicted secreted proteins. We assessed the contribution to virulence of nine of these candidate effector genes in *S. reilianum* f. sp. *zeae* by deleting individual genes, including a homologue of the effector gene *pit2* previously characterized in *U. maydis*. Only the *pit2* deletion mutant was found to be strongly reduced in virulence. Additional experiments are required to understand the molecular mechanisms underlying the selection forces acting on the other candidate effector genes, as well as the large fraction of positively selected genes encoding predicted cytoplasmic proteins.

Key words: positive selection, effector evolution, smut fungi, comparative genomics, virulence.

Introduction

Plants and fungi have a long history of coevolution since the emergence of pioneering land plants ~400 Ma. The development of early plants was likely supported by associations with symbiotic fungi, as suggested by analyses of ribosomal RNAs and fossil records (Remy et al. 1994; Gehrig et al. 1996; Martin et al. 2017). Different forms of plant–fungus interactions have evolved, including mutualistic symbiosis where both plant and fungus benefit (Parniske 2008), and pathogenic interactions where fungal colonization greatly reduces...
Secreted fungal effector proteins are key players in pathogenic interactions as they are involved in protecting and shielding growing hyphae, suppressing plant defense responses and changing plant physiology to support growth of the pathogen (Stergiopoulos and de Wit 2009; de Jonge et al. 2011; Giraldo and Valent 2013). Many effector proteins lack known functional domains, and expression of a subset of effectors is linked to plant colonization (Lo Presti et al. 2015; Toruno et al. 2016; Franceschetti et al. 2017; Lanver et al. 2017). Effector proteins with a strong effect on virulence phenotype are thought to coevolve with their plant targets either in an arms race or a trench-warfare scenario (Brown and Tellier 2011). In the former, fungal effectors manipulating the host are under positive directional selection, and plant targets evolve in response to changes in effector proteins (Rovenich et al. 2014). In the latter scenario, sets of alleles are maintained by balancing selection in both host and pathogen populations (Brown and Tellier 2011; Tellier et al. 2014). Several methods are available for identifying genomic regions under selection (Nielsen 2005; Aguileta et al. 2009, 2010). It has been proposed that genes with signatures of positive selection have important functions during host pathogen interaction or have contributed to host specialization (Tiffin and Moeller 2006). It is therefore expected that the deletion of such genes reduces virulence when tested on a susceptible host.

Depending on the aim of the investigation, studies identifying genes with signatures of positive selection are carried out within or between species. Although studies on the population level focus on recent and ongoing selective processes and are instrumental in the understanding of adaptation, comparative genomic studies employing different species encompass a broader time span and provide insight into the underlying genetic basis of host specialization (Plissonneau et al. 2017). The signature of positive selection in such case typically takes the form of an excess of divergence between species due to increased fixation of mutations by selective sweeps compared with a neutral expectation (Yang and Nielsen 1998). This is commonly measured by the ratio of nonsynonymous (dN) over synonymous (dS) divergence and a dN/dS ratio > 1 is taken as evidence for positive selection under the assumption that synonymous substitutions are neutral while nonsynonymous are not. Positive selection studies in a number of plant pathogen systems revealed that genes encoding secreted effector proteins are enriched in signatures of positive selection (Möller and Stukenbrock 2017). Such studies include investigations in diverse plant pathogens like Microbotryum species causing anther-smut disease of Caryophyllaceae species (Aguileta et al. 2010), the wheat pathogen Zymoseptoria tritici (Stukenbrock et al. 2011), the rust fungus Melampsora lari-culinaris (Haccquard et al. 2012), the rice BLAST fungus Magnaporthe oryzae (Huang et al. 2014), the wheat stem rust fungus Puccinia graminis f. sp. tritici (Sperschneider et al. 2014), the Irish potato famine pathogen Phytophthora infestans (Dong et al. 2014), a group of Fusarium species (Sperschneider et al. 2015), and a group of smut fungi parasitizing different grasses and a dicot host (Sharma et al. 2014, 2015). Yet, as only a few genes under positive selection have been functionally studied, the link between the selected genotypes and their corresponding phenotypes are only beginning to be understood and only a few studies used evolutionary predictions to unravel the molecular mechanisms of host adaptation. For example, the population genomics study in the wheat pathogen Z. tritici which identified candidate effector genes under positive selection (Stukenbrock et al. 2011) was followed up experimentally, and in this case it was shown that the deletion of three of four candidate genes reduced virulence (Poppe et al. 2015). In the gray mold fungus Botrytis cinerea, four positively selected genes were deleted without affecting virulence, and this finding was attributed to functional redundancy, the limited number of tested host plants, or experimental conditions different from natural infections (Aguileta 2012). A study of the oomycete effector protein Epic1 showed that a single amino acid substitution at a site under positive selection affected the binding affinity of different host proteases determining host specificity (Dong et al. 2014).

Smut fungi, belonging to the division of Basidiomycota, are a group of about 550 species parasitizing mostly grasses, including important crops like maize, sorghum, oat, barley, and sugarcane (Begerow et al. 2014). In smut fungi, sexual reproduction is linked to pathogenic development and smut fungi therefore depend on successful plant colonization to complete their life cycle. As biotrophic pathogens, they require living plant tissue for establishing a successful interaction (Martinez-Espinoza et al. 2002). With few exceptions like Ustilago maydis, smut fungi usually develop symptoms only in the female or male inflorescence of their respective host plants. During the last 10 years, quality draft genome sequences of prominent species were obtained, including U. maydis, the causative agent of smut disease on maize and teosinte (Kämper et al. 2006), Sporisorium reilianum causing head smut of maize and sorghum (Schirawski et al. 2010), Ustilago hordei infecting barley (Laurie et al. 2012), and Speronosum scitamineum parasitizing sugarcane (Que et al. 2014; Taniguti et al. 2015; Dutheil et al. 2016). The head smut fungus S. reilianum occurs in two formae speciales that infect maize (S. reilianum f. sp. zeae) or sorghum (S. reilianum f. sp. reilianum) (Zuther et al. 2012). The concept of formae speciales is used in phytopathology to distinguish members of the same species based on their ability to colonize a certain host plant (in this example maize or sorghum).
positive selection for the other candidate genes.

Candidate genes were individually deleted in initial effector genes with signatures of positive selection. S. scitamineum (Munkacsi et al. 2007) together with the genomes of S. reilianum (Khrunyk et al. 2010; Schuster et al. 2016) make this group manipulations (Brachmann et al. 2004; Kämper 2004; Khrunyk et al. 2010; Schuster et al. 2016) to study the evolution of effector genes as well as their contributions to virulence, specialization, and host specificity.

Here, we employed the genome of the recently sequenced strain S. reilianum f. sp. reilianum SRS1_H2-8 (http://www.ebi.ac.uk/ena/data/view/L1795054-L1795076) (Zuther et al. 2012) together with the genomes of U. maydis, U. hordei, S. scitamineum, and S. reilianum f. sp. zeae to identify potential effector genes with signatures of positive selection. Candidate genes were individually deleted in S. reilianum f. sp. zeae and the phenotype of the deletion strains was assessed after infection of maize in order to understand their function with respect to virulence. We report that the deletion of one candidate gene, pit2, led to a strong reduction in virulence and we further discuss hypotheses on the origin of positive selection for the other candidate genes.

Materials and Methods

Construction of Homologous Protein Families

Fungal species used in this study, their number of gene models, number of predicted secreted proteins, and sources of genome data are listed in supplementary table S1, Supplementary Material online. The predicted proteome of the five smut fungi U. hordei, U. maydis, S. scitamineum, S. reilianum f. sp. zeae, and S. reilianum f. sp. reilianum were used to perform an all-against-all BLASTp search (Altschul et al. 1990). The SiLiX algorithm was subsequently used to infer homology relationships based on the BLAST hits (Altschul et al. 1990). The SiLiX algorithm was subsequently used to infer homology relationships based on the BLAST hits (Miele et al. 2011). Two parameters are considered to decide whether a BLAST hit can be taken as evidence for homology: the percent identity between two sequences and the relative coverage. In order to maximize the number of families comprising 1:1 orthologues (that is families that have an equal number of members in each species), SiLiX (Miele et al. 2011) was run with a range for coverage and identity thresholds between 5% and 95% in 5% steps. An identity of 40% and coverage between 5% and 45% lead to the maximum number of families with 1:1 orthologues (5,394; supplementary fig. S1, Supplementary Material online) while settings with 40% identity and 80% coverage lead to 5,326 families with 1:1 orthologues (supplementary fig. S1, Supplementary Material online). Since using a higher coverage had only a cost of 68 core families, the stricter criteria were applied for family clustering. Families with at least two members were aligned on the codon level using MACSE 1.01b (Ranwez et al. 2011) and on the protein level using PRANK v.100802 (Löytynoja and Goldman 2008). The resulting alignments were subsequently compared and column scores (CS) computed for each position in the alignment (Thompson et al. 1999). Only positions with CS of 100% (that is, alignment columns identically found by both methods) and a maximum of 30% gaps were retained for further analysis.

Estimation of Genome-Wide Divergence Values and Divergence Times

The five genomes of U. hordei, U. maydis, S. scitamineum, S. reilianum f. sp. zeae, and S. reilianum f. sp. reilianum were aligned using the Multiz genome aligner from the TBA package (Blanchette et al. 2004) and projected on the U. maydis genome as reference. The resulting multiple genome alignment had a total size of 21 Mb and was further restricted to regions with homologous sequences in the five species (total length after this step: 14.3 Mb) and processed to remove coding regions. The final non-coding sequence alignment had a total length of 2.2 Mb, for which pairwise nucleotide similarities were computed in nonoverlapping windows of 10 kb.

Gene families with exactly one member in each species were concatenated and pairwise protein sequence similarities computed using the seqinr package for R (Charif and Loby 2007). Protein alignments were also used to infer dates of divergence, using a relaxed clock model. The PhyloBayes version 4.1 software (Lartillot et al. 2009) was used with the auto-correlated model of Thorne et al. (Thorne et al. 1998) under a GTR + CAT model. A unique calibration point was used, based on the divergence time of the most divergent lineage U. hordei, previously estimated to have occurred between 27 and 21 Myr (Bakkeren and Kronstad 2007). A uniform prior was used on this interval for the Monte-Carlo Markov Chain. As convergence issues arise when large alignments (>20,000 positions) are used, we followed the PhyloBayes authors’ recommendation to conduct a jackknife procedure. We generated three data sets of circa 20,000 amino acids by randomly sampling families and concatenating the corresponding alignments. Two chains were run in each case and convergence was assessed. Sampling was performed after a burning of 10,000 iterations, and every ten subsequent iterations. Chains were run to ensure that the minimum effective sample size was >50 and maximum relative difference <0.3 in at least one sample. Results are summarized in supplementary table S2, Supplementary Material online, and figure S2, Supplementary Material online, shows the six chains for the three samples. In addition to the convergence of the two chains for each sample, our results reveal extremely consistent results between samples. Figure 1A...
shows estimates from one chain of the third data set, which shows a minimum effective sample size >300.

Detection of Positive Selection

For gene families with at least three members, translated sequences were employed to create maximum likelihood phylogenetic trees using PhyML 3.0 (Guindon et al. 2010) with a minimum parsimony starting tree and the LG amino acid substitution model with a four-classes gamma distribution of site-specific substitution rate (Le and Gascuel 2008). The best tree topology obtained from nearest neighbor interchange (NNI) and subtree pruning and regrafting (SPR) searches was kept (Guindon et al. 2010). BppML (Dutheil and Boussau 2008) was then used to re-estimate branch lengths from the codon alignment using the YN98 substitution model (Nielsen and Yang 1998). We next aimed at inferring the occurrence of positive selection for each gene family. This is typically achieved by measuring the ratio of nonsynonymous versus synonymous substitutions ($d_N/d_S$ ratio) using models of codon sequence evolution (Yang 2006). In particular, nonhomogeneous models of sequence evolution estimate the $d_N/d_S$ ratio independently in different lineages, yet at the cost of potential overparametrization issues. In the manual of the PAML package, the authors state that such models should only be used for hypothesis testing and advise against using them for scans of positive selection. Dutheil et al. (2012) proposed a model selection approach (implemented in the TestNH package) allowing to select for the best nonhomogeneous model supported by the data. They start by fitting the simplest (homogeneous) model and sequentially add parameters to model variation of selective regime among lineages. Because the number of possible models is large even for small data sets, two heuristic approaches have been introduced: the “free”

![Phylogenetic tree](image)

**Fig. 1.**—Phylogeny, divergence estimates, and number of genes under positive selection in five related smut fungal species parasitizing different host plants. (A) Chronogram of the five fungal pathogens as estimated under a relaxed molecular clock. Boxes represent 95% posterior intervals, with corresponding values indicated below. (B) Pairwise sequence differences, for both the noncoding genome and the proteome (nonsynonymous differences). (C) Number of positively selected genes on each terminal branch (total number of genes and genes predicted to encode a secreted protein).
heuristic permits unconnected branches from the tree to evolve under the same regime, whereas the “joint” heuristic restricts model sharing to connected branches (see Dutheil et al. 2012 for details). The choice of models to test is guided by statistics on the patterns of substitutions on the phylogenetic tree, an approach named substitution mapping (Romiguier et al. 2012). Apart from the model selection approach, the underlying models of codon sequence evolution are identical to the one originally described by Yang (Yang 1998; Yang and Nielsen 1998). Model selection was performed with the TestNH software, which contains two programs: 1) MapNH (Romiguier et al. 2012) was used for mapping substitutions on the previously inferred phylogenetic tree and 2) PartNH (Dutheil et al. 2012) was subsequently employed to fit time nonhomogeneous models of codon substitutions. PartNH uses the previously inferred substitution maps in order to perform model comparisons and select a nonhomogeneous model with minimal number of parameters. Both methods “free” and “join” were compared in order to scan for positive selection. Finally, putative secreted effector proteins were identified by predicting secretion using SignalP 4.0 (Petersen et al. 2011) and proteins were considered as secreted if the program indicated the presence of a signal peptide but no transmembrane domain. To detect residues under positive selection in homologues of pit2, the branch-site model with Bayes Empirical Bayes (BEB) analysis as implemented in PAML4 (Yang 2007) was applied. We employed information about family composition, alignment, and phylogeny as outlined above and defined sr10529 and srS_10529 as foreground branches. A posterior probability threshold of >95% was used for the BEB analysis.

Association of Positively Selected Genes with Repeats in U. hordei

We tested whether genes under positive selection are located significantly closer to repetitive elements than average genes in the genome of U. hordei, which shows the highest content of repetitive elements in the group of smut fungi investigated here. For this analysis, only a group of “uncharacterized interspersed repeats” was investigated, because it was shown previously that this is the only category showing a strong association with candidate effector genes (Dutheil et al. 2016). Binary logistic regressions were conducted in R using the rms package (Harrell 2015). The “robcov” function of the rms package was used in order to get robust estimates of each effect. The variable “distance to the closest interspersed repeat” was transformed by \(\log(x + 1)\) because of its extreme distribution.

Comparing \(d_\omega/d_\sigma\) Ratios of Genes Residing in Virulence Clusters

Previous work has identified several virulence gene clusters in U. maydis and some of them play important roles during pathogenic development (Kämper et al. 2006; Schirawski et al. 2010). In total, these clusters contain 163 genes, where 100 reside in clusters without virulence phenotype and 63 reside in clusters with virulence phenotype upon deletion. Both types of clusters contain each 32 genes for which a \(d_\omega/d_\sigma\) ratio could be determined (the missing genes are part of families that do not have at least three members and were therefore not analyzed). The \(d_\omega/d_\sigma\) ratios of all genes in clusters were compared between clusters with and without virulence phenotype (Wilcoxon rank-sum test).

Gene Ontology Terms Enrichment Analysis

All proteins in S. reilianum f. sp. zeae, in S. reilianum f. sp. reilianum, and in U. hordei were considered for Gene Ontology (GO) term enrichment analyses. GO terms were assigned using iprscan 1.1.0 (http://fgblab.org/runiprscan; developed by Michael R. Thon) which links GO information provided by Interpro to each protein. In this way, 1,759 unique GO terms could be assigned to 4,130 proteins in S. reilianum f. sp. zeae, 1,744 unique GO terms could be assigned to 4,124 proteins in S. reilianum f. sp. reilianum, and 1,757 unique GO terms could be assigned to 3,922 proteins in U. hordei (supplementary table S3, Supplementary Material online). The Bioconductor package topGO (Alexa et al. 2006) was then used to link each GO term to the three major categories “Cellular Component,” “Biological Process,” or “Molecular Function.” Enrichment analysis was performed by computing \(P\) values for each GO term using Fisher’s classic test with parent–child correction (Grossmann et al. 2007). Cytoplasmic proteins with and without signatures of positive selection were compared for the three species separately, and differences were considered to be significant at the 5% level.

Strains and Growth Conditions

The Escherichia coli derivative Top10 (Invitrogen, Karlsruhe, Germany) and the Saccharomyces cerevisiae strain BY4741 (MATa his3Δ1 leu2Δ met15Δ ura3Δ; Euroscarf, Frankfurt, Germany; kindly provided by M. Bölk, Marburg) were used for cloning purposes. Soporisorum reilianum f. sp. zeae strains used in this study are listed in supplementary table S4, Supplementary Material online. They are derivatives of the haploid solopathogenic strain JS161 which is capable of plant colonization without the need of a mating partner, because it expresses a compatible pheromone/receptor pair (Schirawski et al. 2010). Escherichia coli was grown in dYT liquid medium (1.6% [w/v] Trypton, 1.0% [w/v] Yeast Extract [Difco (R)], 0.5% [w/v] NaCl), or YT solid medium (0.8% [w/v] Trypton, 0.5% [w/v] Yeast Extract, 0.5% [w/v] NaCl, 1.3% [w/v] agar) supplemented with 100 mg/mL Ampicillin when needed. The yeast S. cerevisiae was maintained in YPD solid medium (1% [w/v] yeast extract, 2% [w/v] (R)-Pepton, 2% [w/v] (R)-Agar, 2% [w/v] glucose) and grown on SC URA- medium (1.7% [w/v] Yeast Nitrogen Base without ammonium sulfate,
0.147% [w/v] dropout-mix without Uracil, 2% [w/v] glucose) for selecting transformants containing the plasmid pRS426 (Sikorski and Hieter 1989) (kindly provided by M. Bölker, Marburg) or derivatives of pRS426. Strains of S. reilianum were grown in liquid YEPSlight medium (1.0% [w/v] yeast extract, 0.4% [w/v] peptone, 0.4% [w/v] sucrose) at 28°C on a rotary shaker at 200 rpm.

Construction of S. reilianum Strains

Polymerase chain reactions were performed using the Phusion High-Fidelity DNA Polymerase (New England Biolabs). Templates were either JS161 genomic or indicated plasmid DNA. Restriction enzymes were obtained from New England Biolabs. Protoplast-mediated transformation was used to transform S. reilianum following a method established for U. maydis (Schulz et al. 1990). Transformants were selected on RegAgar plates (1.0% [w/v] yeast extract, 0.4% [w/v] Bacto(R)-Pepton, 0.4% [w/v] Sucrose, 1 M Sorbitol, 1.5% [w/v] Bacto(R)-Agar) supplemented with 200 µg/mL Geneticin and true resistance was tested by growing single colonies on PD plates (3.9% [w/v] Potato-Dextrose Agar, 1% [w/v] Tris–HCl [1 M, pH 8.0]) supplemented with 50 µg/mL Geneticin. Gene replacements with resistance markers were generated with a PCR-based method employing the previously described SfiI insertion cassette system (Brachmann et al. 2004; Kämper 2004) and were confirmed by Southern blot analysis. Genomic regions residing ~1-kb upstream (left border) or downstream (right border) adjacent to open reading frames of candidate genes were PCR-amplified using the listed primer pairs (supplementary table S5, Supplementary Material online) and genomic DNA of JS161 as template. The resulting fragments were used for cloning plasmids containing the respective deletion constructs.

To obtain deletion constructs for the genes sr10529 and sr14347, PCR fragments containing the left and right borders of each gene were ligated to the hygromycin resistance cassette by ligation via SfiI restriction sites, yielding plasmids pRS426 Δsr10529 Gen #1, pRS426 Δsr14347 Gen #1, pRS426 Δsr10529 Gen #2, pRS426 Δsr14347 Gen #1, pRS426 Δsr10182 Gen #1, pRS426 Δsr10182 Gen #1, pRS426 Δsr14558 Gen #1, and pRS426 Δsr12897 Gen #5, respectively. Gene deletion constructs were PCR-amplified from the respective plasmid using listed primers (supplementary table S5, Supplementary Material online). The obtained deletion constructs were used to transform the S. reilianum strain JS161 to generate the gene deletion strains JS161 Δsr12968, JS161 Δsr14944, JS161 Δsr10059, JS161 Δsr10182, JS161 Δsr14558, and JS161 Δsr12897, respectively.

The drag and drop cloning method in yeast (Jansen et al. 2005) was used to generate plasmids pRS426 Δsr12968 Hyg #1, pRS426 Δsr14944 Hyg #2, pRS426 Δsr10059 Hyg #1, pRS426 Δsr10182 Hyg #1, pRS426 Δsr14558 Hyg #1, and pRS426 Δsr12897 Hyg #1 which contain deletion constructs for deleting the candidate genes sr12968, sr14944, sr10059, sr10182, sr14558, or sr12897. These plasmids are a derivate of plasmid pRS426, which can be maintained in E. coli and S. cerevisiae (Sikorski and Hieter 1989). PCR-amplified left and right borders of each candidate gene and the hygromycin resistance cassette were integrated in pRS426 by homologous recombination in S. cerevisiae. Subsequently, the hygromycin resistance cassette was replaced with the Geneticin resistance cassette by ligation via SfiI restriction sites, yielding plasmids pRS426 Δsr12968 Gen #1, pRS426 Δsr14944 Gen #3, pRS426 Δsr10059 Gen #1, pRS426 Δsr10182 Gen #1, pRS426 Δsr14558 Gen #1, and pRS426 Δsr12897 Gen #5, respectively. The gene deletion constructs were used to transform the S. reilianum strain JS161 to generate the gene deletion strains JS161 Δsr12968, JS161 Δsr14944, JS161 Δsr10059, JS161 Δsr10182, JS161 Δsr14558, and JS161 Δsr12897, respectively.

Virulence Assays

The solopathogenic strain JS161 and deletion mutants thereof were grown in YEPlight liquid medium to an optical density at 600 nm (OD600) of 0.8–1.0 and cell cultures were adjusted to an OD600 of 1.0 with sterile water prior to injection into 1-week-old maize (Zea mays) seedlings of the dwarf cultivar “Gaspe Flint” (kindly provided by B. Burr, Brookhaven National Laboratories and maintained by self-pollination). Plants were sowed in T-type soil of “Frustorfer Pikiererde” (HAWITA, Vechta, Germany) and grown in a temperature-controlled greenhouse (14 h/10 h-light/dark cycle, with 28/20°C and 25,000–90,000 lux during the light period). Virulence symptoms were scored 9–10-week postinfection according to previously described symptoms (Ghareeb et al. 2011) and the following categories were distinguished: the plant did not develop ears, the plant developed healthy ears.
shorter or equal to 1 cm or the plant developed healthy ears longer than 1 cm, the plant developed spiky ears, phylloyd in ears or phylloyd in tassels. Spore formation was only observed occasionally, and rarely the plant died due to the infection. Three independent infections were carried out per strain, and at least three independent deletion strains were tested for virulence.

Results
Candidate Effector Genes Are Less Conserved between Species Compared with Other Genes

We reconstructed families of homologous genes for the five smut fungi *U. hordei*, *U. maydis*, *S. scitamineum*, *S. reilianum* f. sp. *zeae*, and *S. reilianum* f. sp. *reilianum* using the SiLiX clustering algorithm (Miele et al. 2011). We optimized the clustering parameters to maximize the occurrence of orthologues and minimize the number of paralogues within each family. In this way, we were able to reconstruct 8,761 families, among which 5,266 had at least one gene in each species (supplementary table S6, Supplementary Material online). As a consequence, we found at least one homologous sequence in four species for 78% of all genes. About 5,254 gene families are found to have exactly one member in each species and were therefore taken as true orthologues (referred to as “core orthologous set” in the following). Considering that secreted proteins are putative effectors, we used SignalP (Petersen et al. 2011) to predict secretion of the encoded protein for each gene (supplementary table S6, Supplementary Material online). We report that 920 (11%) families contained only genes encoding a predicted secreted protein, whereas 7,657 (87%) contained only genes encoding a protein not predicted to be secreted. The remaining 184 (2%) families contained both predicted secreted and cytoplasmic proteins (supplementary table S6, Supplementary Material online). The occurrence of families with both secreted and cytoplasmic proteins can be explained by 1) false negative predictions for secretion, as truncated C-terminal sequences were not removed from the data set, 2) wrong gene annotations, or 3) gain or loss of a secretion signal peptide during effector evolution (Poppe et al. 2015). Among all predicted secreted proteins, 52% have at least one orthologue in all other species, which is significantly less than the global 78% proportion for all proteins ($\chi^2$ test, $P$ value < 2.2 × 10$^{-16}$). Genes encoding putative effector proteins are therefore less conserved across species than other genes, either because their sequence is evolving faster, preventing the recovery of homologous relationships, or because effector genes are created or lost at a higher rate. In *U. hordei*, we observe several species-specific family expansions. There were 17 families which encompassed 5–25 members, but no orthologue in other species (supplementary table S6, Supplementary Material online). Moreover, we identified three families with up to 62 members in *U. hordei*, but only one member in up to three of the other species (supplementary table S6, Supplementary Material online). Gene duplications in *U. hordei* have been hypothesized to be driven by mobile elements (Laurie et al. 2012).

The Genomes of *S. reilianum* f. sp. *zeae* and *S. reilianum* f. sp. *reilianum* Diverged around 1 Ma

To establish a frame for our comparative analysis, we first calculated sequence similarity of the five smut fungi for non-coding intergenic and protein sequences. In addition, we estimated divergence times by performing a molecular dating analysis based on the core orthologues set of the five pathogens, using advanced models of protein sequence evolution and Bayesian inference as implemented in the PhyloBayes package (Lartillot et al. 2009). As calibration point, we used the divergence time of *U. hordei* and *U. maydis*, previously estimated to be between 27 and 21 Myr (Bakkeren and Kronstad 2007). In alignable intergenic regions *U. hordei* shares 57% identity with *S. reilianum* f. sp. *reilianum* and 77% identity in protein sequences (fig. 1B). Monte-Carlo Markov chains were run for three independent gene samples totaling >20,000 amino acid positions each, and two chains were run in each case to assess convergence. The resulting posterior distribution of divergence times were used to infer 95% posterior intervals. The split between *U. maydis* and the *Sporisorium* species was estimated to have occurred ~20 Myr ago (95% posterior interval 25–12 Myr; fig. 1A and supplementary table S2, Supplementary Material online). *Sporisorium reilianum* f. sp. *reilianum* shares 61% nucleotide identity in alignable intergenic regions with *U. maydis*, and 79% sequence identity at the protein level (fig. 1A). The divergence times of *S. scitamineum* and the two formae speciales of *S. reilianum* were calculated to be 13 Myr ago (95% posterior interval 19–7 Myr; fig. 1A and supplementary table S2, Supplementary Material online), which is consistent with the mean divergence estimated between the hosts sorghum and sugarcane (10 Myr with a posterior interval of 8–13 Myr, average over eight studies, source: timetree.org; Kumar et al. 2017). *Sporisorium reilianum* f. sp. *reilianum* and *S. scitamineum* share 74% noncoding nucleotide identity and 88% identity at the protein level (fig. 1B). Finally, the two *S. reilianum* strains diverged 1.1 Myr ago (95% posterior interval 2.4–0.4 Myr; fig. 1A) and share 98% noncoding nucleotide identity and 99% protein identity (fig. 1B). We note that the estimation of this divergence date varied with the gene set used, and was in some cases found to be older (1.7 Myr, with a 95% posterior interval of 4.7–0.6 Myr, see supplementary table S2, Supplementary Material online). The comparison of the five smut genomes therefore encompasses a broad evolutionary time, and the divergence times obtained are compatible with previous estimates from smaller data sets (Munkacsi et al. 2007). The speciation times of the investigated smut species largely predate the 10,000 years of crop plant domestication, which implies that adaptation to the agricultural host, if any,
To detect positive selection, 6,205 families with at least three members (orthologues and/or paralogues, see supplementary table S6, Supplementary Material online) were, regardless of their species composition, aligned on the codon and amino acid level and a phylogenetic tree was inferred. Obtaining accurate alignments is critical for detecting positive selection since alignment errors frequently inflate the false discovery rate (Schneider et al. 2009; Jordan and Goldman 2012). We therefore developed a stringent bioinformatics pipeline for the filtering of sequence alignments by masking ambiguous alignment positions for further analysis (see Materials and Methods). To scan for positive selection, we employed a nonhomogeneous model of sequence evolution allowing \( d_{N}/d_{S} \) ratios to vary along the phylogeny, in combination with two heuristic model selection procedures to avoid overparametrization issues (Nielsen and Yang 1998; Dutheil et al. 2012). Model parameters could not be fitted by either one of the two methods in 1.7% of branches. The two model selection procedures led to highly consistent estimates of branch-specific \( d_{N}/d_{S} \) ratios (Spearman’s rank correlation coefficient equal to 0.85, \( P < 2.2 \times 10^{-16} \)). The distribution of \( d_{N}/d_{S} \) was highly skewed with a median value of 0.06, demonstrating the strong predominance of purifying selection throughout lineages and genes. The mean value of \( d_{N}/d_{S} \) ratios for lineages undergoing positive selection (\( d_{N}/d_{S} > 1 \)) was 4.1 (median 1.9). Although a \( d_{N}/d_{S} \) ratio \( > 1 \) is indicative of positive selection, the absolute value of the ratio is a poor indicator of the strength of undergoing selection. In particular, high ratio values can be obtained because of low \( d_{S} \) values, and the \( d_{N} \) rate might include neutral substitutions, such as nonsynonymous substitution that are conservative regarding certain biochemical properties of the amino acids involved (Sainudiin et al. 2005). The largest number of genes with signs of positive selection was found in \( S. \ reilianum \) f. sp. zeae (84 genes, of which 25 encode predicted secreted proteins) and \( S. \ reilianum \) f. sp. \( S. \ scitamineum \) (111 genes of which 27 encode predicted secreted proteins) (fig. 1C). In addition, a substantial number of positively selected candidate genes was also found in \( U. \ hordei \) (49, and of these, 22 genes are predicted to code for secreted proteins), but only very few in \( U. \ maydis \) (2 genes) and \( S. \ scitamineum \) (7 genes) (fig. 1C). A list of all proteins with their associated \( d_{N}/d_{S} \) ratios in each species is provided in supplementary table S3, Supplementary Material online. Predicted secreted proteins were significantly enriched in the group of proteins under positive selection in \( U. \ hordei \) and in the two investigated formae speciales of \( S. \ reilianum \) (\( P < 10^{-5} \); Fisher’s exact test). This corroborates results of earlier studies in other pathosystems that showed that predicted secreted proteins are often under positive selection, which can be attributed to their direct interaction with host proteins (Joly et al. 2010; Wicker et al. 2013; Poppe et al. 2015). Notably, all genes found under positive selection in the two strains of \( S. \ reilianum \), in \( S. \ scitamineum \) and in \( U. \ maydis \) share orthologous genes in the other species (supplementary tables S3 and S6, Supplementary Material online). In contrast, genes with signs of positive selection in \( U. \ hordei \) belong largely (36 out of 49 genes) to families showing species-specific expansions (supplementary tables S3 and S6, Supplementary Material online).

**Genes Under Positive Selection in** \( U. \ hordei \) **Are Associated with Uncharacterized Interspersed Repeats**

Among the species compared here, the genome of \( U. \ hordei \) shows the highest fraction of repetitive elements (Laurie et al. 2012; Dutheil et al. 2016). Such elements are known to contribute to gene family expansions (Kazazian 2004), and have been suggested to contribute to adaptation by providing advantageous mutations, for instance by repeat-induced point mutations (RIP) leakage which was revealed in a species complex of \( Leptosphaeria \) (Rouxel et al. 2011; Grandaubert et al. 2014). As sequence signatures of RIP were found in LTR elements of \( U. \ hordei \) (Laurie et al. 2012), we tested whether genes under positive selection in \( U. \ hordei \) are physically associated with repetitive elements. We performed a binary logistic regression with the prediction of positive selection as a response variable (that is, whether the underlying branch has a \( d_{N}/d_{S} \) ratio \( > 1 \)) and we considered three putative explanatory variables for each analyzed gene: 1) whether the gene is predicted to encode a secreted protein, 2) whether the gene is duplicated, and 3) the distance of the gene to the closest interspersed repeat. The complete linear model explains 50% of the observed variance, and the three explanatory variables are all significant at the 0.1% level (supplementary table S7, Supplementary Material online). These results suggest that positively selected genes in \( U. \ hordei \) are associated with duplication events, and positive selection is more likely to occur at genes encoding putative effectors. In addition, the proximity of interspersed repeats increases the odds of positive selection, independently of the two other effects, and is confirmed by a stratification approach: the effect still holds when only duplicated genes are considered, or only genes encoding a secreted protein, or the combination of the two (supplementary table S7, Supplementary Material online). This finding corroborates previous results obtained in other microbial plant pathogens where it was described that effector genes tend to localize in repeat rich regions and where it was suggested that such regions contribute to the rapid evolution of effector genes (Raffaele and Kamoun 2012).
Positively Selected Genes Encoding Cytoplasmic Proteins in *S. reilianum* and *U. hordei*

Although we expect effector genes to be under positive selection, we find that the majority of positively selected genes in *S. reilianum* encode cytoplasmic proteins (fig. 1C). To assess the putative functional role of these genes, we performed a Gene Ontology term enrichment analysis, comparing cytoplasmic proteins under positive selection to cytoplasmic proteins not under positive selection (table 1). This analysis revealed that genes with a potential role in metabolic processes, like sulfur compound metabolism, molybdopterin cofactor metabolic process, RNA metabolic process, organic cyclic compound metabolic process, and oxidoreductase activity, as well as responses to starvation and extracellular stimuli are significantly overrepresented at the 5% level (Fisher's classic test with Parent–Child correction; see table 1). This could indicate that cytoplasmic proteins under positive selection contribute to metabolic changes which might be needed to survive with the limited nutrients available on the surface or in the biotrophic interface of different host plants. A similar analysis for cytoplasmic proteins under positive selection in *U. hordei* was conducted, but only led to top-level categories (DNA integration, DNA metabolic process, and isomerase activity; see table 1).

**Table 1** Gene Ontology Terms Significantly Overrepresented in Positively Selected Genes Encoding Cytoplasmic Proteins in *Sporisorium reilianum* and *Ustilago hordei*

<table>
<thead>
<tr>
<th>Gene Ontology ID</th>
<th>Gene Ontology Description</th>
<th>Category</th>
<th>P values*</th>
<th>Species†</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO: 0030532</td>
<td>Small nuclear ribonucleoprotein complex</td>
<td>CC</td>
<td>0.023</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0045263</td>
<td>Proton-transporting ATP synthase complex, coupling factor F(o)</td>
<td>CC</td>
<td>0.038 and 0.036</td>
<td>Sr and Srz</td>
</tr>
<tr>
<td>GO: 0044225</td>
<td>Membrane part</td>
<td>CC</td>
<td>0.047</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0031668</td>
<td>Cellular response to extracellular stimulus</td>
<td>BP</td>
<td>0.015 and 0.013</td>
<td>Sr and Srz</td>
</tr>
<tr>
<td>GO: 0051186</td>
<td>Cofactor metabolic process</td>
<td>BP</td>
<td>0.015</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0042594</td>
<td>Response to starvation</td>
<td>BP</td>
<td>0.017 and 0.012</td>
<td>Sr and Srz</td>
</tr>
<tr>
<td>GO: 0009267</td>
<td>Cellular response to starvation</td>
<td>BP</td>
<td>0.020 and 0.014</td>
<td>Sr and Srz</td>
</tr>
<tr>
<td>GO: 0006790</td>
<td>Sulfur compound metabolic process</td>
<td>BP</td>
<td>0.022</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0051301</td>
<td>Cell division</td>
<td>BP</td>
<td>0.023</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0007777</td>
<td>Mo-molybdopterin cofactor biosynthetic process</td>
<td>BP</td>
<td>0.027</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0009605</td>
<td>Response to external stimulus</td>
<td>BP</td>
<td>0.027 and 0.012</td>
<td>Sr and Srz</td>
</tr>
<tr>
<td>GO: 0043545</td>
<td>Molybdopterin cofactor metabolic process</td>
<td>BP</td>
<td>0.034</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0043413</td>
<td>Macromolecule glycosylation</td>
<td>BP</td>
<td>0.036</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0022402</td>
<td>Cell cycle process</td>
<td>BP</td>
<td>0.037</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0051189</td>
<td>Prosthetic group metabolic process</td>
<td>BP</td>
<td>0.039</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0006139</td>
<td>Nucleobase-containing compound metabolic process</td>
<td>BP</td>
<td>0.012</td>
<td>Srz</td>
</tr>
<tr>
<td>GO: 0016070</td>
<td>RNA metabolic process</td>
<td>BP</td>
<td>0.023</td>
<td>Srz</td>
</tr>
<tr>
<td>GO: 1901360</td>
<td>Organic cyclic compound metabolic process</td>
<td>BP</td>
<td>0.016</td>
<td>Srz</td>
</tr>
<tr>
<td>GO: 0006725</td>
<td>Cellular aromatic compound metabolic process</td>
<td>BP</td>
<td>0.024</td>
<td>Srz</td>
</tr>
<tr>
<td>GO: 0046483</td>
<td>Heterocycle metabolic process</td>
<td>BP</td>
<td>0.023</td>
<td>Srz</td>
</tr>
<tr>
<td>GO: 0035383</td>
<td>Thioester metabolic process</td>
<td>BP</td>
<td>0.041</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 1902589</td>
<td>Single-organism organelle organization</td>
<td>BP</td>
<td>0.048</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0015074</td>
<td>DNA integration</td>
<td>BP</td>
<td>0.039</td>
<td>Uh</td>
</tr>
<tr>
<td>GO: 0006259</td>
<td>DNA metabolic process</td>
<td>BP</td>
<td>0.023</td>
<td>Uh</td>
</tr>
<tr>
<td>GO: 0010181</td>
<td>FMN binding</td>
<td>MF</td>
<td>0.047 and 0.022</td>
<td>Sr and Srz</td>
</tr>
<tr>
<td>GO: 0030515</td>
<td>snoRNA binding</td>
<td>MF</td>
<td>0.021</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0016491</td>
<td>Oxidoreductase activity</td>
<td>MF</td>
<td>0.028</td>
<td>Srr</td>
</tr>
<tr>
<td>GO: 0016853</td>
<td>Isomerase activity</td>
<td>MF</td>
<td>0.029</td>
<td>Uh</td>
</tr>
</tbody>
</table>

CC, cellular component; BP, biological process; MF, molecular function.

*Calculated by Fisher's classic test with Parent–Child correction. Only entries with P value ≤ 0.05 are shown.

Species in which GO terms were found enriched. Srr, *S. reilianum* f. sp. *reilianum*; Srz, *S. reilianum* f. sp. *zeae*; Uh, *U. hordei*.

Virulence Contribution of Effector Genes Showing Signs of Positive Selection in *S. reilianum*

Candidate effector genes inferred to be under positive selection in a particular species could play a critical role in pathogenicity. Therefore, we sought to assess the contribution to virulence of such candidate genes by creating individual deletion mutants. In total, we tested nine candidate genes with high $d_{NS}/d_{S}$ ratios and predicted to encode secreted proteins: three with signatures of positive selection only in *S. reilianum* f. sp. *zeae*, three with signatures of positive selection in *S. reilianum* f. sp. *zeae* as well as in *S. reilianum* f. sp. *reilianum* and three with signatures of positive selection only in *S. reilianum* f. sp. *reilianum*. All nine chosen candidate genes together with their characteristics are summarized in table 2. Deletion mutants were generated in the haploid...
showed a strong reduction in virulence (table 2 and fig. 2).

S. reilianum selection was only detected in also generated in strain JS161 in cases where positive se-

p osit ives e le cti o n i nbo t h f o r m a e s p e ci a s o f

symptoms were scored in male and female flowers

Gaspe Flint, a dwarf variety of corn, was infected and

colonized maize plants and cause disease without a compatible mating partner (Schirawski et al. 2010) but virulence is much reduced relative to in-

disease without a compatible mating partner (Schirawski et al. 2010) but virulence is much reduced relative to in-

The gene sr10529 in S. reilianum f. sp. zeae is orthologous to the previously identified and characterized gene pit2 (UMAG_01375) in U. maydis.

Pit2 plays an essential role in virulence as inhibitor of a group of maize papain-like cysteine proteases that are secreted to the apoplast (Doehlemann et al. 2011; Mueller et al. 2013). Previous work identified a conserved domain of 14 amino acids (PID14) in Pit2 as required and sufficient for the inhibition of maize cysteine proteases (Mueller et al. 2013). When the branch-site model of PAML 4 (Yang 2007) was used to identify amino acid residues under positive selection in the Pit2 orthologues of the two S. reilianum species, only two residues residing in the PID14 domain were found under positive selection. However, 24 positively selected residues were detected outside this domain in the 57 amino acid long C-terminal part (fig. 3).

Table 2

<table>
<thead>
<tr>
<th>Family</th>
<th>Gene ID</th>
<th>Gene Description</th>
<th>$d_0/d_s$</th>
<th>Species with Detected Positive Selection</th>
<th>Number of Paralogs</th>
<th>Virulence Phenotype of Candidate Gene Deletion</th>
<th>Closest Ortholog in Ustilago maydis</th>
</tr>
</thead>
<tbody>
<tr>
<td>FAM001428</td>
<td>sr10529</td>
<td>Conserved hypothetical protein</td>
<td>31.1469</td>
<td>Srz and Srr</td>
<td>0</td>
<td>Virulence abolished</td>
<td>UMAG_01375 (pit2)</td>
</tr>
<tr>
<td>FAM005472</td>
<td>sr10059</td>
<td>Conserved hypothetical Ustilaginaceae-specific protein</td>
<td>6.53881</td>
<td>Srz and Srr</td>
<td>0</td>
<td>Virulence unaffected</td>
<td>UMAG_05306 (cluster 19A)</td>
</tr>
<tr>
<td>FAM000532</td>
<td>sr10182</td>
<td>Conserved hypothetical protein</td>
<td>1.57473</td>
<td>Srr</td>
<td>10</td>
<td>Virulence unaffected</td>
<td>UMAG_00492</td>
</tr>
<tr>
<td>FAM002067</td>
<td>sr12968</td>
<td>Conserved hypothetical protein</td>
<td>37.9007</td>
<td>Srr</td>
<td>0</td>
<td>Virulence unaffected</td>
<td>UMAG_02006</td>
</tr>
<tr>
<td>FAM003728</td>
<td>sr14558</td>
<td>Conserved hypothetical protein</td>
<td>24.355</td>
<td>Srz</td>
<td>0</td>
<td>Virulence unaffected</td>
<td>UMAG_03564</td>
</tr>
<tr>
<td>FAM004113</td>
<td>sr14944</td>
<td>Conserved hypothetical Ustilaginaceae-specific protein</td>
<td>4.30527</td>
<td>Srz and Srr</td>
<td>0</td>
<td>Virulence unaffected</td>
<td>UMAG_04034 (cluster 11–16)</td>
</tr>
<tr>
<td>FAM003465</td>
<td>sr14347</td>
<td>Conserved hypothetical protein</td>
<td>544.37</td>
<td>Srz</td>
<td>5</td>
<td>Virulence unaffected</td>
<td>UMAG_03349</td>
</tr>
<tr>
<td>FAM001868</td>
<td>sr12897</td>
<td>Conserved hypothetical protein</td>
<td>999</td>
<td>Srr</td>
<td>0</td>
<td>Virulence unaffected</td>
<td>UMAG_01820</td>
</tr>
<tr>
<td>FAM000842</td>
<td>sr12084</td>
<td>Conserved hypothetical protein</td>
<td>999</td>
<td>Srz</td>
<td>2</td>
<td>Virulence unaffected</td>
<td>UMAG_00792 (cluster 1–32)</td>
</tr>
</tbody>
</table>

*Species are S. reilianum f. sp. zeae (Srz) and S. reilianum f. sp. reilianum (Srr).

*Based on BLASTp search with an e-value cutoff of 0.001.

*Maximum value allowed by the PartNH program.

*Infinity due to low value of $d_s$.

The ten paralogs include: sr13431, sr11876, sr16607, sr11405, sr10621, sr16723, sr16877, 13293, sr11163.2 and sr15970.

The five paralogs include: sr12257, sr11661, sr13976, sr14607, and sr11273.

The two paralogs include: sr11085 and sr12086.

As described in Doehlemann et al. (2011) and Mueller et al. (2013).

As described in Köpfer et al. (2006).

As described in Schirawski et al. (2010).
**FIG. 2.**—Virulence phenotypes of single deletion mutants of positively selected genes in *Sporisorium reilianum* f. sp. zeae. Plants of the maize variety “Gaspe Flint” were infected with the solopathogenic strain JS161 or independent deletion mutants of candidate genes as indicated below each bar. Deletion
We used evolutionary comparative genomics of five related smut fungi infecting four different host plants to identify genes with a signature of positive selection during species divergence, with a focus on genes encoding predicted secreted proteins, as such genes were suggested to contribute to virulence in various plant pathogenic microbes (Aguileta et al. 2010; Stukenbrock et al. 2011; Hacquard et al. 2012; Dong et al. 2014; Huang et al. 2014; Sharma et al. 2014). Our analysis revealed that positive selection is found between paralogous genes in *U. hordei*, where they belong to families with species-specific expansions. In contrast, genes under positive selection in the other four species belong to families of orthologous sequences. Although we find evidence for a large set of genes under positive selection detected in the respective species and purple residues indicate sites found under positive selection in both species. The yellow shaded area is orthologous to the previously identified conserved PID14 domain, which is required and sufficient for inhibition of a group of papain-like cysteine proteases. Green sequences indicate secretion signal peptides and bold numbers above the alignment indicate positions in *UHOR_02064*.

**Discussion**

We used evolutionary comparative genomics of five related smut fungi infecting four different host plants to identify genes with a signature of positive selection during species divergence, with a focus on genes encoding predicted secreted proteins, as such genes were suggested to contribute to virulence in various plant pathogenic microbes (Aguileta et al. 2010; Stukenbrock et al. 2011; Hacquard et al. 2012; Dong et al. 2014; Huang et al. 2014; Sharma et al. 2014). Our analysis revealed that positive selection is found between paralogous genes in *U. hordei*, where they belong to families with species-specific expansions. In contrast, genes under positive selection in the other four species belong to families of orthologous sequences. Although we find evidence for a large set of genes under positive selection detected in the *S. reilianum* species, signatures for positive selection are hardly detectable in the more distant relatives *U. hordei*, *U. maydis*, and *S. scitamineum* that diverged earlier. Finding evidence for positive selection over time spans of several millions of years is notoriously difficult (Gillespie 1994) because of two main reasons: 1) periods where genes are evolving under positive selection may occur episodically and may be followed by long episodes of purifying selection, leading to an average $d_N/d_S < 1$ on long periods of time and 2) fast evolving genes may diverge to an extent where their homology is difficult to infer and where they can no longer be aligned reliably. To overcome this problem, more genome information of species with intermediate branching points is needed (Gillespie 1994).

Predicted secreted proteins were about three times over-represented in the set of positively selected genes, which illustrates the importance of secreted proteins in adaptation processes of smut fungi. This also corroborates results in other plant pathogenic microbes like *Melampsora sp.*, *Z. tritici* and the wheat powdery mildew *Blumeria graminis* (Joly et al. 2010; Stukenbrock et al. 2011; Wicker et al. 2013). However, the majority of positively selected genes encodes...
cytoplasmic proteins (fig. 1C), suggesting that both secreted and nonsecreted proteins are important targets of adaptation. A Gene Ontology analysis in *S. reilianum* showed that mainly processes related to metabolism and its regulation as well as responses to starvation and external stimuli are enriched in cytoplasmic proteins under positive selection. This points at a role of these proteins in adaptation to differences in nutrient availability in the respective host plants maize and sorghum as well as responses to cues originating from the respective host (Haueisen and Stukenbrock 2016). A study conducted in *U. maydis* has shown that the fungus induces major metabolic changes in the host plant upon infection during establishment of biotrophy and undergoes a series of developmental transitions during host colonization that are likely influenced by the host environment (Doehlemann et al. 2008). It is thus conceivable that the two *S. reilianum* accessions have adapted to their different hosts that differ significantly for example in their amino acid and vitamin composition (Etuk et al. 2012). Furthermore, recent studies in *U. maydis* suggested that intracellular changes of metabolism influence virulence, and therefore the underlying proteins could be targets of positive selection (Kretschmer et al. 2012; Goulet and Saville 2017).

Out of nine deletions of positively selected genes, only one mutant, lacking sr10529, was affected in virulence. Although six of the deleted genes are single genes in *S. reilianum* f. sp. zeae for which we failed to identify paralogs, sr12084 has two paralogs, sr14347 has five paralogs, and sr10182 has ten paralogs. We restricted our analyses to generating deletion mutants in some of the genes under positive selection. This leaves open the possibility that the paralogous genes have redundant functions in virulence. Adapting the CRISPR-Cas9 technology allowing multiplexing (Schuster et al. 2017) to *S. reilianum* will be instrumental in testing this hypothesis in future studies. Alternatively, the candidate effectors we investigated may be needed under conditions which differ from those tested here. For example, *S. reilianum* f. sp. zeae can also systemically colonize maize plants via root infection (Mazaheri-Naeini et al. 2015), a colonization route we have not assessed in our experimental setup. Moreover, we employed only one maize cultivar for infection assays. Results from other pathosystems suggest that virulence effects can strongly depend on the host and pathogen genotypes, in particular in the presence of *R* and *avr* genes (Petit-Houdenet and Fudal 2017). No *avr-R* gene interaction was described so far in the *S. reilianum* f. sp. zeae-maize pathosystem. Instead, quantitative virulence differences are observed when different host cultivars are infected (Lübberstedt et al. 1999). Knowing the expression profile of effector genes may assist the identification of differences in development of the mutants compared with wild-type strains. Since we lack this information, we scored disease symptoms only in the inflorescences ~9 weeks after infection. Additionally, it may be possible that small differences in virulence between JS161 and deletion mutants of candidate genes remain undetected due to the weak infection behavior of JS161. For example, spore formation is only rarely observed after infecting maize plants with the solopathogenic strain. In contrast, infections resulting from infections with two compatible haploid strains show spores in ~40% of the infected plants (Zuther et al. 2012). This means that defects related to spore formation will not be evident in mutants of JS161. In three cases, positive selection was detected in orthologous genes in *S. reilianum* f. sp. *reilianum* while candidate effector genes were for experimental reasons deleted in *S. reilianum* f. sp. zeae. Therefore, it cannot be excluded that these effectors might have a virulence function in *S. reilianum* f. sp. *reilianum*. In this case, the positively selected effector genes might have evolved during adaptation to the sorghum host and present host specificity genes. In summary, our virulence assays leave open the possibility that the eight candidate genes which did not show a contribution to virulence could play a role in pathogenicity under conditions not tested here. Alternatively, candidate effector proteins might also be positively selected for traits that are not directly linked to pathogenicity. Such traits could for instance involve competition with large numbers of other plant colonizing microbes (Zhan and McDonald 2013; Rovenich et al. 2014). Secreted proteins of *S. reilianum* could act for example as toxin or could efficiently utilize resources from the environment and thereby limit the growth of other microbes. In these cases, a contribution to virulence is not expected to be observed in the employed infection assay with the effector gene mutants. Moreover, our molecular dating analysis showed that the common ancestors of the investigated smut species originated before the beginning of crop domestication. Therefore, positive selection, whose signs we detect by our approach, has most likely occurred on ancestral host plants and not on the domesticated host maize. Consequently, some of the candidate effector genes under positive selection might not be important for the colonization of crop plants, but for infection of related wild species.

In *U. maydis*, we note that effector genes residing in clusters whose deletion affected virulence (Kämper et al. 2006; Schirawski et al. 2010) have similar *dN/dS* ratios as effector genes in clusters where the deletion had no effect on virulence (median *dN/dS* ratio 0.0619 vs. 0.1094; Wilcoxon rank test with *P* value = 0.1848). Furthermore, orthologues of the effectors Pep1, Stp1, and Cmu1, which were shown to have important roles in pathogenicity of *U. maydis* (Doehlemann et al. 2009; Schipper 2009; Djamei et al. 2011) showed no signature of positive selection. These observations could suggest that certain fungal effector proteins are under evolutionary constraint and are therefore not free to accumulate nonsynonymous mutations. Such effectors are conserved over long time spans (Schirawski et al. 2010; Hemetsberger et al. 2015; Sharma et al. 2015) and this illustrates that they are instrumental for successful infections in a large group of smut fungi. They probably target molecules shared by
several host plants, for example, housekeeping functions that cannot easily evolve in response to the binding of an effector.

One candidate gene (sr10529) under positive selection in both formae speciales of *S. reilianum* showed a strong contribution to virulence upon deletion. It is orthologous to the previously described protease inhibitor Pit2 in *U. maydis*, where the deletion also abolished virulence (Doehlemann et al. 2011; Mueller et al. 2013). Positively selected residues in the PID14 domain of Pit2 might reflect that different proteases need to be inhibited in maize and sorghum. Pit2 might thus contribute to determining the host range of the respective species. A role of cysteine protease inhibitors in host specificity was demonstrated in *Phytophthora infestans*, a pathogen of potato and its sister species *P. mirabilis*, which infects the ornamental plant *Mirabilis jalapa*. Positively selected orthologous protease inhibitors were shown to inhibit proteases specific to the respective host plants and this specificity could be traced back to a single amino acid substitution (Dong et al. 2014). Surprisingly, 24 positively selected sites in Pit2 were detected outside the PID14 domain in the 57 amino acid long C-terminal part in both *S. reilianum* f. sp. zeae and *S. reilianum* f. sp. *reilianum*. This finding raises the intriguing possibility that the C-terminus of Pit2 might possess a second function that is independent of protease inhibition. Earlier work has shown that the *pit1* gene encoding a transmembrane protein is located next to the *pit2* effector gene and both genes contribute similarly to virulence (Doehlemann et al. 2011). Furthermore, *pit1* and *pit2* are divergently transcribed, which makes it likely that the expression of *pit1* and *pit2* is coregulated. In addition, this gene arrangement of *pit1* and *pit2* is conserved in *U. hordei*, *U. maydis*, *S. scitamineum*, and *S. reilianum* (Sharma et al. 2015). This finding has led to the speculation that Pit1 and Pit2 somehow act together to govern virulence of *U. maydis* and related smut fungi. It was hypothesized that Pit2 shuttles apoplastic maize proteins toward Pit1, thereby scavenging damage-associated molecules (Doehlemann et al. 2011). In this scenario, the positively selected amino acids in the C-terminus of Pit2 could have been selected for scavenging such molecules as adaptation to the two hosts. In future studies, it will be highly interesting to complement the *pit2* mutant of *S. reilianum* f. sp. *zeae* with the *pit2* orthologue of *S. reilianum* f. sp. *reilianum* to see if this promotes virulence on sorghum.

**Conclusions**

Screens for genes with signs of positive selection are commonly used to identify candidate effector genes in various plant pathogenic microbes. However, it is currently largely open whether positively selected effector genes play indeed a role in virulence. Here, we used comparative genomics of five smut fungi and showed that only one out of nine genes under positive selection contributes to virulence of *S. reilianum*. Moreover, the majority of positively selected genes did not encode predicted secreted proteins. Our results leave open the possibility that many genes with signatures of positive selection contribute to virulence under conditions not tested in this study or are selected in traits that are not directly related to pathogenicity.

**Supplementary Material**

Supplementary data are available at Genome Biology and Evolution online.

**Acknowledgments**

Our work was supported through the LOEWE program of the state of Hesse through SYNMICRO and through the Max Planck Society. G.S. was funded by the International Max Planck Research School (IMPRS) for Molecular, Cellular and Environmental Microbiology. We thank Michael Bölker for providing the yeast strain BY4741 and the plasmid pRS426 which were used for cloning purposes, and Benoit Nabholz and Nicolas Lartillot for guidance with the dating analysis. We are grateful to Markus Rampp (Max Planck Computing Center Garching) for assistance with computational work and to Stefan Schmidt for expert greenhouse service, as well as all present and former group members for lively discussions and Eva H. Stukenbrock for her comments on the manuscript.

**Literature Cited**


Associate editor: Rebecca Zufall