The insertion of a mitochondrial selfish element into the nuclear genome and its consequences

Julien Y. Dutheil[1,2,3,*], Karin Münch[2], Klaas Schotanus[2,4,5], Eva H. Stukenbrock[1,2,4], Regine Kahmann[2]

1. Max Planck Institute for Evolutionary Biology
August-Thienemann-Str. 2
24306 Plön, Germany

2. Max Planck Institute for Terrestrial Microbiology
Karl-von-Frisch-Str. 10
35043 Marburg, Germany

3. Institute of Evolutionary Sciences
CNRS - University of Montpellier – IRD - EPHE
Place Eugène Bataillon, 34095 Montpellier, France.

4. Christian Albrechts University of Kiel
24118 Kiel, Germany

5. Current address: Department of Molecular Genetics and Microbiology (MGM).
Duke University Medical Center. 213 Research Drive, 317 CARL Building, Box 3546 Durham, NC 27710 USA

Correspondence:
* Julien Y. Dutheil, dutheil@evolbio.mpg.de

Running head: Insertion of a mitochondrial homing endonuclease into the nuclear genome

Keywords: homing endonuclease, mitochondrion, intron, gene birth, gene transfer

Abstract

Homing endonucleases (HE) are enzymes capable of cutting DNA at highly specific target sequences, the repair of the generated double-strand break resulting in the insertion of the HE-encoding gene ("homing" mechanism). HEs are present in all three domains of life and viruses; in eukaryotes, they are mostly found in the genomes of mitochondria and chloroplasts, as well as nuclear ribosomal RNAs. We here report the case of a HE that accidentally integrated into a telomeric region of the nuclear genome of the fungal maize pathogen *Ustilago maydis*. We show that the gene has a mitochondrial origin, but its original copy is absent from the *U. maydis* mitochondrial genome, suggesting a subsequent loss or a horizontal transfer from a different species. The telomeric HE underwent mutations in its active site and lost its original start codon. A potential other start codon was retained downstream, but we did not detect any significant transcription of the newly created open reading frame, suggesting that the inserted gene is not functional. Besides, the insertion site is located in a putative RecQ helicase gene, truncating the C-terminal domain of the protein. The truncated helicase is expressed during infection of the host, together with other homologous telomeric helicases. This unusual mutational event altered two genes: the integrated HE gene subsequently lost its homing activity, while its insertion created a truncated version of an existing gene, possibly altering its function. As the insertion is absent in other field isolates, suggesting that it is recent, the *U. maydis* 521 reference strain offers a snapshot of this singular mutational event.

36 ## *Introduction*

37 The elucidation of the mechanisms at the origin of genetic variation is a longstanding goal
38 of molecular evolutionary biology. Mutation accumulation experiments - together with comparative
39 analysis of sequence data - are instrumental in studying the processes shaping genetic diversity at
40 the molecular level (Kondrashov and Kondrashov 2010; Eyre-Walker and Keightley 2007). They
41 revealed that the spectrum of mutations ranges from single nucleotide substitutions to large scale
42 chromosomal rearrangements, and encompasses insertions, deletions, inversions, and duplication of
43 genetic material of variable length (Lynch et al. 2008). Mutation events may result from intrinsic
44 factors such as replication errors and repair of DNA damage. In some cases, however, mutations can
45 be caused or favored by extrinsic factors, such as mutagenic environmental conditions or parasitic
46 genome entities like viruses or selfish mobile elements. Such particular sequences, able to replicate
47 and invade the host genome, may have various effects including inserting long stretches of DNA
48 that do not encode any organismic function, but also disrupting, copying and moving parts of the
49 genome sequence. These selfish element-mediated mutations can significantly contribute to the
50 evolution of their host: first, the invasion of mobile elements creates "junk" DNA that can
51 significantly increase the genome size (Lynch 2007), and some of this material can be ultimately
52 domesticated and acquire a new function, beneficial to the host (Kaessmann 2010; Volff 2006).
53 Second, the genome dynamics resulting from the activity of mobile elements can generate novelty
54 by gene duplication (Ohta 2000; Dutheil et al. 2016) or serve as a mechanism of parasexuality and
55 compensate for the reduced diversity in the absence of sexual reproduction (Dong et al. 2015;
56 Möller and Stukenbrock 2017). Finally, control mechanisms (such as repeat-induced point
57 mutations in fungi (Gladyshev 2017)) may also incidentally affect genetic diversity (Grandaubert et
58 al. 2014).

59 Intron-borne homing endonuclease genes (HEG) constitute a class of selfish elements

60 whose impact on genome evolution is less well documented. They encode a protein able to
61 recognize a particular genomic DNA sequence and cut it (homing endonuclease, HE). The resulting
62 double-strand break is subsequently repaired by recombination using the homologous sequence
63 containing the HEG itself as a template, resulting in its insertion in the target location (Stoddard
64 2005). As the recognized sequence is typically large, its occurrence is rare and the insertion
65 typically happens at a homologous position. In this process, a *heg*$^{+}$ element containing the
66 endonuclease gene converts a *heg*$^{-}$ allele (devoid of HEG but harbouring the recognition sequence)
67 to *heg*$^{+}$, a mobility mechanism referred to as *homing* (Dujon et al. 1989). After the insertion, the
68 host cell is homozygous *heg*$^{+}$, and the HEG segregates at a higher frequency than the Mendelian
69 rate (Goddard and Burt 1999). The open reading frame of the HEG is typically associated with a
70 sequence capable of self-splicing, either at the RNA (group-I introns) or protein (inteins) level,
71 avoiding disruption of functionality when inserted in a protein-coding gene (Chevalier and Stoddard
72 2001; Stoddard 2005). The dynamic of HEGs has been well described, and involves three stages: (i)
73 conversion from *heg*$^{-}$ to *heg*$^{+}$ by homing activity, (ii) degeneration of the HEG leading to the loss of
74 homing activity, but still protecting against a new insertion because the target is altered by the
75 insertion event and (iii) loss of the HEG leading to the restoration of the *heg*$^{-}$ allele (Gogarten and
76 Hilario 2006; Barzel et al. 2011). This cycle leads to recurrent gains and losses of HEG at a given
77 genomic position, and ultimately to the loss of the HEG at the population level unless new genes
78 invade from other locations or by horizontal gene transfer (Gogarten and Hilario 2006).

79 HEGs are found in all domains of life as well as in the genomes of organelles,
80 mitochondria and chloroplasts (Stoddard 2005; Lambowitz and Belfort 1993; Belfort and Roberts
81 1997). In several fungi, HEGs are residents of mitochondria. Here, we study the molecular
82 evolution of a HEG from the fungus *Ustilago maydis*, which serves as a model for the elucidation
83 of (i) fundamental biological processes like cell polarity, morphogenesis, organellar targeting, and
84 (ii) the mechanisms allowing biotrophic fungi to colonize plants and cause disease (Steinberg and

85 Perez-Martin 2008; Djamei and Kahmann 2012; Ast et al. 2013). *U. maydis* is the most well-studied
86 representative of smut fungi, a large group of plant pathogens, because of the ease by which it can
87 be manipulated both genetically and through reverse genetics approaches (Vollmeister et al. 2012).
88 Besides, its compact, fully annotated genome comprises only 20.5 Mb and is mostly devoid of
89 repetitive DNA (Kämper et al. 2006). The genome sequences of several related species,
90 *Sporisorium reilianum*, *S. scitamineum,* and *U. hordei* causing head smut in corn, smut whip in
91 sugarcane and covered smut in barley, respectively, provide a powerful resource for comparative
92 studies (Schirawski et al. 2010; Laurie et al. 2012; Dutheil et al. 2016). Here, we report the case of a
93 mitochondrial HEG that integrated into the nuclear genome of *U. maydis*. This singular mutation
94 event created two new genes: first the original endonuclease activity of the integrated HEG was
95 inactivated by a deletion in the active site, leading to a frameshift and a new open reading frame
96 containing the DNA-binding domain of the HEG (Derbyshire et al. 1997). Second, the integration
97 of the HEG occurred within another protein-coding gene, leading to its truncation.

98 *Results*

99 We report the analysis of the nuclear gene *UMAG_11064* from the smut fungus *U.*
100 *maydis*, which was identified as an outlier in a whole-genome analysis of codon usage. We first
101 provide evidence that the gene is a former HEG and then reconstruct the molecular events that led
102 to its insertion in the nuclear genome using comparative sequence analysis. Finally, we assess the
103 phenotypic impact of the insertion event.

104 The *UMAG_11064* nuclear gene has a mitochondrial codon usage.

105 We studied the synonymous codon usage in protein-coding genes of the smut fungus *U.*
106 *maydis*, using within-group correspondence analysis. As opposed to other methods, within-group
107 correspondence analysis allows the comparison of codon usage while adequately taking into
108 account confounding factors such as variation in amino-acid usage (Perrière and Thioulouse 2002).

109 We report a distinct synonymous codon usage for nuclear genes and mitochondrial genes (Figure
110 1A), with the notable exception of the nuclear gene *UMAG_11064,* which displays a typical
111 mitochondrial codon usage. The *UMAG_11064* gene is located in the telomeric region of
112 chromosome 9, with no further downstream annotated gene (Figure 1B). It displays a low GC
113 content of 30%, which contrasts with the GC content of the flanking regions (50%) and the rather
114 homogeneous composition of the genome sequence of *U. maydis* as a whole. It is, however, in the
115 compositional range of the mitochondrial genome (Figure 1B). Altogether, the synonymous codon
116 usage and GC content of *UMAG_11064* suggest a mitochondrial origin.

117 In order to confirm the chromosomal location of *UMAG_11064*, we amplified and
118 sequenced three regions encompassing the gene using primers within the *UMAG_11064* gene and
119 primers in adjacent chromosomal genes upstream and downstream of *UMAG_11064* (Figure S1).
120 The sequences of the amplified segments were in full agreement with the genome sequence of *U.*
121 *maydis* (Kämper et al. 2006), thereby ruling out possible assembly artefacts in this region. As both
122 the GC content and synonymous codon usage of *UMAG_11064* are indistinguishable from the ones
123 of mitochondrial genes and have not moved toward the nuclear equilibrium, the transfer of the gene
124 to its nuclear position is likely to have occurred relatively recently.

125 The *UMAG_11064* gene contains parts of a former GIY-YIG homing 126 endonuclease

127 To gain insight into the nature of the *UMAG_11064* gene, its predicted nucleotide
128 sequence was searched against the NCBI non-redundant nucleotide sequence database. Surprisingly,
129 the sequence of *UMAG_11064* has no match in the mitochondrial genome of *U. maydis* itself
130 (GenBank entry NC_008368.1), but high similarity matches were found in the mitochondrial
131 genome of three other smut fungi (Supplementary Table S1): *S. reilianum* (87% nucleotide
132 identity), *S. scitamineum* (79%), and *U. bromivora* (76%). Two other very similar sequences were

133 found in the mitochondrial genome of two other smut fungi, *Tilletia indica* and *Tilletia walkeri*
134 (69% nucleotide identity), as well as in mitochondrial genomes from other basidiomycetes (e.g.
135 *Laccaria bicolor*, 72%) and ascomycetes (e.g. *Leptosphaeria maculans*, 69%, see Supplementary
136 Table S1). The protein sequence of *UMAG_11064* shows high similarity with fungal HEGs, in
137 particular of the so-called GIY-YIG family (Supplementary Table S2) (Stoddard 2005). The closest
138 fully annotated protein sequence matching *UMAG_11064* corresponds to the GIY-YIG HEG located
139 in intron 1 of the *cox1* gene of *Agaricus bisporus* (I-AbiIII-P, 54% nucleotide identity). The amino-
140 acid sequence of UMAG_11064 matches the N-terminal part of this protein containing the DNA-
141 binding domain of the HE (Derbyshire et al. 1997).

142 We performed a codon alignment of the *UMAG_11064* gene together with the most
143 similar sequences identified by Blast, using the Macse codon aligner to infer sequence alignment in
144 the presence of frameshifts (Ranwez et al. 2011). The sequences from *S. scitamineum* and *U.*
145 *bromivora* appeared to have several frameshifts introducing stop codons, suggesting that these
146 sequences are pseudogenes. We reconstructed the phylogeny of the nucleotide sequences after
147 removing ambiguously aligned regions (Figure 2). The resulting tree shows that the closest relative
148 of the *UMAG_11064* gene is the intronic sequence from *S. reilianum*.

149 As the GC profile of *UMAG_11064* suggests that the upstream region also has a
150 mitochondrial origin (Figure 1B), we performed a codon alignment of the 5' region with the full
151 intron sequences of S. *reilianum*, *T. indica* and *T. walkeri* as well as the sequence of I-AbIII-P from
152 *A. bisporus* in order to search for putative traces of the activity domain of the HE (Figure 3). We
153 found that the intergenic region between *UMAG_11065* and *UMAG_11064* is similar to the activity
154 domain of other GIY-YIG HE, and contains remnants of the former active site of the type GVY-
155 YIG (Figure 3). Compared to I-AbiIII-P and homologous sequences in *Tilletia*, however, a
156 frameshift mutation has occurred in the active site (a 7 bp deletion). The predicted gene model for
157 *UMAG_11064* starts at a conserved methionine position, 14 amino-acids downstream of the former

158 active site (Figure 3) and contains the *helix-turn-helix* DNA-binding domain of the original HE.

159 A branch model of codon sequence evolution was fitted to the codon sequence alignment
160 of *UMAG_11064* and its identified homologs, with the exception of the putative pseudogenes from
161 *S. scitamineum* and *U. bromivora*. The proteins appear to be evolving under purifying selection
162 (dN/dS < 1) on most branches of the tree, with the exception of the branches leading to the two
163 *Tilletia* sequences, as well as the branch leading to *UMAG_11064* (Figure 2). We note, however,
164 that the codeml program suffers from convergence issues on these particular branches, as witnessed
165 by the variance in final estimates after 10 independent runs (Figure 2). Such convergence failures
166 likely result from the branch model being an over-parametrized model, here fitted on relatively
167 short sequences. To further assess whether the high dN/dS ratio measured in the *UMAG_11064* gene
168 could be explained both by relaxed purifying selection or positive selection, we fitted a branch-site
169 model allowing specifically for sites in the *UMAG_11064* gene to evolve under positive selection
170 (foreground branch), which we contrasted with a null model where all sites evolve under purifying
171 selection or neutral evolution. The likelihood ratio test was not significant (p-value equal to
172 0.1585), even after removing the *Tilletia* sequences (p-values equal to 0.2183), and does not reject
173 the hypothesis that the *UMAG_11064* gene is evolving under a nearly neutral scenario. The higher
174 dN/dS in the branch leading to *UMAG_11064* might, therefore, be the result of relaxed purifying
175 selection.

176 Altogether, these results suggest that *UMAG_11064* is a former HE that inserted into the
177 nuclear genome from the mitochondrion, was then inactivated by a deletion in its active site and
178 acquired a new start codon, allowing it to code for a protein sequence with the former nucleotide
179 binding domain of the HE.

The *UMAG_11064* gene is similar to an intronic mitochondrial sequence of *S. reilianum*

The closest homologous sequence of *UMAG_11064* was found in the first intron of the *cox1* gene of the smut fungus *S. reilianum* while this sequence was absent in the mitochondrial genome of *U. maydis*. The *cox1* genes of *S. reilianum* and *U. maydis* both have eight introns, of which only seven are homologous in position and sequence (Figure 4). *S. reilianum* has one extra intron in position 1, while *U. maydis* has one extra intron in position 6. In *U. maydis* all introns but the sixth one are reported to contain a HEG. A blast search of this intron's sequence, however, revealed similarity with a homing endonuclease of type LAGLIDADG (Supplementary Table S4). In *S. reilianum*, intron 1 (the putative precursor of *UMAG_11064)* and intron 2 are not annotated as containing a HEG. Blast searches of the corresponding sequences, however, provided evidence for homology with a GIY-YIG HE (Supplementary Table S5) and a LAGLIDADG HE, respectively (Supplementary Table S6). Like many other species of fungi (Stone et al. 2018; Jalalzadeh et al. 2015; Pogoda et al. 2019), plants (Cho et al. 1998) and even animals (Fukami et al. 2007; Schuster et al. 2017), the *cox1* gene seems to be a hotspot of HEG-encoding introns in smut fungi.

Lastly, no homolog of intron 1 in *S. reilianum* was detected in the mitochondrial genome of *U. maydis*. A closer inspection showed that the ORF could be aligned with related HEs in other fungi (Figure 3). This alignment revealed an insertion of four amino-acids, a deletion of the first glycine residue in the active site plus several frameshifts at the beginning of the gene, which suggests that this gene has been altered and might not encode a functional HE any longer.

UMAG_11064 inserted into a gene encoding a RecQ helicase

In order to study the effect of the HEG insertion in the nuclear genome, we looked at the genomic environment of the *UMAG_11064* gene. Downstream of *UMAG_11064* are telomeric repeats, while the next upstream gene, *UMAG_11065*, is uncharacterized. A similarity search for

204 *UMAG_11065* detected 13 paralogous sequences in the *U. maydis* genome (including one,
205 *UMAG_12076*, on an unmapped contig), but only low-similarity matches in other sequenced smut
206 fungi (see Methods). The closest non-smut related sequence comes from a gene from *Fusarium*
207 *oxysporum*. We inferred the evolutionary relationships between the 14 genes by reconstructing a
208 maximum likelihood phylogenetic tree, and found that the *UMAG_11065* gene is closely related to
209 *UMAG_04486*, located on chromosome 14 (Figure 5 and Table 1). The *UMAG_04486* gene,
210 however, is predicted to be almost six times as long as *UMAG_11065*. We note that the downstream
211 region of *UMAG_11064* does not show any similarity with the 3' part of the *UMAG_04486* gene,
212 suggesting that the insertion of the HEG did not lead to the formation of an intron in the
213 *UMAG_11065* gene, but rather to its truncation. A search for similar sequences of *UMAG_11065*
214 and its relatives in public databases revealed homology with so-called RecQ helicases
215 (Supplementary Table S3), enzymes known to be involved in DNA repair and telomere expansion
216 (Singh et al. 2012). While this function is only predicted by homology, we note that all 12
217 chromosomal *recQ* related genes are located very close to telomeres in *U. maydis* (Table 1),
218 suggesting a role of these genes in telomere maintenance (Sánchez-Alonso and Guzmán 1998).
219 Lastly, we tested whether the truncation of *UMAG_11065* was followed by positively selected
220 mutations in the remaining part of the gene. We inferred a dN/dS ratio equal to 0.342, which
221 suggests that the *UMAG_11065* gene evolved mostly under purifying selection since divergence
222 from the *UMAG_04486* gene. The insertion of UMAG_11064, therefore, was not followed by
223 positive selection, or was too recent for sufficient positively selected substitutions to occur.

224 *U. maydis* populations shows structural polymorphism in the telomeric 225 region of chromosome 9

226 Because the *UMAG_11064* gene still displays a strong signature of its mitochondrial
227 origin (codon usage and GC content), its transfer may have occurred recently. In order to provide a

228 timeframe for the insertion event, we examined the structure of the genomic region of the insertion
229 in other *U. maydis* and *S. reilianum* isolates, as well as the structure of the *cox1* exons 1, 2 and 7.
230 The regions that could be amplified and their corresponding sizes are listed in Supplementary
231 Figure S2 and the inferred genome organisations are summarized in Figure 6. The *UMAG_11064*
232 gene is present in the FB1-derived strain SG200, as well as in the Holliday strains 518 and 521, but
233 is absent in the nuclear and mitochondrial genome sequences of a recent *U. maydis* isolate from the
234 US, strain 10-1, as well as from 5 Mexican isolates (I2, O2, P2, S5 and T6, Figure S2A).
235 Conversely, the *UMAG_11072* gene, which is located further away from the telomere on the same
236 chromosome arm, could be amplified in all strains. This positive control demonstrates that the lack
237 of amplification of *UMAG_11064* in some strains is not due to any issue with the quality of the
238 extracted DNA (Figure S2B). These results suggest that, either the *UMAG_11064* gene was
239 ancestral to all tested strains and subsequently lost in the Mexican and 10-1 strains, or it inserted in
240 an ancestor of the three strains 518, 521 and SG200, after the divergence from other *U. maydis*
241 strains, an event that occurred after the domestication of maize and the spread of the associated
242 pathogen, 10,000 to 6,000 years ago (Munkacsi et al. 2008). Moreover, all *U. maydis* strains possess
243 intron 6 in the mitochondrial *cox1* gene, which is absent in *S. reilianum*. While the three *S.*
244 *reilianum* strains tested carry intron 1, the most direct descendant of the progenitor of the HEs, it
245 was absent in all *U. maydis* strains tested (Figure S2C-E and Figure 6).

246 Functional characterization

247 To shed light on the functional implication of the translocation of the HEG and
248 subsequent mutations we (i) assessed the expression profile of these genes and (ii) generated a
249 deletion strain and phenotyped it. For the expression analysis we relied on a previously published
250 RNASeq data set (Lanver et al. 2018), from which we extracted the expression profiles of genes in
251 the telomeric region of chromosome 9 (Figure 7A). While the expression of *UMAG_11064*
252 remained close to zero in the three replicates, expression of *UMAG_11065* increased during plant

253 infection. The telomeric region was highly heterogeneous in terms of expression profile: while
254 *UMAG_11066* and *UMAG_03393* did not show any significant level of expression, *UMAG_03392*
255 was down-regulated at twelve hours post-infection, while *UMAG_03394*, another RecQ-encoding
256 gene homologous to *UMAG_11065,* displayed constitutively high levels of expression (Figure 7A).
257 All homologs of *UMAG_11065* show a significantly higher expression during infection (Tukey's
258 posthoc test, false discovery rate of 5%, Figure 7B). The comparison of expression profiles revealed
259 two main classes of genes (Figure 7C): highly expressed genes (upper group), and moderately
260 expressed genes (lower group), to which *UMAG_11065* belongs. We further note that the
261 differences in expression profiles do not mirror the protein sequence similarity of the genes (Mantel
262 permutation test, p-value = 0.566).

263 To assess the functional role of *UMAG_11064* and *UMAG_11065*, these genes were
264 simultaneously deleted in SG200, a solopathogenic haploid strain that can cause disease without a
265 mating partner (Kämper et al. 2006) using a single-step gene replacement method (Kämper 2004).
266 Gene deletion was verified by Southern analysis (Figure S3). Virulence assays, conducted in
267 triplicate revealed no statistically different symptoms of the double deletion strain,
268 SG200Δ11065Δ11064, compared to SG200 in infected maize plants (Figure 8A, Chi-square test, p-
269 value = 0.453). Since RecQ helicases contribute to dealing with replication stress (Kojic and
270 Holloman 2012) we also determined the sensitivity of the mutant to various stressors including UV,
271 hydroxyurea and Congo Red. (Figure 8B). We report that the deletion strain shows increased
272 sensitivity to cell wall stress induced by Congo Red and increased resistance to UV stress. Since
273 *UMAG_11064* does not show any detectable level of expression, we hypothesize that the deletion of
274 *UMAG_11065* is responsible for this phenotype.

275 *Discussion*

276 The codon usage and GC content of the *UMAG_11064* gene, as well as its similarity to
277 known mitochondrial HEGs, points at a recent transfer into the nuclear genome of *U. maydis*.

278 Moreover, the precursor of this gene is absent from the mitochondrial genome of this species. Two
279 possible scenarios can explain this pattern, which we detail below.

280 The first scenario involves a transfer of the gene to the nuclear genome followed by a loss
281 of the mitochondrial copy (Figure 9). Under this scenario, the mitochondrial HEG was present in
282 the *U. maydis* ancestor. Two evolutionary events are invoked: the insertion of the HEG into the
283 nuclear genome, on the one hand, creating a HEG^+ genotype at the nuclear locus (designated
284 $[HEG^+]_{nuc}$), and the loss of the mitochondrial copy, creating a HEG^- genotype at the mitochondrial
285 locus (designated $[HEG^-]_{mit}$). These two events may have happened at distinct time points, but,
286 under this scenario, the former cannot have happened after the fixation of the $[HEG^-]_{mit}$ genotype in
287 the population. The $[HEG^+]_{nuc}$ / $[HEG^-]_{mit}$ genotype could be generated by a cross between two
288 individuals, one $[HEG^+]_{nuc}$ and the other $[HEG^-]_{mit}$, given that mitochondria are uniparentally
289 inherited in *U. maydis* (Basse 2010). Importantly, the segregation of the $[HEG^+]_{nuc}$ and $[HEG^-]_{mit}$
290 variants could be purely neutral and driven by genetic drift only. In case the $[HEG^-]_{nuc}$ allele
291 contained a recognition sequence of the HE, the $[HEG^+]_{nuc}$ allele may have initially benefited from a
292 genetic drive effect. Any putative selective advantage/disadvantage of the $[HEG^+]_{nuc}$ or $[HEG^-]_{mit}$
293 alleles may have favored their fixation, or on the contrary, acted against it.

294 In the second scenario, the mitochondrial HEG was not ancestral to *U. maydis*, but was
295 horizontally transferred from *S. reilianum* (or a related species). The high similarity of the
296 *UMAG_11064* gene to the *S. reilianum* mitochondrial HEG (Figure 3) supports this hypothesis,
297 given the relatively high nucleotide divergence between the two species, which diverged around 20
298 My ago (Schweizer et al. 2018). We note, however, that intronic HEGs have also been reported to
299 show reduced nucleotide substitution rates, which can potentially explain their comparatively low
300 divergence (Jalalzadeh et al. 2015). Group I introns have been reported to be highly mobile and to
301 undergo frequent horizontal gene transfers (HGT) within metazoans (Schuster et al. 2017), plants
302 (Sanchez-Puerta et al. 2008) and fungi (Jalalzadeh et al. 2015; Férandon et al. 2010). Group I

303 introns in metazoans and plants are also thought to originate from a fungal donor (Schuster et al.
304 2017; Sanchez-Puerta et al. 2008, 2011). In this respect, a transfer from another *cox1* intron-
305 carrying smut fungus to *U. maydis* is not unlikely, given that *U. maydis* and *S. reilianum* share the
306 same host, and that hybridization between smut species has been reported (Fischer 1957; Boidin
307 1986).

308 The insertion of the HEG in the nuclear genome poses the question of the underlying
309 mechanisms, independently of the origin of the HEG. First, the HEG could be encoding a fully
310 functional HE and the $[HEG^-]_{nuc}$ allele contained a HE recognition sequence. Under this scenario,
311 the insertion event was a homing event and the inactivation of the HEG occurred after the insertion.
312 Therefore, several generations must have passed since the insertion event in order for the
313 inactivating mutations to occur. Alternatively, the $[HEG^-]_{nuc}$ allele might not have contained a
314 recognition sequence and the insertion of the HEG occurred by an unknown mechanism. Lastly, a
315 possibility is that the inserted sequence already encoded an inactivated HE, which was then inserted
316 by an unknown mechanism. In this latter case, the insertion could have occurred very recently,
317 possibly a few generations in the past.

318 HEGs are found in eukaryotic nuclei but are usually restricted to small and large
319 ribosomal RNA subunit genes (Lambowitz and Belfort 1993; Dunin-Horkawicz et al. 2006). While
320 transfer of DNA segments and functional genes from organellar genomes to the nucleus is well
321 documented (Sun and Callis 1993; Thorsness and Weber 1996; Lloyd and Timmis 2011; Fuentes et
322 al. 2012), established examples of HEG insertions at other genomic locations than rRNA genes is
323 very scarce. Several questions remain unanswered regarding the mechanisms of insertion into the
324 nuclear genome, providing it happened as a homing event. For the event to happen, a template
325 sequence containing the HEG is required for repairing the break initiated by the HE. This template
326 must have, therefore, "leaked" from the mitochondrion. The recognition motif of the original
327 *UMAG_11064* HEG is unknown, and the very short flanking regions surrounding the insertion site

328 do not allow any comparison with known motifs, preventing further conclusions to be made
329 regarding the nature of the insertion event of *UMAG_11064*.

330 Interestingly, Louis and Haber (Louis and Haber 1991) reported a similar transfer of a
331 HEG into a telomeric region of *Saccharomyces cerevisiae*. The authors argue that signatures of such
332 insertion could be found because (i) it had no deleterious effect and (ii) the occurrence of
333 heterologous recombination between telomeres favours the maintenance of elements that would
334 otherwise be lost. Contrasting with this result, the insertion of the GIY-YIG HEG that inserted into
335 the ancestor of the *UMAG_11065* gene potentially had non-neutral effects, resulting in an expressed
336 truncated protein. Several mutations were found within the active site of the inserted HEG that led
337 to the *UMAG_11064* gene, suggesting that the encoded protein is unlikely to act as a HE any longer.
338 However, a putative alternative start codon was detected, downstream the active site, followed by
339 an uninterrupted peptide sequence containing the helix-turn-helix binding domain of the original
340 HE. Furthermore, we could not detect any significant level of expression of the *UMAG_11064* gene
341 in various laboratory conditions. Comparative sequence analysis further suggests that
342 *UMAG_11064* is evolving under relaxed purifying selection, indicating that it might be undergoing
343 pseudogenization. These results, therefore, suggest that the *UMAG_11064* gene is not functional.
344 However, as this mitochondrial HEG inserted into a nuclear *U. maydis* gene, it might have had
345 phenotypic consequences not directly due to the HEG gene itself. The *UMAG_11065* gene appeared
346 truncated by the HEG insertion, which removed the C-terminal part of the encoded protein, a likely
347 RecQ helicase, and the truncated *UMAG_11065* is expressed during infection. The truncation likely
348 did not have a strong negative impact, possibly because of the existence of multiple potentially
349 functionally redundant paralogs of *UMAG_11065,* including on the same telomeric region of
350 chromosome 9, with *UMAG_03394* being located 4 genes upstream (Table 1). While we were
351 unable to detect a contribution to virulence, our results point at a putative role of the truncated RecQ
352 helicase into stress tolerance, as its deletion increases resistance to UV radiation but makes the

353 fungus more susceptible to cell wall stress, at least under laboratory conditions. How the truncated
354 UMAG_11065 RecQ helicase could improve coping with cell wall stress and increases the
355 sensitivity to UV simultaneously, however, remains to be investigated, as well as the potential
356 fitness benefit or cost of these phenotypes. Furthermore, a possibility remains that the observed
357 phenotype of *UMAG_11065* is ancestral and not due to the truncation itself, which could be neutral.
358 In order to elucidate the putative adaptive role of the truncation of *UMAG_11065*, knowledge of the
359 ancestral, non-truncated *UMAG_11065* allele is needed, as well as its distribution in natural
360 populations.

361 Conclusions

362 In this study, we report instances of two stages of the life cycle of HEGs. Intron 1 of the
363 mitochondrial *cox1* gene of *S. reilianum* was shown to contain a degenerated GIY-YIG HEG, while
364 the homologous position in the *U. maydis* gene displays no intron. Besides, in the telomeric region
365 of chromosome 9 of the nuclear genome of *U. maydis*, we found evidence of a recent insertion of a
366 very similar GIY-YIG HEG. Phenotypic assay of the mutant strain containing a double-deletion of
367 the HEG and the helicase gene where it inserted reveals enhanced stress sensitivity *in vitro*. The
368 absence of a GIY-YIG HEG in any field isolates of *U. maydis* sequenced so far, however, suggests
369 that either the mutation was lost in natural populations and only maintained under laboratory
370 conditions, or that it is only present in a so far unsampled population. The *UMAG_11064* gene
371 offers a snapshot of evolution taken soon after a mutation event occurred. As such, it can provide
372 insights into the mechanisms of HEG mobility and horizontal transfer. These results further
373 demonstrate that HEGs can generate genetic diversity not only via their duplication, but also by
374 drastically modifying the local genome architecture where they insert.

375 ## *Material and Methods*

376 Data sets and scripts necessary to reproduce the statistical analyses in this work are
377 available as Supplementary File S1 and at https://gitlab.gwdg.de/molsysevol/umag_11064 .
378

379 ## Analysis of codon usage and GC content

380 *Ustilago maydis* gene models (genome version 2.0) were retrieved from the MIPS
381 database (Mewes et al. 2011). Mitochondrial genes were extracted from the *U. maydis* full
382 mitochondrial genome (Genbank accession number: NC_008368.1). Within-group correspondence
383 analysis of synonymous codon usage was performed using the ade4 package for R, following the
384 procedure described in (Charif et al. 2005). The proportion of G and C nucleotides was computed
385 along with the first 10 kb of *U. maydis* chromosome 9, using 300 bp windows slid by 1 bp. The
386 corresponding R code is available as Supplementary File S1.

387 ## Strains, growth conditions and virulence assays

388 The *Escherichia coli* strains DH5α (Bethesda Research Laboratories) and TOP10 (Life
389 Technologies, Carlsbad, CA, USA) were used for the cloning and amplification of plasmids. *U.*
390 *maydis* strains 518 and 521 are the parents of FB1 and FB2 (Banuett and Herskowitz 1989). SG200
391 is a hapoid solopathogenic strain derived from FB1 (Kämper et al. 2006). 10-1 is an uncharacterized
392 haploid *U. maydis* strain isolated in the US and kindly provided by G. May. I2, O2, P2, S5, and T6
393 are haploid *U. maydis* strains collected in different parts of Mexico (Valverde et al. 2000). The
394 haploid *S. reilianum* strains SRZ1 and SRZ2 as well as the solopathogenic strain JS161 derived
395 from SRZ1 have been described (Schirawski et al. 2010). Deletion mutants were generated by gene
396 replacement using a PCR-based approach and verified by Southern analysis (Kämper 2004).
397 pRS426Δum11064+11065 is a pRS426-derived plasmid containing the *UMAG_11064*/

398 *UMAG_11065* double deletion construct which consists of a hygromycin resistance cassette flanked
399 by the left border of the *UMAG_11064* and right border of the *UMAG_11065* gene. The left border
400 of *UMAG_11064* and the right border of *UMAG_11065* were PCR amplified from SG200 gDNA
401 with primers um11064_lb_fw/um11064_lb_rv and um11065_rb_fw/um11065_rb_rv
402 (Supplementary Table S7). The hygromycin resistance cassette was obtained from SfiI digested
403 pHwtFRT (Khrunyk et al. 2010). The pRS426 EcoRI/XhoI backbone, both borders and the
404 resistance cassette were assembled using yeast drag and drop cloning (Christianson et al. 1992). The
405 fragment containing the deletion cassette was amplified from this plasmid using primers
406 um11064_lb_fw and um11065_rb_rv (Supplementary Table S7), transformed into SG200 and
407 transformants carrying a deletion of *UMAG_11064* and *UMAG_11065* were identified by southern
408 analysis (Figure S3).

409 *U. maydis* strains were grown at 28°C in liquid YEPSL medium (0.4% yeast extract, 0.4%
410 peptone, 2% sucrose) or on PD solid medium (2.4% Potato Dextrose broth, 2% agar). Stress assays
411 were performed as described in (Krombach et al. 2018). Transformation and selection of *U. maydis*
412 transformants followed published procedures (Kämper et al. 2006). To assess virulence, seven day
413 old maize seedlings of the maize variety Early Golden Bantam (Urban Farmer, Westfield, Indiana,
414 USA) were syringe-infected. At least three independent infections were carried out and disease
415 symptoms were scored according to Kämper et al. (Kämper et al. 2006). Consistence of replicates
416 was tested using a chi-squared test and p-values were computed using 1,000,000 permutations. As
417 no significant difference between replicates was observed (p-value = 0.347 for the wildtype and p-
418 value = 0.829 for the deletion strain), observation were pooled between all replicates for each strain
419 before being compared.

420 Blast searches and gene alignment

421 We performed BlastN and BlastP (Altschul et al. 1990) searches using the (translated)
422 sequence of *UMAG_11064* as a query using NCBI online blast tools. The non-redundant nucleotide

423 and protein sequence databases were selected for BlastN and BlastP, respectively. Results were
424 further processed with scripts using the NCBIXML module from BioPython (Cock et al. 2009). The
425 Macse codon aligner (Ranwez et al. 2011) was used in order to infer the position of putative
426 frameshifts in the upstream region of *UMAG_11064*. The alignment was depicted using the
427 Boxshade software and was further manually annotated. The sequences of *U. maydis cox1* intron 6,
428 as well as *S. reilianum cox1* introns 1 and 2 were used as query and searched against the protein non
429 redundant database using NCBI BlastX, excluding environmental samples and model sequences.
430 The *cox1* genes from *U. maydis* and *S. reilianum* were aligned and pairwise similarity was
431 computed in non-overlapping 100 bp windows (Supplementary File S1). The gene structure,
432 synteny and local pairwise similarity was depicted using the genoPlotR package for R (Guy et al.
433 2010).

434 Phylogeny estimation, estimation of dN/dS ratios and tests of positive 435 selection

436 The nucleotide sequence of *UMAG_11064*, the first intron of the *cox1* gene of *S.*
437 *reilianum*, and the eight non-redundant, most similar matches from BlastP (Table S2) were aligned
438 using the Macse codon aligner (Ranwez et al. 2011) together with the unannotated but similar
439 nucleotide sequences from *S. scitamineum*, *U. bromivora*, *T. indica* and *T. walkiri*, using the NCBI
440 codon translation table 4 "mitochondrial mold". Columns in the alignment were manually selected
441 to discard ambiguously aligned regions and a phylogeny was inferred using PhyML (Guindon et al.
442 2010) with a General Time Reversible (GTR) model of nucleotide evolution and a 4-classes discrete
443 gamma distribution of rates. The tree topology was inferred using the "best of nearest-neighbor-
444 interchange (NNI) and subtree-pruning-regrafting (SPR)" option, and 100 non-parametric bootstrap
445 replicated were obtained. The final tree was rooted using the midpoint method. Analyses were
446 performed using the Seaview software (Gouy et al. 2010). For the positive selection analysis, the *S.*

447 *scitamineum* and *U. bromivera* sequences were discarded as they contained multiple frameshifts. A
448 phylogenetic tree was estimated using PhyML from the remaining species after translation using a
449 Le and Gascuel model of protein evolution (Le and Gascuel 2008), and other parameters as for the
450 nucleotide model. Nodes with bootstrap support values lower than 65% were collapsed. A branch
451 model of codon evolution was fitted on the alignment and the inferred phylogenetic tree using
452 PAML 4.9d (Yang 2007), keeping selected positions that may contain missing data. The F3X4
453 codon frequencies model was selected, and one dN/dS ratio was estimated per branch. A branch-site
454 model (Zhang et al. 2005) was fitted by specifying the branch leading to the *UMAG_11064* gene as
455 the "foreground" group, putatively evolving under positive selection. Test for selection was
456 performed as suggested in the PAML manual, comparing to a model where the omega2 parameter is
457 fixed to a value of 1. A similar test was conducted after excluding the two *Tilletia* sequences from
458 the "background" branches, as they were found to have each a branch with dN/dS > 1.

459 Amplification of the *UMAG_11064* regions in several *U. maydis* strains

460 Amplification of DNA fragments via polymerase chain reaction (PCR) was done using
461 the Phusion High Fidelity DNA_Polymerase (Thermo Fisher Scientific, Waltham, USA). The PCR
462 reactions were set up in a 20 µl reaction volume using DNA templates indicated in the respective
463 experiments and buffer recommended by the manufacturer containing a final concentration of 3%
464 DMSO. The PCR programs used are represented by the following scheme: Initial denaturation –
465 [denaturation – annealing – elongation] x number cycles – final elongation. *UMAG_11072* was
466 amplified with primers um11072_ORF_fw x um11072_ORF_rv using 98 °C/3 m - [98 °C/10 s – 65
467 °C/30 s - 72 °C/45 s] x 30 cycles - 72 °C/10 m. *UMAG_11064* was amplified with primers
468 um11064_ORF_fw x um11064_ORF_rv using 98 °C/3 m - [98 °C/10 s – 65 °C/30 s - 72 °C/45 s] x
469 30 cycles - 72 °C/10 m. The *cox1* exons 1+2 were amplified with primers cox1_ex1_rv x
470 cox1_ex2_fw using 98 °C/3 m - [98 °C/10 s – 63 °C/30 s - 72 °C/90 s] x 33 cycles - 72 °C/10 m.

471 cox1 exon 7 was amplified with primers cox1_ex7_fw X cox1_ex7_rv using 98 °C/3 m - [98 °C/10
472 s – 67 °C/30 s - 72 °C/60 s] x 30 cycles - 72 °C/10 m. Parts of the genomic region containing
473 *UMAG_11064*, *UMAG_11065* and *UMAG_11066* were amplified with primer pairs um11064_fw1 x
474 um11064_rv1, um11064_fw1 x um11064_rv2; and um11064_ fw2 x um11064_rv2 using 98 °C/3 m
475 - [98 °C/10 s – 65 °C/30 s - 72 °C/150 s] x 32 cycles - 72 °C/10 m. The list of all primer sequences
476 is provided in Supplementary Table S7. PCR results are shown in Figures S1 and S2.

477 History of the *UMAG_11065* family

478 The sequence of the *UMAG_11065* protein was used as a query for a search against
479 several smut fungi (*U. maydis*, *U. hordei*, *S. reilianum*, *S. scitamineum*, *Melanopsichum*
480 *pennsylvanicum*, *Pseudozyma flocculosa*) complete proteome using BlastP (Altschul et al. 1990).
481 The search finds 17 hits within the *U. maydis* genome with an E-value below 0.0001, as well as two
482 genes in *S. scitamineum* (*SPSC_04622* and *SPSC_05783*) and two genes in *P. flocculosa*
483 (*PFL1_06135* and *PFL1_02192*). Using NCBI BlastP, we found several sequences from *Fusarium*
484 *oxyparum* with high similarity. We selected the sequence *FOXG_04692* as a representative and
485 added it to the data set. The Guidance web server with the GUIDANCE2 algorithm was then used
486 to align the protein sequences and assess the quality of the resulting alignment. Default options
487 from the server were kept, selecting the MAFFT aligner (Katoh et al. 2002). Several sequences
488 appeared to be of shallow alignment quality and were discarded. The remaining sequences were
489 realigned using the same protocol. Four iterations were performed until the final alignment had a
490 quality good enough for phylogenetic inference. The final alignment contained 14 sequences and
491 had a global score of 0.79. These 14 alignable sequences contained 13 *U. maydis* sequences
492 (including *UMAG_11065*), and the *F. oxysporum* gene, other sequences from smut genomes were
493 too divergent to be unambiguously aligned. Using Guidance, we further masked columns in the
494 alignment with a score below 0.93 (a maximum of one position out of 14 in the column was
495 allowed to be uncertain).

496 A phylogenetic analysis was conducted using the program Seaview 4 (Gouy et al. 2010).
497 First, a site selection was performed in order to filter regions with too many gaps, leaving 506 sites.
498 Second, a phylogenetic tree was built using PhyML within Seaview (Guindon et al. 2010) (Le and
499 Gascuel protein substitution model (Le and Gascuel 2008) with a four-classes discretized gamma
500 distribution of rates, the best tree of Nearest Neigbour Interchange (NNI) and Subtree Pruning and
501 Regrafting (SPR) topological searches was kept). Support values were computed using the
502 approximate likelihood ratio test (aLRT) method (Anisimova and Gascuel 2006).
503 A test for positive selection was conducted using a combination of branch and branch-site
504 models using PAML (Yang 2007). The final GUIDANCE alignment was used, realigned using the
505 Macse codon aligner (Ranwez et al. 2011), and ambiguously aligned sites and shorter sequences
506 were manually filtered. The final alignment contained the following sequences: *UMAG_03394*,
507 *UMAG_11065*, *UMAG_04486*, *UMAG_06506*, *UMAG_04094*, *UMAG_10585*, *UMAG_06474*,
508 *UMAG_10980*, *UMAG_05977*, *FOXG_04692*. We used the PhyML software with the same options
509 as described above to reconstruct a phylogenetic tree with this subset of sequences. The branch
510 toward the *UMAG_11065* gene was used as a foreground group in the branch site model.

511 Gene expression

512 RNASeq normalized expression counts for the *UMAG_11064* and *UMAG_11065*, as well
513 as of neighbouring genes and paralogs elsewhere in the genome, were extracted from the Gene
514 Expression Omnibus data set GSE103876 (Lanver et al. 2018). Gene clustering based on expression
515 profiles was conducted using a hierarchical clustering with an average linkage on a Canberra
516 distance, suitable for expression counts, as implemented in the 'dist' and 'hclust' functions in R (R
517 Core Team 2018). The resulting clustering tree was converted to a distance matrix and compared to
518 the inferred phylogeny of the genes using a Mantel permutation test, as implemented in the 'ape'
519 package for R (Paradis et al. 2004). Differences in expression between time points were assessed by
520 fitting the linear model "expression ~ time * gene", testing the effect of time while controlling for

521 interaction with the "gene" variable. Residuals were normalized using a Box-Cox transform as
522 implemented in the MASS package for R. Tukey's posthoc comparisons were conducted on the
523 resulting model, allowing for a 5% false discovery rate.

524 Acknowledgments

525 We thank all members of our groups for stimulating discussions. We are grateful to
526 Georgiana May and Octavio Paredes-López for providing field isolates of *U. maydis* from the US
527 and Mexico, respectively. We acknowledge the generous support by the Max Planck Society.

528 Conflict of interest disclosure

529 The authors of this study declare that they have no financial conflict of interest with the
530 content of this article.

531 References

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

Anisimova M, Gascuel O. 2006. Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative. *Syst Biol* **55**: 539–552.

Ast J, Stiebler AC, Freitag J, Bölker M. 2013. Dual targeting of peroxisomal proteins. *Front Physiol* **4**: 297.

Banuett F, Herskowitz I. 1989. Different a alleles of *Ustilago maydis* are necessary for maintenance of filamentous growth but not for meiosis. *Proc Natl Acad Sci U S A* **86**: 5878–5882.

Barzel A, Obolski U, Gogarten JP, Kupiec M, Hadany L. 2011. Home and away- the evolutionary dynamics of homing endonucleases. *BMC Evol Biol* **11**: 324.

Basse CW. 2010. Mitochondrial inheritance in fungi. *Curr Opin Microbiol* **13**: 712–719.

Belfort M, Roberts RJ. 1997. Homing endonucleases: keeping the house in order. *Nucleic Acids Res* **25**: 3379–3388.

Boidin J. 1986. Intercompatibility and the species concept in the saprobic basidiomycotina. *Mycotaxon* **XXVI**: 319–336.

Charif D, Thioulouse J, Lobry JR, Perrière G. 2005. Online synonymous codon usage analyses with

the ade4 and seqinR packages. *Bioinforma Oxf Engl* **21**: 545–547.

Chevalier BS, Stoddard BL. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. *Nucleic Acids Res* **29**: 3757–3774.

Cho Y, Qiu Y-L, Kuhlman P, Palmer JD. 1998. Explosive invasion of plant mitochondria by a group I intron. *Proc Natl Acad Sci* **95**: 14244–14249.

Christianson TW, Sikorski RS, Dante M, Shero JH, Hieter P. 1992. Multifunctional yeast high-copy-number shuttle vectors. *Gene* **110**: 119–122.

Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al. 2009. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinforma Oxf Engl* **25**: 1422–1423.

Derbyshire V, Kowalski JC, Dansereau JT, Hauer CR, Belfort M. 1997. Two-domain structure of the td intron-encoded endonuclease I-TevI correlates with the two-domain configuration of the homing site. *J Mol Biol* **265**: 494–506.

Djamei A, Kahmann R. 2012. *Ustilago maydis*: dissecting the molecular interface between pathogen and plant. *PLoS Pathog* **8**: e1002955.

Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. *Curr Opin Genet Dev* **35**: 57–65.

Dujon B, Belfort M, Butow RA, Jacq C, Lemieux C, Perlman PS, Vogt VM. 1989. Mobile introns: definition of terms and recommended nomenclature. *Gene* **82**: 115–118.

Dunin-Horkawicz S, Feder M, Bujnicki JM. 2006. Phylogenomic analysis of the GIY-YIG nuclease superfamily. *BMC Genomics* **7**: 98.

Dutheil JY, Mannhaupt G, Schweizer G, Sieber CMK, Münsterkötter M, Güldener U, Schirawski J, Kahmann R. 2016. A Tale of Genome Compartmentalization: The Evolution of Virulence Clusters in Smut Fungi. *Genome Biol Evol* **8**: 681–704.

Eyre-Walker A, Keightley PD. 2007. The distribution of fitness effects of new mutations. *Nat Rev Genet* **8**: 610–618.

Férandon C, Moukha S, Callac P, Benedetto J-P, Castroviejo M, Barroso G. 2010. The Agaricus bisporus cox1 Gene: The Longest Mitochondrial Gene and the Largest Reservoir of Mitochondrial Group I Introns. *PLoS ONE* **5**. http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2987802/ (Accessed December 19, 2012).

Fischer C. 1957. *Biology and Control of Smut Fungi*. John Wiley & Sons Canada, Limited.

Fuentes I, Karcher D, Bock R. 2012. Experimental reconstruction of the functional transfer of intron-containing plastid genes to the nucleus. *Curr Biol CB* **22**: 763–771.

Fukami H, Chen CA, Chiou C-Y, Knowlton N. 2007. Novel group I introns encoding a putative homing endonuclease in the mitochondrial cox1 gene of Scleractinian corals. *J Mol Evol* **64**: 591–600.

Gladyshev E. 2017. Repeat-Induced Point Mutation and Other Genome Defense Mechanisms in Fungi. *Microbiol Spectr* **5**.

Goddard MR, Burt A. 1999. Recurrent invasion and extinction of a selfish gene. *Proc Natl Acad Sci U S A* **96**: 13880–13885.

Gogarten JP, Hilario E. 2006. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. *BMC Evol Biol* **6**: 94.

Gouy M, Guindon S, Gascuel O. 2010. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol Biol Evol* **27**: 221–224.

Grandaubert J, Lowe RGT, Soyer JL, Schoch CL, Van de Wouw AP, Fudal I, Robbertse B, Lapalu N, Links MG, Ollivier B, et al. 2014. Transposable element-assisted evolution and adaptation to host plant within the Leptosphaeria maculans-Leptosphaeria biglobosa species complex of fungal pathogens. *BMC Genomics* **15**: 891.

Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**: 307–321.

Guy L, Kultima JR, Andersson SGE. 2010. genoPlotR: comparative gene and genome visualization in R. *Bioinforma Oxf Engl* **26**: 2334–2335.

Jalalzadeh B, Saré IC, Férandon C, Callac P, Farsi M, Savoie J-M, Barroso G. 2015. The intraspecific variability of mitochondrial genes of Agaricus bisporus reveals an extensive group I intron mobility combined with low nucleotide substitution rates. *Curr Genet* **61**: 87–102.

Kaessmann H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res* **20**: 1313–1326.

Kämper J. 2004. A PCR-based system for highly efficient generation of gene replacement mutants in *Ustilago maydis*. *Mol Genet Genomics MGG* **271**: 103–110.

Kämper J, Kahmann R, Bölker M, Ma L-J, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O, et al. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444**: 97–101.

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.

Khrunyk Y, Münch K, Schipper K, Lupas AN, Kahmann R. 2010. The use of FLP-mediated recombination for the functional analysis of an effector gene family in the biotrophic smut fungus *Ustilago maydis*. *New Phytol* **187**: 957–968.

Kojic M, Holloman WK. 2012. Brh2 domain function distinguished by differential cellular responses to DNA damage and replication stress. *Mol Microbiol* **83**: 351–361.

Kondrashov FA, Kondrashov AS. 2010. Measurements of spontaneous rates of mutations in the recent past and the near future. *Philos Trans R Soc Lond B Biol Sci* **365**: 1169–1176.

Krombach S, Reissmann S, Kreibich S, Bochen F, Kahmann R. 2018. Virulence function of the *Ustilago maydis* sterol carrier protein 2. *New Phytol* **220**: 553–566.

Lambowitz AM, Belfort M. 1993. Introns as mobile genetic elements. *Annu Rev Biochem* **62**: 587–622.

Lanver D, Müller AN, Happel P, Schweizer G, Haas FB, Franitza M, Pellegrin C, Reissmann S, Altmüller J, Rensing SA, et al. 2018. The Biotrophic Development of *Ustilago maydis* Studied by RNA-Seq Analysis. *Plant Cell* **30**: 300–323.

Laurie JD, Ali S, Linning R, Mannhaupt G, Wong P, Güldener U, Münsterkötter M, Moore R, Kahmann R, Bakkeren G, 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.

Le SQ, Gascuel O. 2008. An improved general amino acid replacement matrix. *Mol Biol Evol* **25**: 1307–1320.

Lloyd AH, Timmis JN. 2011. The origin and characterization of new nuclear genes originating from a cytoplasmic organellar genome. *Mol Biol Evol* **28**: 2019–2028.

Louis EJ, Haber JE. 1991. Evolutionarily recent transfer of a group I mitochondrial intron to telomere regions in Saccharomyces cerevisiae. *Curr Genet* **20**: 411–415.

Lynch M. 2007. *The Origins of Genome Architecture*. Sinauer Associates.

Lynch M, Sung W, Morris K, Coffey N, Landry CR, Dopman EB, Dickinson WJ, Okamoto K, Kulkarni S, Hartl DL, et al. 2008. A genome-wide view of the spectrum of spontaneous mutations in yeast. *Proc Natl Acad Sci U S A* **105**: 9272–9277.

Mewes HW, Ruepp A, Theis F, Rattei T, Walter M, Frishman D, Suhre K, Spannagl M, Mayer KFX, Stümpflen V, et al. 2011. MIPS: curated databases and comprehensive secondary data resources in 2010. *Nucleic Acids Res* **39**: D220-224.

Möller M, Stukenbrock EH. 2017. Evolution and genome architecture in fungal plant pathogens. *Nat Rev Microbiol* **15**: 756–771.

Munkacsi AB, Stoxen S, May G. 2008. *Ustilago maydis* populations tracked maize through domestication and cultivation in the Americas. *Proc Biol Sci* **275**: 1037–1046.

Ohta T. 2000. Evolution of gene families. *Gene* **259**: 45–52.

Paradis E, Claude J, Strimmer K. 2004. APE: Analyses of Phylogenetics and Evolution in R language. *Bioinforma Oxf Engl* **20**: 289–290.

Perrière G, Thioulouse J. 2002. Use and misuse of correspondence analysis in codon usage studies. *Nucleic Acids Res* **30**: 4548–4555.

Pogoda CS, Keepers KG, Nadiadi AY, Bailey DW, Lendemer JC, Tripp EA, Kane NC. 2019. Genome streamlining via complete loss of introns has occurred multiple times in lichenized fungal mitochondria. *Ecol Evol* **9**: 4245–4263.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing https://www.R-project.org/.

Ranwez V, Harispe S, Delsuc F, Douzery EJP. 2011. MACSE: Multiple Alignment of Coding SEquences accounting for frameshifts and stop codons. *PloS One* **6**: e22594.

Sánchez-Alonso P, Guzmán P. 1998. Organization of chromosome ends in *Ustilago maydis*. RecQ-like helicase motifs at telomeric regions. *Genetics* **148**: 1043–1054.

Sanchez-Puerta MV, Abbona CC, Zhuo S, Tepe EJ, Bohs L, Olmstead RG, Palmer JD. 2011. Multiple recent horizontal transfers of the cox1 intron in Solanaceae and extended co-conversion of flanking exons. *BMC Evol Biol* **11**: 277.

Sanchez-Puerta MV, Cho Y, Mower JP, Alverson AJ, Palmer JD. 2008. Frequent, phylogenetically local horizontal transfer of the cox1 group I Intron in flowering plant mitochondria. *Mol Biol Evol* **25**: 1762–1777.

Schirawski J, Mannhaupt G, Münch K, Brefort T, Schipper K, Doehlemann G, Di Stasio M, Rössel N, Mendoza-Mendoza A, Pester D, et al. 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* **330**: 1546–1548.

Schuster A, Lopez JV, Becking LE, Kelly M, Pomponi SA, Wörheide G, Erpenbeck D, Cárdenas P. 2017. Evolution of group I introns in Porifera: new evidence for intron mobility and implications for DNA barcoding. *BMC Evol Biol* **17**: 82.

Schweizer G, Münch K, Mannhaupt G, Schirawski J, Kahmann R, Dutheil JY. 2018. Positively Selected Effector Genes and Their Contribution to Virulence in the Smut Fungus Sporisorium reilianum. *Genome Biol Evol* **10**: 629–645.

Singh DK, Ghosh AK, Croteau DL, Bohr VA. 2012. RecQ helicases in DNA double strand break repair and telomere maintenance. *Mutat Res* **736**: 15–24.

Steinberg G, Perez-Martin J. 2008. *Ustilago maydis*, a new fungal model system for cell biology. *Trends Cell Biol* **18**: 61–67.

Stoddard BL. 2005. Homing endonuclease structure and function. *Q Rev Biophys* **38**: 49–95.

Stone CL, Frederick RD, Tooley PW, Luster DG, Campos B, Winegar RA, Melcher U, Fletcher J, Blagden T. 2018. Annotation and analysis of the mitochondrial genome of Coniothyrium glycines, causal agent of red leaf blotch of soybean, reveals an abundance of homing endonucleases. *PLOS ONE* **13**: e0207062.

Sun CW, Callis J. 1993. Recent stable insertion of mitochondrial DNA into an Arabidopsis polyubiquitin gene by nonhomologous recombination. *Plant Cell* **5**: 97–107.

Thorsness PE, Weber ER. 1996. Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. *Int Rev Cytol* **165**: 207–234.

Valverde ME, Vandemark GJ, Martínez O, Paredes-López O. 2000. Genetic diversity of *Ustilago maydis* strains. *World J Microbiol Biotechnol* **16**: 49–55.

Volff J-N. 2006. Turning junk into gold: domestication of transposable elements and the creation of

new genes in eukaryotes. *BioEssays News Rev Mol Cell Dev Biol* **28**: 913–922.

Vollmeister E, Schipper K, Baumann S, Haag C, Pohlmann T, Stock J, Feldbrügge M. 2012. Fungal development of the plant pathogen *Ustilago maydis*. *FEMS Microbiol Rev* **36**: 59–77.

Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**: 1586–1591.

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.

532

533 ## *Tables*

534 **Table 1**: *UMAG_11065* paralogs in *U. maydis*, together with a homolog from *F. oxysporum* for

535 comparison.

Gene	Chr / Scaffold / Contig	Start	End	Length of Chr / Scaffold / Contig	Number of introns	Length of protein	Relative position (1)
UMAG_06476	Chromosome 3	1641500	1642057	1642070	0	185	99.98%
UMAG_06474	Chromosome 3	1639598	1640203	1642070	0	201	99.87%
UMAG_06506	Chromosome 7	951043	954234	957188	5	983	99.52%
UMAG_10585	Chromosome 4	883585	884046	884984	0	153	99.87%
UMAG_11065	Chromosome 9	1886	1263	733962	0	207	0.21%
UMAG_03394	Chromosome 9	8836	5960	733962	0	958	1.01%
UMAG_03869	Chromosome 10	687301	690648	692354	7	937	99.51%
UMAG_04094	Chromosome 11	688670	689965	690620	0	431	99.81%
UMAG_04486	Chromosome 14	605233	609089	611467	2	1175	99.30%
UMAG_04308	Chromosome 14	1241	87	611467	0	384	0.11%
UMAG_05977	Chromosome 20	523510	523884	523884	0	124	99.96% (2)
UMAG_10980	Chromosome 22	398220	400499	403590	0	759	98.95%
UMAG_12076	Contig 1.265	4214	5343	5343	0	376	89.43%
FOXG_04692	Supercontig 2.5	9736	6398	2688632	0	1112	0.30%

536 (1) Position reported to the length of the chromosome or contig.

537 (2) N-terminal fragment only.

538

539 **Supplementary Table S1:** Homology search results using *UMAG_11064* as a query on the NCBI

540 non-redundant nucleotide database, using BlastN. All hits with an E-value lower than 1E-04 are

541 included, alongside with corresponding alignment length and percentage of sequence identity.

542

543 **Supplementary Table S2:** Homology search results using *UMAG_11064* as a query on NCBI non-

544 redundant protein database, using BlastP. All hits with an E-value lower than 1E-04 are included,

545 alongside with corresponding alignment length and percentage of sequence identity.

546

547 **Supplementary Table S3:** Homology search results using *UMAG_11065* as a query on NCBI non-

548 redundant protein database, using BlastP. All hits with an E-value lower than 1E-04 are included,

549 alongside with corresponding alignment length and percentage of sequence identity.

550

551 **Supplementary Table S4:** Homology search results using *U. maydis cox1* intron 6 as a query on a
552 NCBI non-redundant protein database, using BlastX. All hits with an E-value lower than 1E-04 are
553 included, alongside with corresponding alignment length and percentage of sequence identity.

554

555 **Supplementary Table S5:** Homology search results using *S. reilianum cox1* intron 1 as a query on
556 a NCBI non-redundant protein database, using BlastX. All hits with an E-value lower than 1E-04
557 are included, alongside with corresponding alignment length and percentage of sequence identity.

558

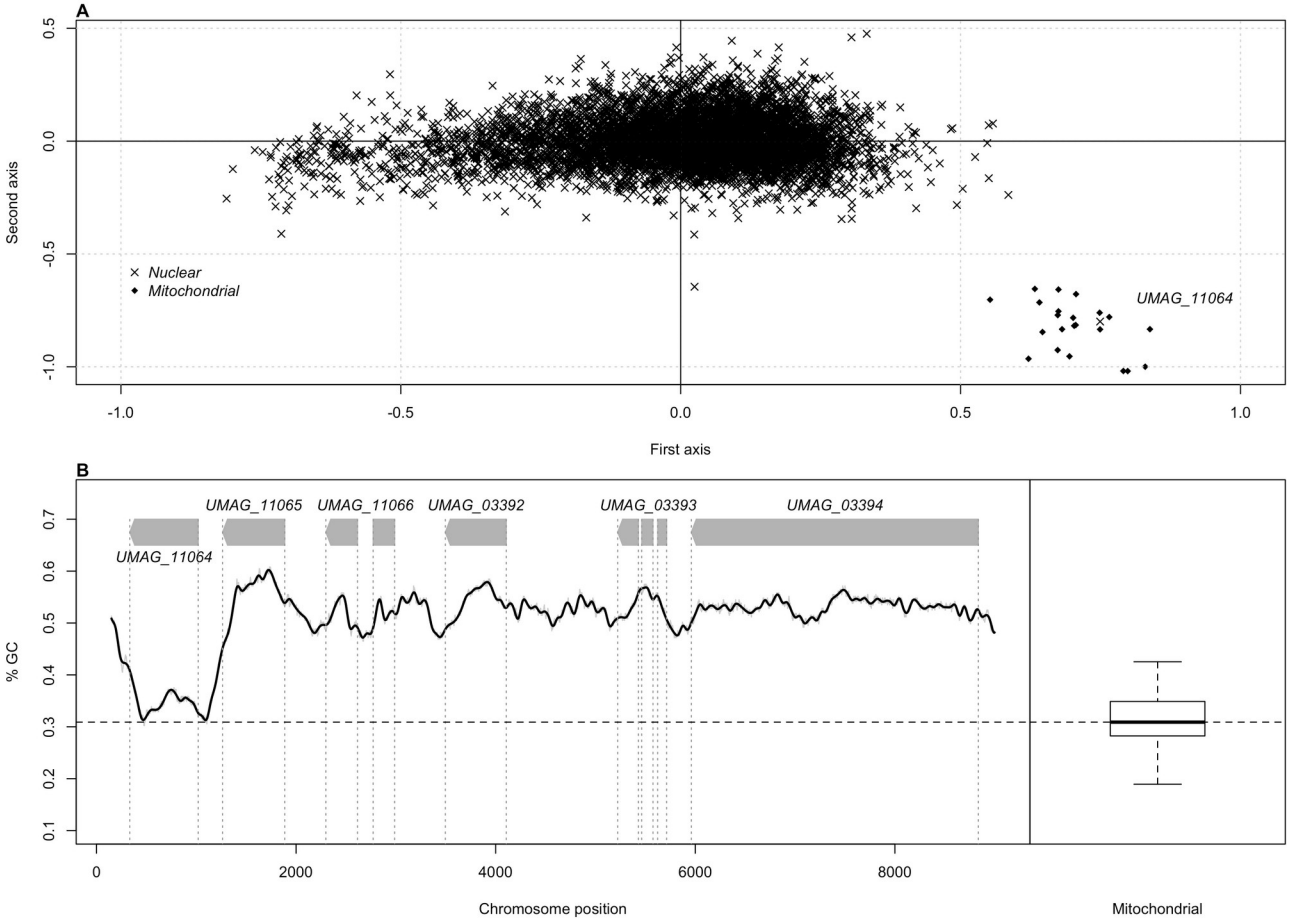
559 **Supplementary Table S6:** Homology search results using *S. reilianum cox1* intron 2 as a query on
560 a NCBI non-redundant protein database, using BlastX. All hits with an E-value lower than 1E-04
561 are included, alongside with corresponding alignment length and percentage of sequence identity.

562

563 **Supplementary Table S7:** Primers used in this study.

564

565 ***Figures***



566 **Figure 1:** Identification of the *UMAG_11064* gene. A) Within-group correspondence analysis of *U.*
567 *maydis* codon usage. Each gene is represented according to its coordinates along the first two
568 principal factors. The genomic origin of each gene is indicated by a cross for nuclear genes and a
569 dot for mitochondrial genes. B) Genomic context of the gene *UMAG_11064*. GC content in 300 bp
570 windows sliding by 1 bp, and distribution of GC content in 300 bp windows of mitochondrial
571 genome of *U. maydis*. The dash line represents the median of the distribution.

572

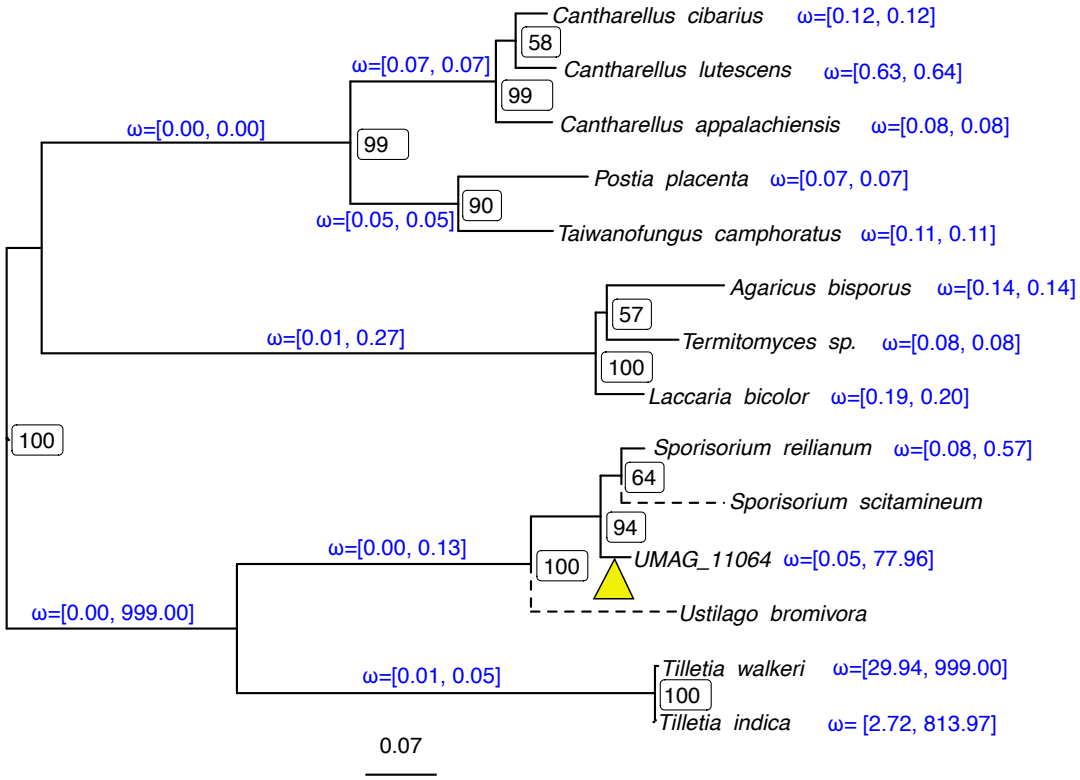


Figure 2: Phylogeny of *UMAG_11064* and its homologous sequences. Maximum likelihood tree inferred from nucleotides sequences. Node labels show bootstrap support values. Discontinuous branches indicate that the corresponding sequence is pseudogenized, with multiple frameshifts and insertions/deletions. Branch annotations show the minimum and maximum dN/dS ratio (ω) estimated from 10 independent runs of the codeml program. The yellow triangle indicates the supposed branch where the frameshift within the active domain of the ancestral HE occurred (See Figure 3).

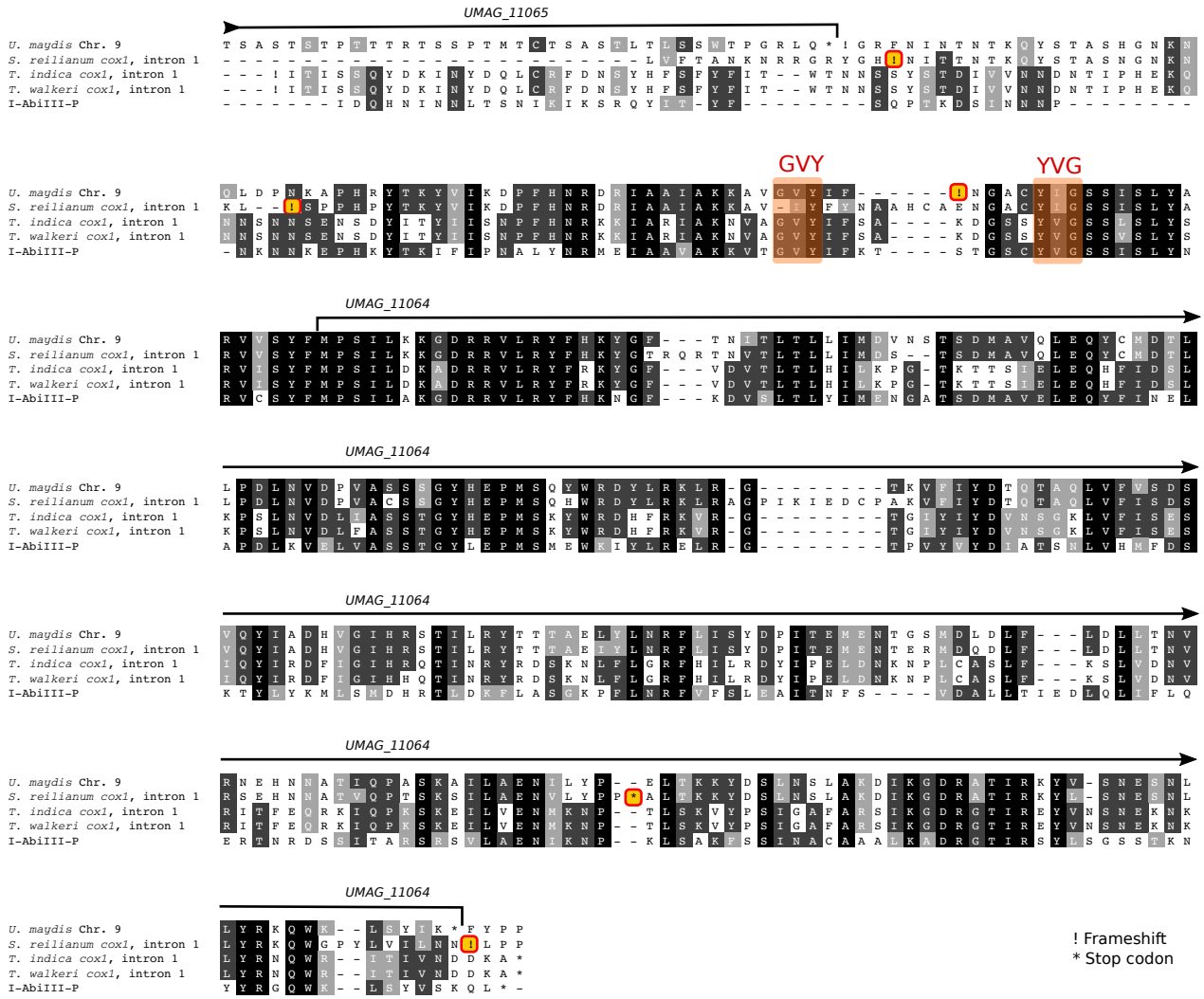
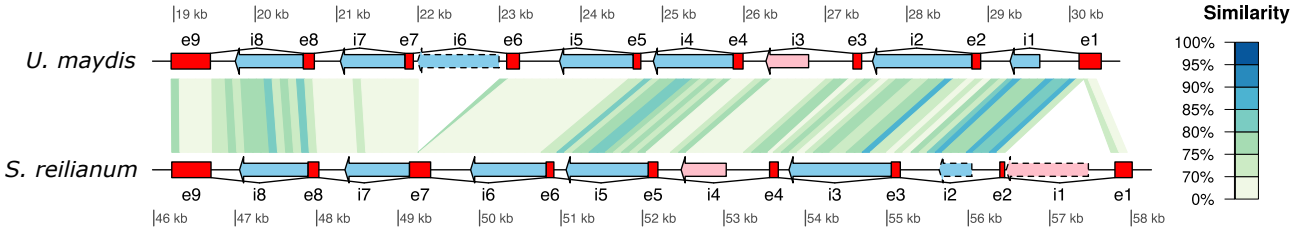


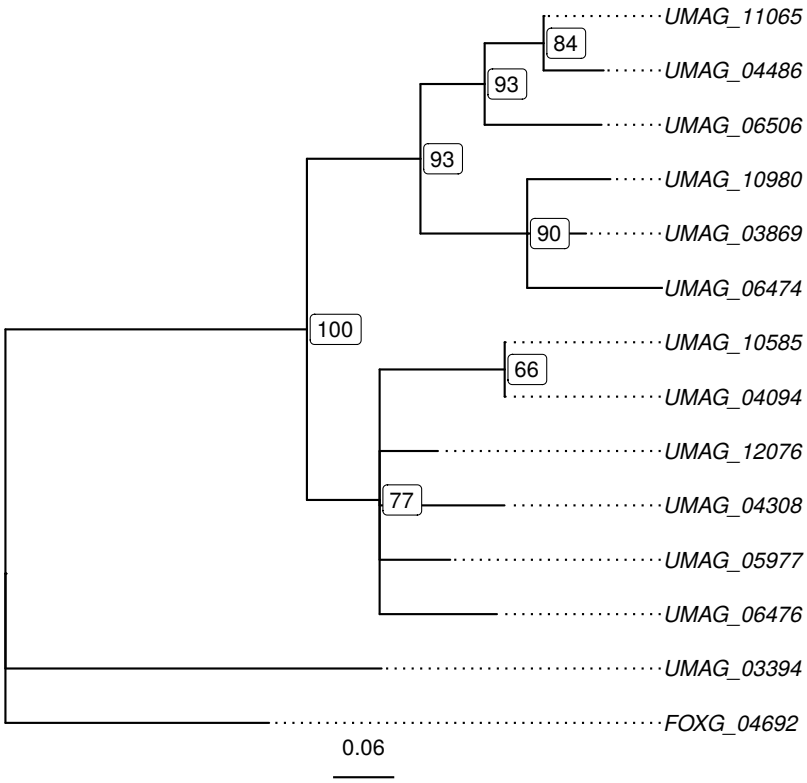
Figure 3: Alignment of *UMAG_11064* and its upstream sequence with intron 1 from the *cox1* gene of *S. reilianum*, *T. indica* and *T. walkeri*, as well as the coding sequence of the *A. bisporus* HE. Shading indicates the level of amino-acid conservation, showing conserved residues (in black) and residues with similar biochemical properties (grayscale). Amino-acids noted as 'X' have incomplete codons due to frameshifts. Highlighted exclamation marks denote inferred frameshifts and '*' characters stop codons. The location of the active site of the HE (GVY-YVG) is highlighted.

587

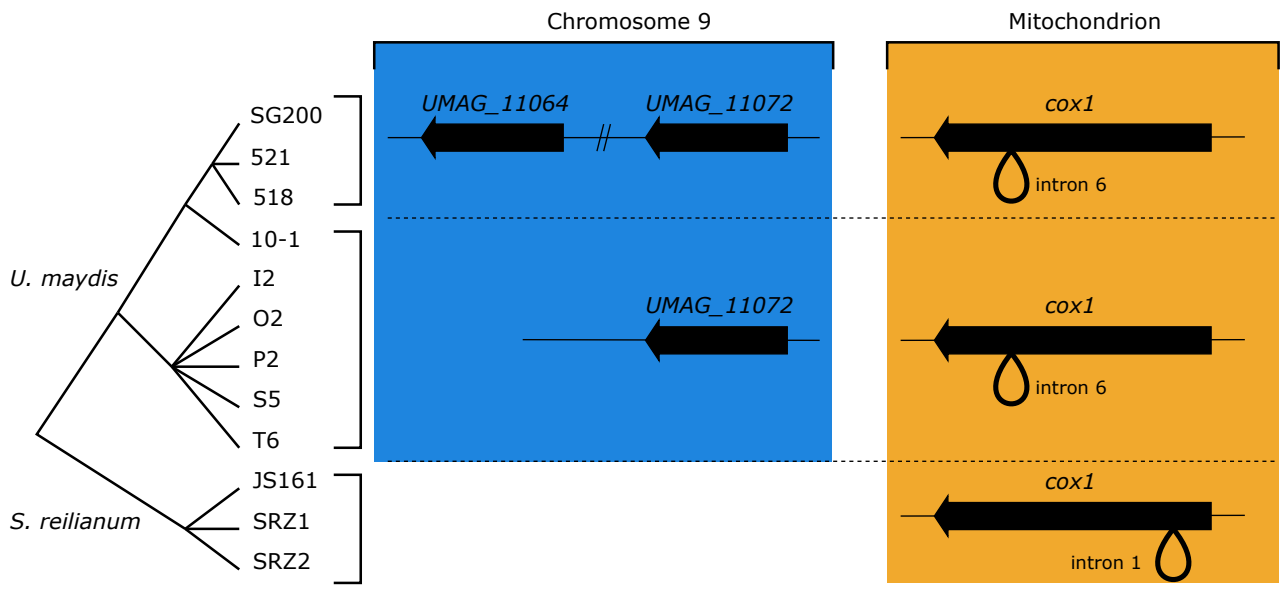


588 **Figure 4:** Intron structure of the *cox1* gene in *U. maydis* and *S. reilianum*. Annotated HEs are
589 indicated. Red boxes depict *cox1* exons, numbered from *e1* to *e9*. Introns are represented by
590 connecting lines and numbered *i1* to *i8*. Arrows within introns show LAGLIDADG (light blue) and
591 GIY-YIG HEs (pink). Dashed arrows correspond to HEGs inferred by blast search, while solid
592 arrows correspond to the annotation from the GenBank files. Piecewise sequence similarity between
593 *U. maydis* and *S. reilianum* is displayed with a color gradient.

594



595 **Figure 5:** Maximum likelihood phylogeny of *UMAG_11065 U. maydis* paralogs together with the
596 closest homolog from *F. oxysporum* (see Table 1). Nodes labels indicate support values as
597 percentage. Nodes with support values lower than 60% have been collapsed.



598 **Figure 6:** Presence of the *UMAG_11064* gene and structure of the *cox1* gene in several *U. maydis*
599 and *S. reilianum* strains, as assessed by PCR, together with their phylogeny. The *UMAG_11072*
600 gene, located 90 kb downstream the *UMAG_11064* gene on chromosome 9, was used as a positive
601 control. Strains 521 and 518 are two strains resulting from the same spore from a field isolate from
602 the U.S.A. SG200 is a genetically engineered strain derived from a cross between the 518 and 521
603 strains. Strain 10-1 is another field isolate from the U.S.A. Strains I2, O2, P2, S5 and T6 from field
604 isolates from Mexico.

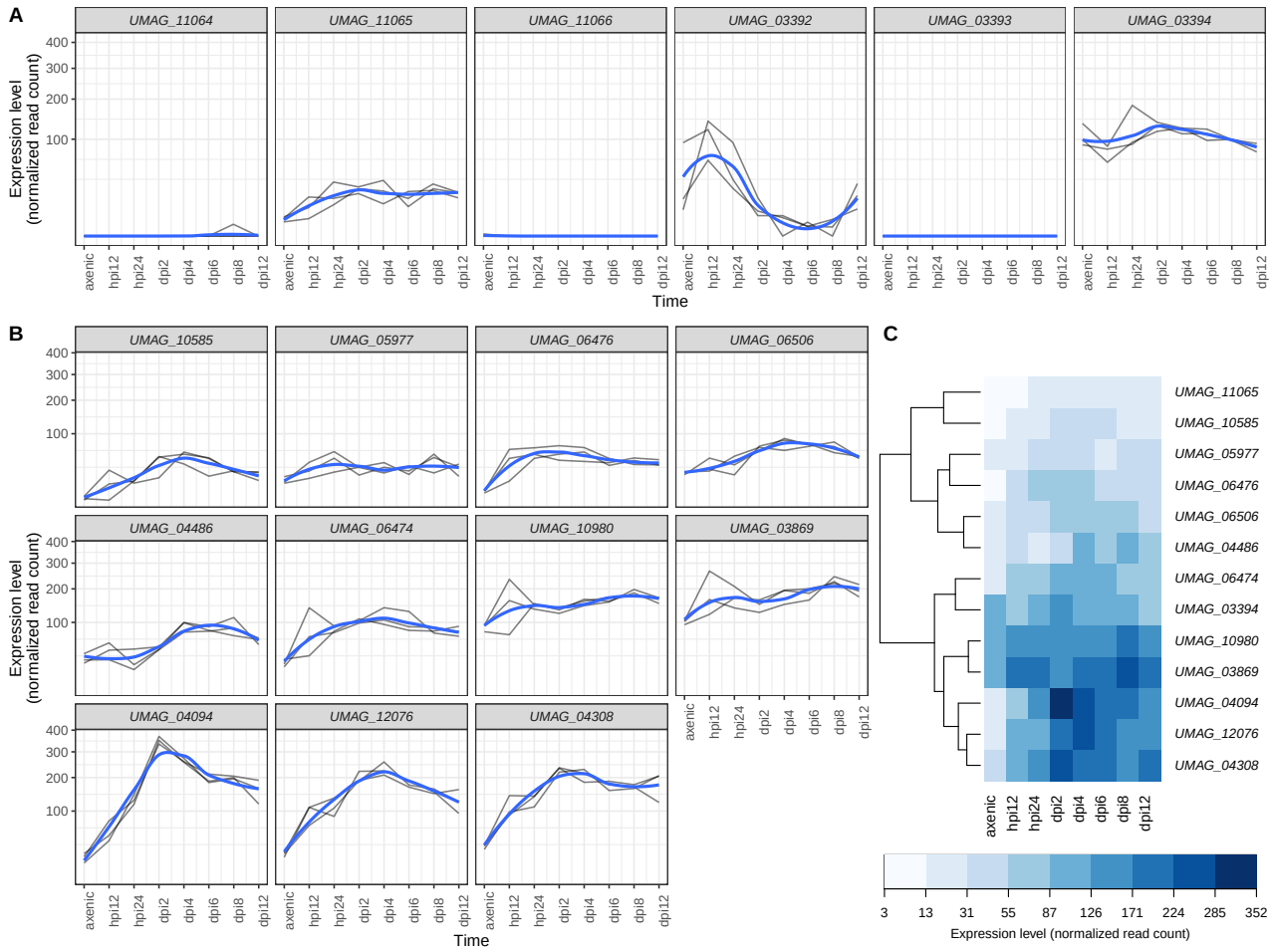


Figure 7: Patterns of gene expression for *UMAG_11064* and *UMAG_11065*, together with neighboring and homologous genes. A) Gene expression profiles for genes in the chromosome 9 telomeric region (as depicted on Figure 1B). Straight lines represent three independent replicates, while the blue curve depicts the smoothed conditional mean computed using the LOESS method. B) Gene expression profiles for the *UMAG_11065* homologs (Figure 5). Legends as in A. C) Clustering of the *UMAG_11065* homologs based on their averaged expression profile (see Methods). Hpi: hours post-infection. Dpi: days post-infection.

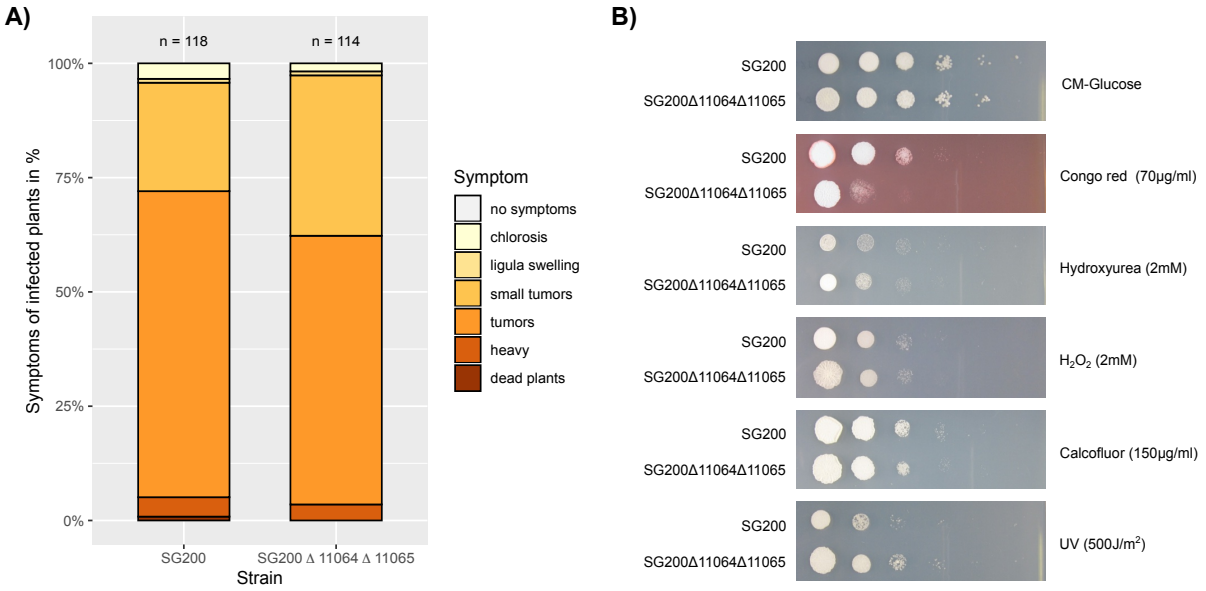
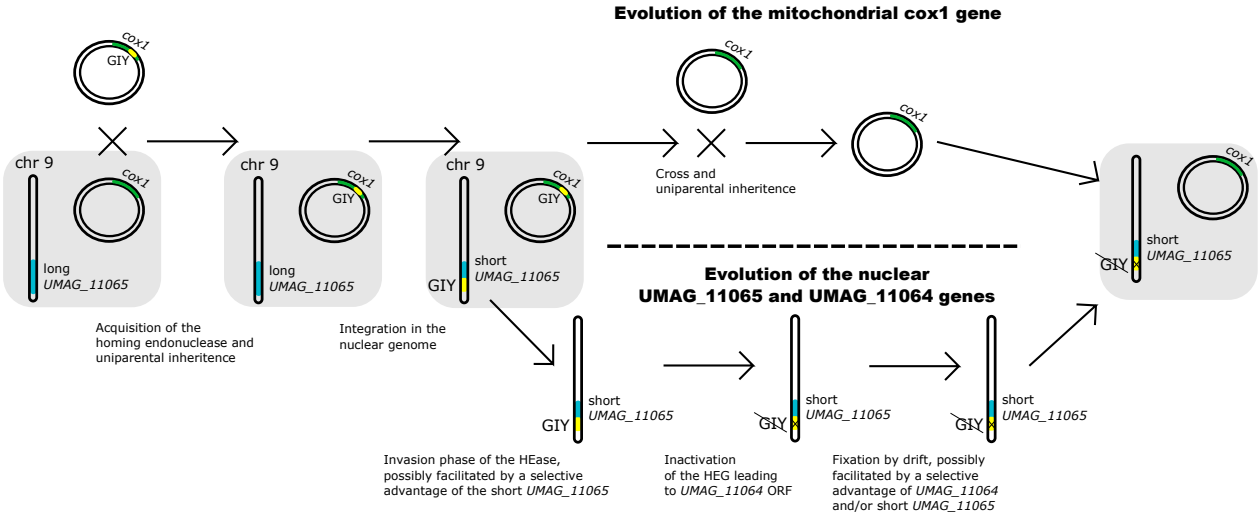


Figure 8: Phenotype assessment of the double deletion strain. A) The simultaneous deletion of *UMAG_11064* and *UMAG_11065* does not affect virulence. Maize seedlings were infected with the indicated strains. Disease symptoms were scored at 12 dpi according to Kämper et al. (Kämper et al. 2006) using the color code depicted on the right. Colors reflect the degree of severity, from brown-red (severe) to light yellow (mild). Data represent mean of n = 3 biologically independent experiments. Total numbers of infected plants are indicated above the respective columns. B) Stress assay of the double deletion strain (Δ11064Δ11065), lacking both genes *UMAG_11064* and *UMAG_11065*, compared to the parental SG200 strain. Assays were repeated at least three times with comparable results.



624 **Figure 9:** Possible evolutionary scenario recapitulating the events leading to the formation of the
625 *UMAG_11064* and *UMAG_11065 U. maydis* genes. Importantly, each step in this model may have
626 occurred by genetic drift alone. Positive selection may have favored – but is not required to explain
627 – the spread of the nuclear and/or the loss of the mitochondrial HEGs.
628

629 **Supplementary Figure S1:** Amplification of the candidate region in the telomeric region of
630 chromosome 9. A) Genomic context based on the *U. maydis* reference genome, and location PCR
631 primers. B) PCR results with corresponding expected fragment sizes. Forward (fw) and reverse (rv)
632 primer sequences are provided in Table S7.
633
634 **Supplementary Figure S2:** Amplification of A) *UMAG_11064* , B) *UMAG_11072* and *cox1* exons
635 C) 1 and D) 7 in several *U. maydis* and *S. reilianum* strains. Forward (fw) and reverse (rv) primer
636 sequences are provided in Table S7. E) Summary table of the results. Plus and minus signs indicate
637 whether the corresponding gene could be amplified or not. Numbers indicate the size of the
638 amplified region in base pairs. Strains are labelled as in Figure 6.
639
640 **Supplementary Figure S3**: Verification of the deletion of *UMAG_11064* and *UMAG_11065*. A)
641 Schematic map of the genomic region containing *UMAG_11064* and *UMAG_11065* in SG200 and
642 SG200Δ11064Δ11065. Primers used to amplify the left and right border sequences are indicated. B)
643 DNA of SG200 and SG200Δ11064Δ11065 was cleaved with Fsp1 and subjected tho southern blot
644 analysis using a mixture of Probes 1 and 2 indicated in A). The 2.94 kb fragment is diagnostic for
645 SG200 while the 4.19 kb fragment is diagnostic for the deletion of *UMAG_11064* and
646 *UMAG_11065.*
647

648 ## *Supplementary file:*

649 **Supplementary File S1**: Scripts used to conduct the phylogenetic and statistical analyses, As well
650 as R code used to generate figures 1, 3, 4, 5 and 6.