High nucleotide substitution rates associated with retrotransposon proliferation drive dynamic secretome evolution in smut pathogens

Depotter JRL[a], Ökmen B[a], Ebert MK[a], Beckers J[a], Kruse J[bc], Thines M[bc], Doehlemann G[a]*

[a]University of Cologne, CEPLAS, Institute for Plant Sciences, Cologne, Germany

[b]Senckenberg Biodiversity and Climate Research Centre (BiK-F), Frankfurt a. M., Germany.

[c]Institute of Ecology, Evolution and Diversity, Goethe University Frankfurt, Frankfurt a. M., Germany.

*Corresponding author

E-mail: g.doehlemann@uni-koeln.de (GD)

Abstract

Transposable elements (TEs) play a pivotal role in shaping diversity in eukaryotic genomes. The covered smut pathogen on barley, *Ustilago hordei*, encountered a recent genome expansion. Using long reads, we assembled genomes of 6 *U. hordei* strains and 3 sister species, to study this genome expansion. We found that larger genome sizes can mainly be attributed to a higher genome fraction of long terminal repeat retrotransposons (LTR-RTs). In the studied smut genomes, LTR-RTs fractions are the largest in *U. hordei* and are positively correlated to the mating-type locus sizes, which is up to ~560 kb in *U. hordei*. Furthermore, LTR-RTs were found to be associated with higher nucleotide substitution levels, as these higher levels occur more clustered in smut species with a recent LTR-RT proliferation. Moreover, genes in genome regions with higher nucleotide substitution levels generally reside closer to LTR-RTs than other genome regions. Genome regions with many nucleotide substitutions encountered an especially high fraction of CG substitutions, which is not observed for LTR-RT sequences. The high nucleotide substitution levels particularly accelerate the evolution of secretome genes, as their more flexible nature results that substitutions often lead to amino acid alterations.

Importance

Genomic alteration can be generated through various means, in which transposable elements (TEs) can play a pivotal role. Their mobility causes mutagenesis in itself and can disrupt the function of the sequences they insert into. Indirectly, they also impact genome evolution as their repetitive nature facilitates non-homologous recombination. Furthermore, TEs have been linked to specific epigenetic genome organizations. We report a recent TE proliferation in the genome of the barley covered smut fungus, *Ustilago hordei*. This proliferation is associated with a distinct nucleotide substitution regime that has a higher rate and a higher fraction of CG substitutions. This different regime shapes the evolution of genes in subjected genome regions. Our findings highlight that TEs may influence the error-rate of DNA polymerase in a hitherto unknown fashion.

Key words: Ustilago, transposable element, genome expansion, DNA polymerase, mating-type locus

INTRODUCTION

Transposable elements (TEs) play a pivotal role in the genome evolution of eukaryotic organisms, including fungi (1). Fungal genomes can vary considerably in size, which is often determined by the extent and recency of TE proliferations (2, 3). On one side of the spectrum, Microsporidia, a diverse group of obligate intracellular parasitic fungi, contain members with extremely small genomes down to 2.3Mb that lack TEs (4, 5). In contrast, rust plant pathogens from the order Pucciniales contain members with genome sizes that are among the largest in the fungal kingdom (6, 7). For instance, the wheat stripe rust pathogen *Puccinia striiformis* f.sp. *tritici* has an estimated genome size of 135 Mb, which more than half consists of TE sequences (8). Mutations caused by TE transposition predominantly have a neutral or negative impact, but in particular cases they can also improve fungal fitness (3, 9). For plant pathogenic fungi, TE transposition can be a source of mutagenesis to evade host immunity and/or lead to an optimized host interaction (10). TEs can also passively contribute to mutagenesis, as their transpositions increase homologous genome sequences that are prone to ectopic recombination (11, 12). Pathogens evolve by host jumps, radiation and subsequent arms races with their hosts (13), in which the latter attempts to detect pathogen ingress through the recognition of so-called invasion patterns (14). One invasion pattern that is typically detected are effectors, i.e. secreted proteins that facilitate host colonization (15). To quickly adapt to effector-triggered immunity and yet continue host symbiosis, effector genes often reside in genome regions that facilitate mutagenesis (13, 16), such as those rich in TEs (17). TE-rich genome regions may not only encounter higher mutation rates, but may also have a higher chance to fix mutations due to their functionally more accessory nature (18).

TEs are a diverse group of mobile nucleotide sequences that are categorized into two classes (1). Class I comprises retrotransposons that transpose through the reverse transcription of their messenger RNA (mRNA). Class II are DNA transposons that transpose

without mRNA intermediate. TEs are then further classified based on their sequence structure (19). Retrotransposons with direct repeats at each end of their sequence are long terminal repeat retrotransposons (LTR-RTs) (20). LTR-RTs can encode the structural and enzymatic machinery for autonomous transposition. However, they may lose this ability through mutagenesis, but still be able to transpose using proteins of other TEs (21). LTR-RTs can then be further classified into superfamilies including *Copia* and *Gypsy*, which differ in the order of their reverse transcriptase and their integrase domain (19).

Smut fungi are a diverse group of plant pathogenic, hemibiotrophic basidiomycetes of which many infect monocot plants, in particular grasses. They live saprophytically as yeasts and mate in order to switch to the diploid, filamentous stage that enables them to infect their host (22). Smut pathogens are very host-specific and generally have small genomes in comparison to other plant pathogens (23, 24). The corn smut species *Ustilago maydis* and *Sporisorium reilianum* are closely related and have genome sizes of 19.8 Mb and 18.4 Mb, respectively (25, 26). This is partly due to their low level of repetitive sequences, including TEs. In total, only 2.1 and 0.5% of the genome consists of TEs for the *U. maydis* and *S. reilianum*, respectively (27). The covered smut pathogen of barley, *Ustilago hordei*, and the Brachipodieae grass smut, *U. brachipodii-distachyi*, are two related smut fungi and have genome assemblies of 21.15 Mb and 20.5 Mb, respectively (27, 28). These larger genome assembly sizes correlate to their higher TE content, which is 11.8% and 14.3% for *U. hordei* and *U. brachipodii-distachyi*, respectively (27). The assembled genome of *U. brachipodii-distachyi* is originally published under the species name *U. bromivora* (27). *U. brachipodii-distachyi* infects members from the tribe Brachipodieae, whereas *U. bromivora* affects bromes from the supertribe Triticodae (29, 30). Considering the host specific nature of smut pathogens, we prefer to refer to this assembly as *U. brachipodii-distachyi* instead of *U. bromivora*, as the assembled strain infects *Brachypodium* species (27, 29).

Mating in grass-parasitic smut fungi is tetrapolar in *U. maydis* and *S. reilianum*, whereas *U. hordei* and *U. brachipodii-distachyi* have a bipolar mating system. In the bipolar system, there is one mating-type locus where recombination is suppressed (27, 31, 32). This locus is flanked by the *a* locus, which contains pheromone/receptor genes, and the *b* locus, which encodes the two homeodomain proteins bEast and bWest (33, 34). In the bipolar mating-type system, there are two mating-type alleles, *MAT-1* and *MAT-2*, which are in *U. hordei* ~500kb and ~430kb in size, respectively (32). A large fraction of the mating-type loci consists of repetitive sequences, i.e. ~45% repeats for *U. hordei* (28, 35). In contrast, the tetrapolar smuts, *U. maydis* and *S. reilianum*, have their *a* and *b* loci on different chromosomes causing them to segregate independently during meiosis (26, 31).

Recently, the complete genome of the reference *U. hordei* strain Uh364 was re-sequenced using the long-read PacBio technology and, instead of the previous 21.15 Mb assembly (28), a 27.1 Mb assembly was obtained (36). Thus, the *U. hordei* genome underwent a genome expansion as it is significantly larger than other sequenced smut species (27). This finding triggered us to study the *U. hordei* genome more in depth and use recently developed long-read sequencing technologies to unravel how its genome recently expanded.

Results

LTR-RTs is an important determinant for *U. hordei* genome size

To study the expansion in genome size of *U. hordei*, we sequenced and assembled 6 *U. hordei* strains of different geographic origins (Figure S1). Five contained a *MAT-1* locus and one (Uh1278) a *MAT-2*. The assemblies were composed of 23-46 contigs and ranged from 25.8-27.2 Mb in size (Table 1). Strain Uh805 was assembled into 23 contigs that are homologous to the 23 chromosomes of *U. brachipodii-distachyi* indicating that inter-chromosomal rearrangements did not occur during the divergence of these two species (27). *MAT-1* loci, regions between and including the *a* and *b* loci, ranged from 536 to 564 kb in size, whereas the *MAT-2* locus was 472 kb (Table S1). In the *U. hordei* genome assemblies, class I TE sequences are over 6 times more abundant than class II TE sequences (Table 1, S2). More than 90% of the class I TEs consist of LTR-RTs, which is a total sequence amount ranging between 4,326 and 5,272 kb (Table 1). The number of LTR-RT sequences is positively correlated to the assembly sizes (r = 0.94, p-value = 0.0051). Moreover, using strain Uh805 as a reference, 56-79% of the differences in assembly size with other strains can be attributed to differences in LTR-RT content. Thus, the variation in genome size between *U. hordei* strains can largely be attributed to intraspecific differences in LTR-RT proliferation and/or retention. More than 75% of the mating-type loci consist of repetitive sequences and over 29% are classified as LTR-RTs. The *MAT-1* and *MAT-2* loci and their flanking regions only have 27% one-to-one best homology to each other (Figure 1A), which is mainly due to mating-type specific repeats as only 6% of the repeats are shared between the two mating types. In contrast, 41 of the 47 expressed mating-type locus genes are shared between the two alleles (Figure 1B). Homologous recombination is suppressed in the mating-type region, which makes that TE transpositions within these regions are by definition mating-type specific (Figure 1B) (32).

Table 1. Genome statistics of various smut genome assemblies.

Species	*U. hordei*						*U. nuda*	*U. brachipodii-distachyi*(27)	*U. tritici*	*U. loliicola*	*U. maydis*(25)	*S. reilianum*(37, 38)
Strain	Uh359	Uh805	Uh811	Uh818	Uh1273	Uh1278	DE_29490	UB2112	Ut_3	Us_530	521	SRS1_H2-8
Assembly size (Mb)	27.0	25.8	26.2	26.2	27.2	26.6	21.4	20.4	20.4	20.8	19.7	18.5
Contigs	46	23	26	25	37	27	31	23	32	41	27	23
BUSCOs (%)	98.9	98.9	98.9	98.9	98.6	99.0	98.9	99.1	98.9	98.8	98.8	98.5
Telomers*	14	22	19	20	23	23	45	37	47	43	1	0
Total repeats (%)	38.2	35.3	36.4	36.5	38.9	36.0	22.6	17.0	16.4	8.9	4.6	3.6
Class I TEs (kb)§	5,625	4,611	4,985	4,897	5,663	4,940	1,786	672	739	102	199	5
LTR (kb)	5,272	4,326	4,615	4,549	5,208	4,607	1,537	463	462	9	185	5
Gypsy (kb)	2,066	1,688	1,679	1,653	2,225	1,873	484	144	127	3	4	3
Copia (kb)	2,732	2,331	2,561	2,554	2,587	2,531	1,1019	292	289	6	182	1
Class II TEs (kb)§	781	746	708	791	731	692	395	473	285	482	5	103

§ Only repetitive sequences that were larger than 500 bp were classified.

* "TAACCC" or "GGGTTA" repeats at the end of a contig

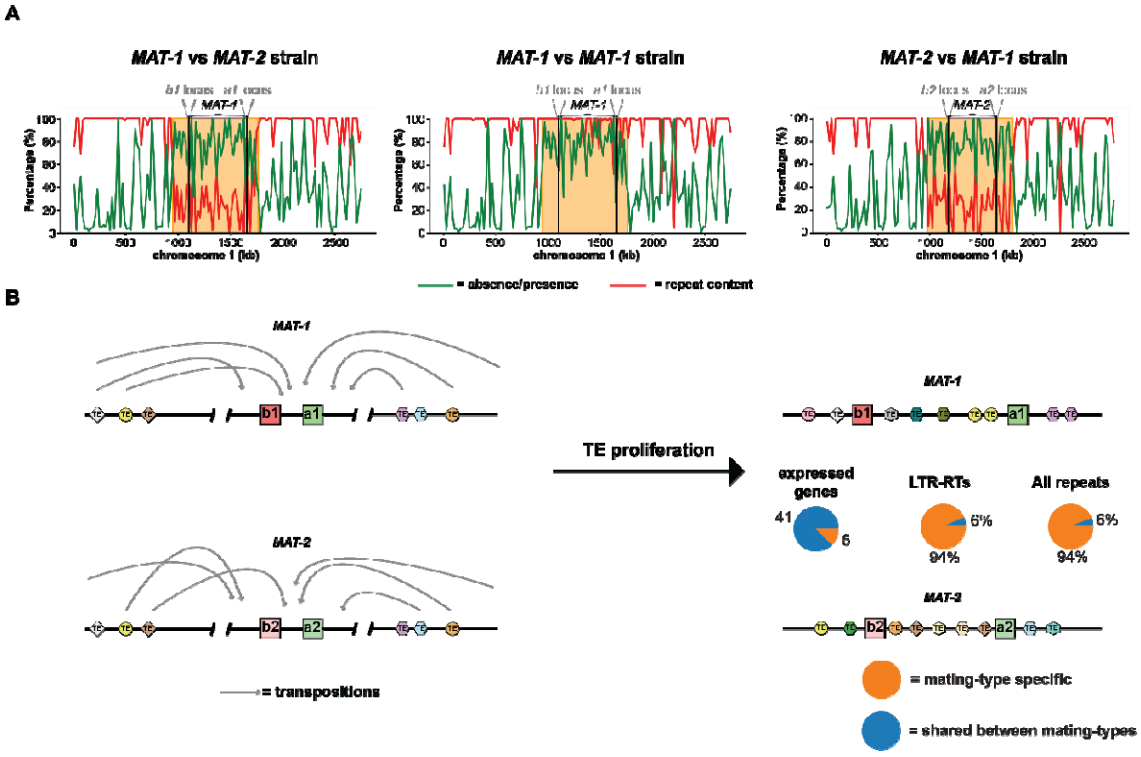
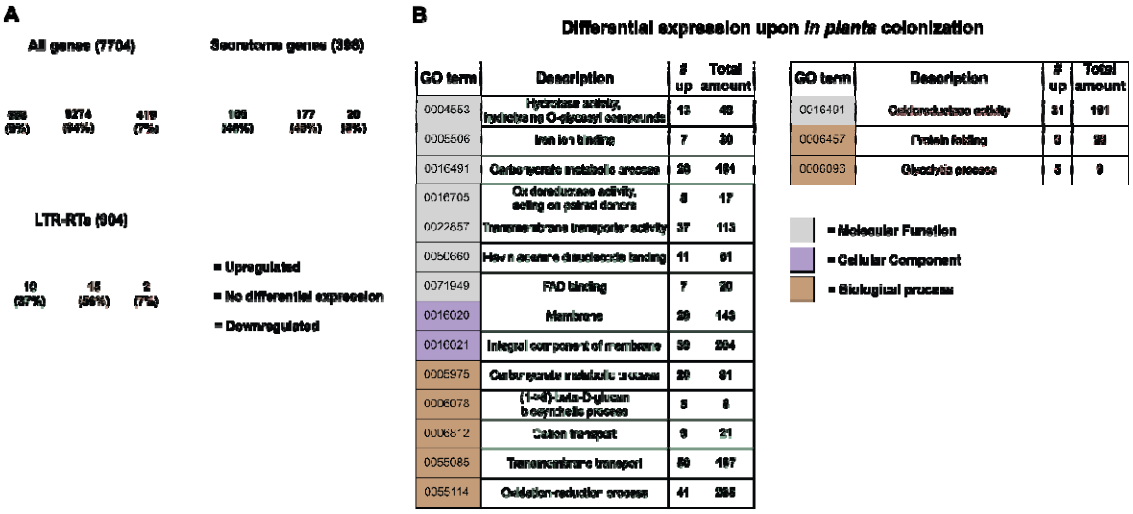


Figure 1: The mating-type specificity of *MAT-1* and *MAT-2* loci sequences. (A) As references, the *MAT-1* strain Uh805 and the *MAT-2* strain Uh1278 were used. Repeat content and presence/absence polymorphisms were calculated for 20 kb windows. Presence-absence polymorphisms were determined between the *MAT-1* and *MAT-2* reference strains, in addition to the *MAT-1* strains Uh805 and Uh811. The orange squares encompass the mating-type loci and indicate genome regions where the repeat contents are high and sequences are generally mating-type specific. (**B**) Model that explains the mating-type specificity of sequences within and flanking the mating-type loci. The absence of recombination within the mating-type loci and their flanking regions makes that transpositions within these regions become mating-type specific.

The U. hordei secretome is activated upon plant colonization, whereas LTR-RTs are generally inactive

To study gene and LTR-RT expression, RNA-seq was done from from *U. hordei* grown in liquid medium, and in barley leaves at 3 days post infection. In total, 6229 of the 7704 (81%) predicted gene loci in Uh803 were expressed in either of the two *U. hordei* growth conditions, whereas only 27 of the 904 (3%) LTR-RTs displayed expression (Figure 2B).

Moreover, only 7 of the expressed LTR-RTs displayed expression in more than half of their sequence. Of these 7, there was one *Copia* and one *Gypsy* LTR-RT that can be autonomous, as functional domains for aspartyl protease, reverse transcriptase and integrase could be identified. In conclusion, almost all LTR-RT sequences were inactive in the two tested environmental conditions. For the genes, 558 (9%) were upregulated *in planta*, whereas 419 (7%) were downregulated (Figure 2). Up- and downregulated genes were screened for Gene Ontology (GO) term enrichments, to see which biological processes are affected by plant colonization. In total, 14 and 3 GO terms were enriched in *in planta* up- and down-regulated genes, respectively (Figure 2). Generally, processes associated with the fungal membrane, including transmembrane transport were upregulated *in planta*. In correspondence with these results, 165 of the 558 genes upregulated *in planta* encode secreted proteins. Thus, 45% (165/369) of the expressed genes that encode secreted proteins were upregulated *in planta*, which is a significant enrichment (Fisher exact test, *p*-value = 1.87e-83) (Figure 2). In contrast, only 6% of the genes encoding a secreted protein were downregulated. Of these downregulated genes, 35% (7/20) was predicted to have a carbohydrate-active (CAZyme) function, whereas this was 18% (29/165) for *in planta* upregulated secretome genes. Thus, the *U. hordei* transmembrane transport system and secretome genes are strongly activated upon plant colonization, whereas hardly any LTR-RTs display expression. In total, 24% (median) of the 20kb flanking regions secretome genes consist of repeats, which is the same for non-secretome genes (t-test, *p*-value = 0.51) (Figure S2). Secretome genes upregulated *in planta* have a median of 21%, which is not significantly lower than non-secretome genes (t-test, *p*-value > 0.01). Thus, in contrast to some other filamentous plant pathogens (17), secretome genes are not especially associated with repeat-rich genome regions in *U. hordei*.



177

178 **Figure 2. Differential expression of *U. hordei* loci upon plant colonization. (A)** Comparison of *U. hordei*
179 locus expression between growth in liquid culture medium and *in planta*. The numbers between brackets
180 indicate how many genes and LTR-RTs have been annotated in the genome. The significance of differential
181 expression was calculated using a threshold of log2-fold-change. **(B)** Gene Ontology (GO) term enrichments in
182 differently regulated *Ustilago hordei* genes. In green and red are GO terms that are enriched in *in plant* up- and
183 down-regulated genes, respectively. *p*-values were calculated with the Fisher's exact test. For the whole figure
184 significance was determined with a *p*-value < 0.01 and corrected for multiple-testing with the Benjamini-
185 Hochberg method.

186

187 **Higher LTR-RT contents in genomes of smuts with a bipolar mating-type system**

188 As LTR-RTs played a predominant role in the genome expansion of *U. hordei*, we also
189 studied the impact of TE dynamics on the genome evolution of *U. hordei* sister species. We
190 sequenced genomes of *Ustilago nuda*, *Ustilago tritici* and *Ustilago loliicola*, which are smut
191 species that are close relatives of *U. hordei* and *U. brachipodii-distachyi* (29, 39). Assemblies
192 of 21.4, 20.8 and 20.4 Mb were obtained in 31, 41 and 32 contigs for *U. nuda*, *U. loliicola*
193 and *U. tritici*, respectively (Table 1). A phylogenetic tree was constructed, which included the
194 newly sequenced species as well as *U. brachipodii-distachyi*, *U. maydis* and *S. reilianum*
195 (Figure 3) (25, 27, 37, 38). *U. hordei*, *U. nuda*, *U. brachipodii-distachyi* and *U. tritici* cluster
196 together with *U. loliicola* being the closest outgroup species. Within the cluster, *U. hordei*

197 diverged the most recently from *U. nuda*, which also infects *Hordeum* species (Figure 3).
198 Synteny between the different contigs was also investigated and the ancestral gene order
199 reconstructed. *S. reilianum* and *U. loliicola* do not have inter-chromosomal rearrangement in
200 comparison to their reconstructed last common ancestor (Figure 3). The *U. maydis* genome
201 has one inter-chromosomal rearrangement with respect to its last common ancestor with *S.*
202 *reilianum*. *U. hordei*, *U. nuda*, *U. brachipodii-distachyi* and *U. tritici* share one inter-
203 chromosomal rearrangement that occurred after their divergence from *U. loliicola* (Figure 3).
204 As previously reported, this rearrangement resulted in the mating-type polarity switch from
205 tetrapolar to bipolar due to the linkage of the *a* and *b* mating-type loci (31). This inter-
206 chromosomal rearrangement is the only one observed in the assemblies of *U. brachipodii-*
207 *distachyi*, *U. nuda* and *U. hordei*, whereas the *U. tritici* assembly has one additional inter-
208 chromosomal rearrangement. The smut species with a bipolar mating-type generally have a
209 higher repeat content (16.4-38.9%), than the tetrapolar ones (3.6-8.9%) (Table 1). This
210 increase in repeat content can largely be attributed to LTR-RT sequences, which comprise
211 4,326 kb in *U. hordei* (Uh805) in contrast to only 5 kb for *S. reilianum* (Table 1). Thus,
212 repeats have increased after the polarity switch, mainly due to higher LTR-RT contents.
213 Furthermore, repeat and the LTR-RT contents of smut genomes with a bipolar mating type
214 positively correlate to mating-type loci sizes (r = 0.98, p-value = 0.02, using strain Uh805 for
215 *U. hordei*), which ranges from 190 kb for *U. brachipodii-distachyi* to 560 kb for *U. hordei*
216 (Figure 3, Table S1). In conclusion, the proliferation and/or retention of TEs seems to be an
217 important determinant of the eventual size of mating-type loci.

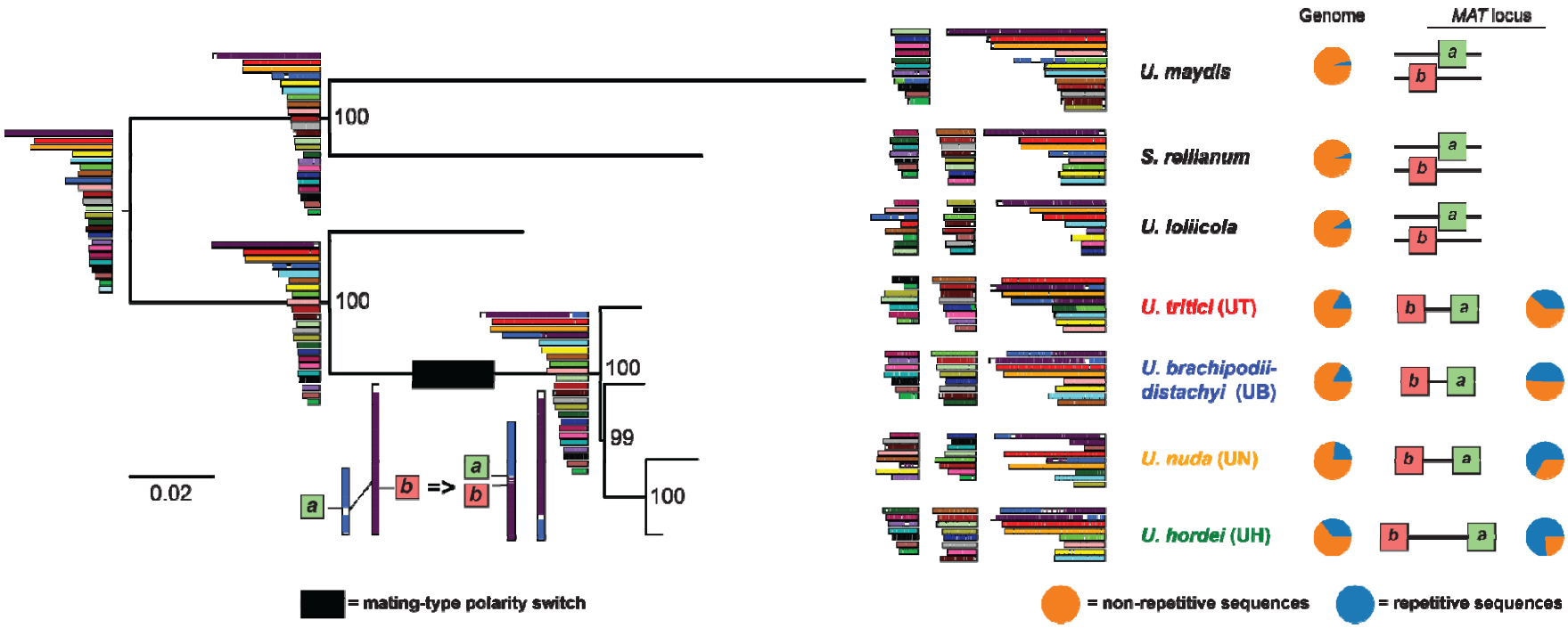


Figure 3: Genome evolution of smut species with bi-and tetrapolar mating systems. (**A**) Phylogenetic relationship between smut pathogens based on Benchmarking Universal Single-Copy Orthologs (BUSCOs). Phylogenetic relationship between newly and previously sequences smut species was constructed with the *Ustilago maydis*/*Sporisorium reilianum* branch as an outgroup. In total, 1,667 BUSCOs were used for tree construction. For *U. hordei*, strain Uh805 was used in the tree. The robustness of the trees was assessed using 100 bootstrap replicates. The colours of the contigs indicate the synteny with the ancestor contigs. The blue section of the circles indicate the repeat fraction that is present in the genome assembles and within the mating-type loci for species with a bipolar mating-type system.

The timepoint of most recent LTR-RT proliferation differs between smut species

Although species with a bipolar mating system collectively encountered an increase in LTR-RT content, there are large interspecific differences as *U. hordei* has more than 9 times the amount of LTR-RT sequences than *U. tritici* (Table 1). To study the relative timepoint of the most recent LTR-RT proliferation, the nucleotide sequence identity distributions of the best reciprocal paralogous and orthologous LTR-RT sequences were calculated (Figure 4A). This was on the one hand done for the species with the highest LTR-RT contents, *U. hordei* and *U. nuda*, and on the other hand for *U. brachipodii-distachyi* and *U. tritici*. The distribution of the paralogous LTR-RTs in *U. brachipodii-distachyi* and *U. tritici* displayed two maxima, i.e. at 82-83% and at 88-90% (Figure 4A). The maximum of the orthologous LTR-RTs between *U. brachipodii-distachyi* and *U. tritici* was at 93%. Thus, orthologous LTR-RTs generally have a higher identity than paralogous ones, which indicates that LTR-RTs mainly proliferated before the last common ancestor of *U. brachipodii-distachyi* and *U. tritici* (Figure 4A). Orthologous LTR-RTs between *U. hordei* and *U. nuda* displayed a maximum at 89%, whereas for paralogous LTR-RTs a maximum at 93% was present for *U. nuda* and two maxima at 94 and 97% for *U. hordei* (Figure 4A). Thus, in contrast to *U. brachipodii-distachyi* and *U. tritici*, paralogous LTR-RTs generally have a higher identity than orthologous ones, which means that LTR-RTs continued to proliferated after the last common ancestor of *U. nuda* and *U. hordei*.

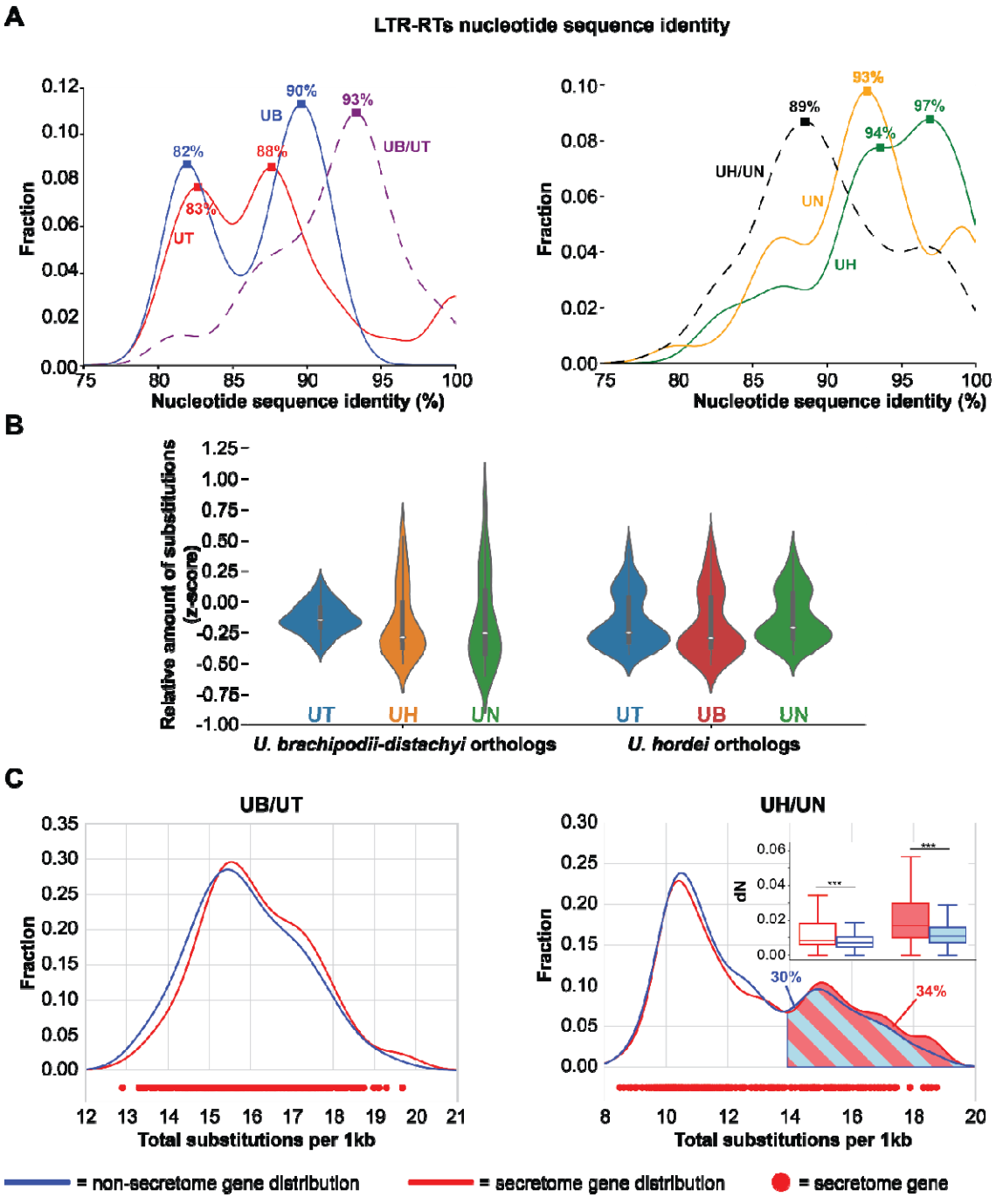


Figure 4: Interspecific comparison in long terminal repeat retrotransposon (LTR-RT) proliferation and local gene nucleotide substitution levels (**A**) Nucleotide sequence identity distribution of best reciprocal paralogous (full lines) and orthologous (striped lines) LTR-RT sequences. Squares on the lines display maxima with the corresponding sequence identity value. (**B**) The normalized sequence identity (z-score) was calculated for *U. brachipodii-distachyi* and *U. hordei* genes with orthologs of other bipolar mating-type species. The sequence identity was determined for non-overlapping sliding windows of 75 genes. (**C**) The distribution of the

sequence divergence between *U. brachipodii-distachyi*/*U. tritici* and *U. hordei*/*U. nuda* ortholog windows (75 genes) are depicted for secretome and non-secretome genes in red and blue, respectively. The *U. hordei*/*U. nuda* distribution displays two peaks. The number of nonsynonymous substitutions per nonsynonymous site (dN) was compared between secretome and non-secretome genes for genes in the first and second distribution peak. Significance was determined with an unequal variance t-test. ***: p-value < 0.001.

High nucleotide substitution levels affect secretome proteins

As TE-active genome regions have been associated with distinct nucleotide substitution regimes (11, 40, 41), we studied if different extents of LTR-RT fractions are associated with different nucleotide substitution regimes. We calculated the median number of substitutions between orthologs in windows of 75 genes. To ensure that genes are transcriptionally active, we only analysed *U. hordei* genes that displayed expression from here onwards. The variation in the normalized number of nucleotide substitutions (z-score) between *U. brachipodii-distachyi* and *U. tritici* ortholog windows is around 5.3 and 8.8 times less than *U. brachipodii-distachyi* ortholog windows with *U. hordei* and *U. nuda*, respectively (Figure 4B). In contrast, nucleotide substitutions of *U. hordei* ortholog windows with the other bipolar mating-type species display a more constant variation as the most varying ortholog windows (with *U. brachipodii-distachyi*) have only a 0.5 times higher variation than the least varying (with *U. nuda*) (Figure 4B). Thus, since their last common ancestor, gene nucleotide sequence divergence occurred more evenly across the genomes of *U. brachipodii-distachyi* and *U. tritici* than in *U. hordei* and *U. nuda*, where the divergence is more clustered. Correspondingly, substitutions between *U. brachipodii-distachyi* and *U. tritici* ortholog windows have a unimodal distribution, whereas the distribution between *U. hordei* and *U. nuda* have two distinct peaks (Figure 4C). For both comparisons, the distributions of secretome genes generally corresponds to that of non-secretome genes (Figure 4C). For *U. hordei*/*U. nuda* ortholog windows, the second peak in the distribution contains 30% of the

non-secretome and 34% of the secretome genes, which is not significantly different (Fisher exact test, *p*-value = 0.10). Thus, high nucleotide substitution levels are not especially associated with secretome genes. Also for the second distribution peak, no GO terms enrichments could be found (*p*-value < 0.01). Furthermore, nucleotide substitution levels are negligibly positively correlated (Pearson's r = 0.14 and *p*-value = 0.0026) with the fraction of species-specific genes (*U. hordei* genes without *U. maydis* ortholog) (Figure S3). In conclusion, genes in genome regions with high nucleotide substitution levels could not be associated with a particular function or more clear accessory nature. However, higher nucleotide substitution levels have a different impact on genes depending on their function. Substitutions that lead to amino acid alterations are more frequently fixed in secretome genes than in non-secretome genes (Figure 4C). The median number of nonsynonymous substitutions per nonsynonymous site (dN) for secretome genes is 18% higher than for non-secretome genes in the first peak of the *U. hordei* and *U. nuda* secretome distribution, whereas this is 55% for the second peak. Thus, the more flexible nature of secretome proteins makes that a higher nucleotide substitution rate speeds up their evolution.

High substitution levels are association with high fractions of CG substitutions

We then analysed which type of substitutions (AC, AG, AT, CG, CT, GT) occur across the different substitution levels. The number of all substitution types are positively correlated with the number of total substitutions. Transitions (AG and CT substitutions) are responsible for 56% of the different substitution levels between *U. brachipodii-distachyi*/*U. tritici* ortholog windows (Figure 5A). In total, 27% of the variance can be attributed to CG substitutions, whereas the other transversions ranged from 4 to 7%. Similarly, for *U. hordei*/*U. nuda* ortholog windows, the number of all substitutions types display a positive correlation with the number of total substitutions. Here, CG substitutions are responsible for

47% of the variation in nucleotide substitution levels, whereas the contributions of other substitution types range from 5% (GT) to 16% (CT) (Figure 5A). The fraction of CG substitutions varies from 4% to 27% across the ortholog windows, whereas this is 3% to 16% for *U. brachipodii-distachyi*/*U. tritici* ortholog windows (Figure S4A). Correspondingly, the number of all substitution types in intergenetic regions are positively correlated with the total number of gene substitutions (Figure 5B). Similar to the coding regions, transitions contributed 52% to the intergenic substitution variation, whereas this was 23% for CG substitutions and 8-9% for the other transversion in *U. brachipodii-distachyi*/*U. tritici* ortholog windows. In contrast, *U. hordei*/*U. nuda* ortholog windows, transitions only contributed 40% to the variation of intergenic substitution levels (Figure 5B, S4B). All substitution types considered, CG displayed the highest variation and was responsible for 24% of the total nucleotide substitution variation. Although CG has, with 24%, the highest variation, this contrast with the 47% of coding regions. This discrepancy may be due to the difference in selection regime between coding and non-coding genome regions. The dN for every individual substitution type is positively correlated with the total substitution level. The correlation slope is the highest for CG substitutions, which is 3.5 times higher than for the second highest slope (AT). Similarly, the number of synonymous substitutions per synonymous site (dS) also has the steepest correlation slope for CG. However, this slope is only 1.5 times greater than the second highest slope (CT). In conclusion, *U. hordei* and *U. nuda* encountered more variation in their local nucleotide regimes than *U. brachipodii-distachyi* and *U. tritici*. For *U. hordei* and *U. nuda*, genome regions with higher nucleotide substitution levels encountered a relatively higher fraction of CG substitutions, which, after selection, is especially apparent in coding regions. Conceivably, different contributions of substitutions types impact codon frequencies and consequently amino acid compositions of proteins. Encoded proteins of genes that reside in genome regions with higher substitution

levels are Cys, Gln, His, Leu richer and Asp, Gly, Phe, Val poorer than regions with lower substitution levels (Figure S5). Moreover, these specific amino acid tendencies have become more aggravated since the *U. hordei* divergence from *U. brachipodii-distachyi* (Figure S5).

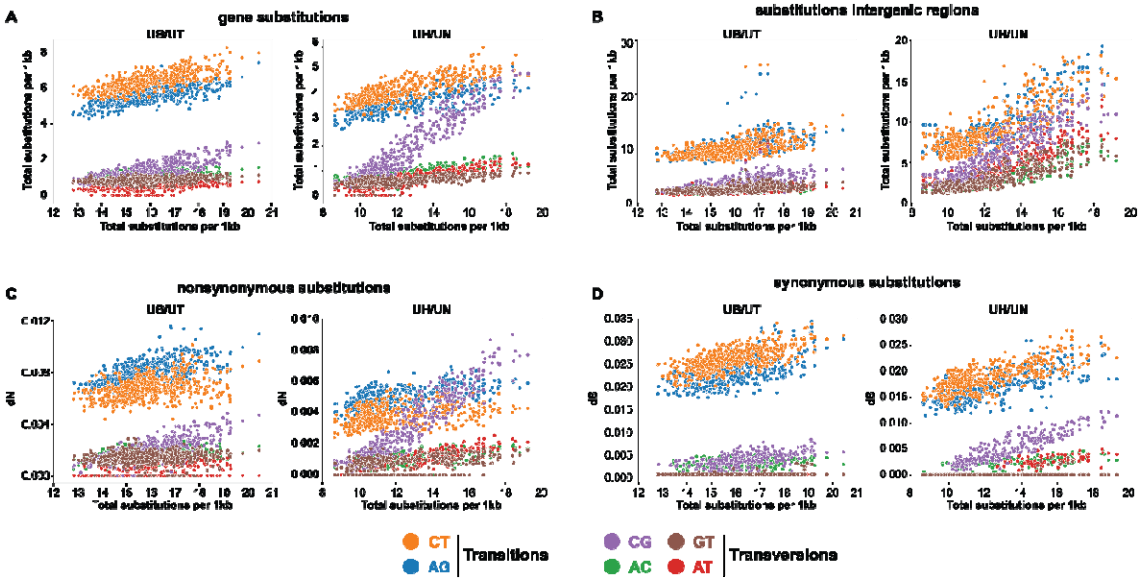


Figure 5: Comparison of nucleotide substitution regimes for *U. brachipodii-distachyi*/*U. tritici* and *U. hordei*/*U. nuda* ortholog windows. The nucleotide substitutions were calculated for windows of 75 genes with a sliding step of 10. The x-axis consistently displays the total substitutions per 1 kb for these windows. (**A**) The y-axis depicts the median number of every substation type (CT, AG, CG, AC, GT, AT) of ortholog windows. (**B**) The y-axis depicts the median number of every substation type for the intergenetic regions of ortholog windows. (**C**) The y-axis depicts the median fraction of nonsynonymous substitutions per nonsynonymous site (dN) for every substitution type in ortholog windows. (**D**) The y-axis depicts the median fraction of synonymous substitutions per synonymous site (dS) for every substitution type in ortholog windows.

High local nucleotide substitution levels are associated with LTR-RT proliferation

As higher nucleotide substitution levels with distinct substitution patterns occur in *U. hordei* and *U. nuda*, which are species with more recent LTR-RT proliferations than *U. brachipodii-distachyi* and *U. tritici*, we looked for a direct association with LTR-RTs. The median distance of *U. hordei* genes to their closest LTR-RT is significantly, negatively correlated to the median substitution level (with *U. nuda* orthologs) of ortholog windows (Pearson's r = -0.27, *p*-value = 4.74e-9) (Figure 6A). A correlation coefficient of -0.27 points towards a weak correlation. In conclusion, genes in genome regions with higher nucleotide substitution levels generally reside closer to LTR-RTs.

To study the LTR-RT nucleotide substitution regime, we constructed ancestor LTR-RT sequences of LTR-RT families, using the convention that TE family members share at least 80% sequence identity in at least 80% of their sequence with one other family member (19). To facilitate the sequence alignment and ancestor sequence construction, we only took a subset of the LTR-RTs and excluded the terminal repetitive sequences (more details in Material and Methods). In total, ancestors of 13 LTR-RT families were reconstructed using 252 LTR-RT sequences (Figure 6B). We then determined clades in the phylogenetic tree that solely consists of very similar *U. hordei* LTR-RTs and, thus, recently proliferated in *U. hordei*. In relation to their ancestor sequences, LTR-RTs substitutions consisted 91% (median) of transitions (Figure S6). In contrast, nucleotide substitutions of their 20kb flanking regions (excluding repetitive sequences) consisted 62% of transitions (compared to *U. nuda*). Here, CG comprised the highest fraction of transversions with a median of 14% of the total substitutions (Figure 6). In contrast, only 1% of the substitutions between LTR-RTs and their ancestors were CG. In conclusion, LTR-RTs are not subjected to the nucleotide substitution regime with a high fraction of CG substitutions.

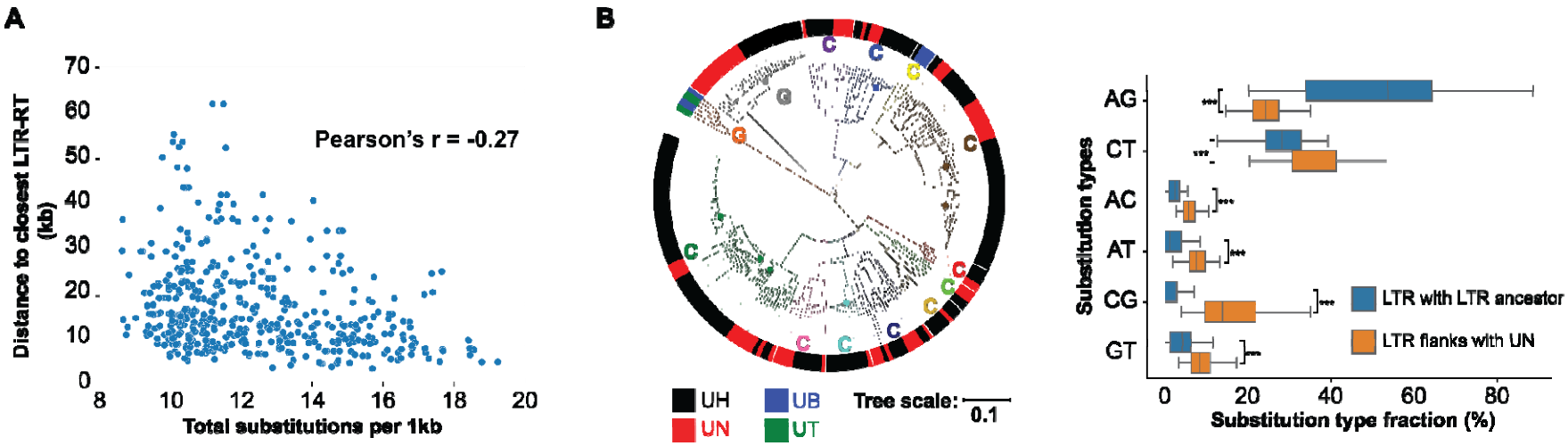


Figure 6: High local nucleotide substitution levels are associated with long terminal repeat retrotransposons (LTR-RTs). (**A**) The relation between the median number of nucleotide substitutions and the median distance between *U. hordei* genes and the closest LTR-RT for ortholog windows of 75 genes with a sliding step of 10. (**B**) In total, 252 LTR-RTs are included in the phylogenetic tree and their species of origin is indicated by the outer band colour (UH = *U. hordei*, UN = *U. nuda*, UB = *U. brachipodii-distachyi*, UT = *U. tritici*). LTR-RTs families are indicated by the circle segments in different colour. Families indicated with "G" are gypsy-type families and "C" are copia-type families. For recently proliferated *U. hordei* LTR-RTs, the fractions of the different substitution types were determined with their LTR-RT ancestor that is indicated with a circle on the phylogenetic tree. Substitution fractions of the 20 kb flanking regions (40 kb in total) of the LTR-RT with *U. nuda*, excluding repetitive sequences, were also determined. Significant differences between LTR and flanking regions were determined for every substitution type individually with an unequal variance t-test. ***: *p*-value < 0.001.

Discussion

Nucleotide substitution rates are unevenly distributed across genomes and can be influenced by numerous factors, including neighbouring nucleotides, recombination frequencies and TE activity (41–43). Nucleotide divergence in *U. hordei* and *U. nuda* occurred more clustered in their genomes compared to *U. brachipodii-distachyi* and *U. tritici* (Figure 4B-C). These differences in regional substitution rates can be directly or indirectly caused by distinct LTR-RT dynamics, as *U. hordei* and *U. nuda* encountered a more recent LTR-RT proliferation than *U. brachipodii-distachyi* and *U. tritici* (Figure 4A, Table 1). Moreover, another association between LTR-RTs and nucleotide substitution rates was found, as gene nucleotide substitution levels are weakly, negatively correlated to the distance of the closest LTR-RT in *U. hordei* (Figure 6A). Conceivably, the purge of LTR-RTs from the genome impacts this correlation considerably, as purged LTR-RTs cannot be detected, but may have had an impact on the local nucleotide substitution regime. High nucleotide substitution levels are accompanied with a high fraction of CG substitutions (Figure 5,S4). A relatively high fraction of CG substitutions is found in the flanking regions of recently proliferated LTR-RTs, but not for LTR-RTs themselves (Figure 6B). A mechanism to how LTR-RTs may impact local nucleotide substitution regimes remains elusive. The relation might be indirectly, and caused by different epigenetic regimes in the genome (44). Distinct methylation and/or histon modification patterns may occur in LTR-RT-rich genome regions, which leads to a more erroneous DNA polymerase with high CG substitutions. However, LTR-RTs themselves are not subjected to a high fraction of CG substitutions. Possibly, DNA methylation may specifically target LTR-RT sequences, which cause a distinct nucleotide substitution regime that is different from the LTR-RT flanking regions. Alternatively, the distinct nucleotide substitution regime may not have an epigenetic origin and originates from a more erroneous DNA polymerisation of the single stranded LTR-RT flanking regions

during LTR-RT insertion. This mechanism has been previously suggested in rice, where higher nucleotide substitutions levels occur close to TE insertion sites (41). TE insertion causes cuts in the host DNA, which are then ligated by the host (45, 46). However, the cut host DNA might become a target for 3'->5' exonuclease resulting in a segment of single-stranded DNA (41). The complementary strand of this stretch of DNA would then be synthesized by a replication complex with lower DNA polymerase fidelity and mismatch repair. This hypothesis could explain why the nucleotide substitution regime with high CG fractions affect LTR-RT neighbouring regions, but not LTR-RTs themselves.

Higher levels of nucleotide substitutions impact the evolution of the genes that reside in the affected genome regions. Particularly the occurrence of nonsynonymous CG substitutions strongly increases with higher substitution levels (Figure 5C). These shifts in nucleotide substitution regime change the amino acid composition of proteins (Figure S5). High nucleotide substitution levels especially lead to amino acid alteration in secretome genes, as they are more flexible to amino acid changes than other genes (Figure 4C). Although the effect of nucleotide substitutions affected secretome genes more, enrichments of particular gene functions could not be found for genome regions with high nucleotide substitution levels. Thus, the high substitution levels are not in line with the two-speed genome model (18, 24), as they do not specifically affect genome regions that are rich in secretome genes, which include effector gene candidates. More generally, repeat content were not more frequently found in the proximity of secretome genes compared to other genes (Figure S2). The specificity and the universality of the two-speed genome model for filamentous plant pathogens has recently been contested (47, 48). More plant pathogens have been reported where effector candidates do not especially reside in gene-poor/repeat-rich regions, such as the leaf spot pathogen *Ramularia collo-cygni* on barley, the earlier

421 mentioned *P. striiformis* f. sp. *tritici* and the barley powdery mildew pathogen *Blumeria*
422 *gramin*is f. sp. *hordei* (49–51).

423 LTR-RTs are mainly responsible for the *U. hordei* genome expansion (Table 1). The
424 expansion occurred especially in the mating-type locus that increased almost three times in
425 size in comparison to *U. brachipodii-distachyi* (Table S1). The lack of the purifying
426 recombination ability in this genome region can be the reason why LTR-RTs especially
427 accumulated in the mating-type and flanking genome regions (28, 32). Conceivably, this
428 process is reinforced by the increasing presence of repetitive sequences as the transposition
429 into a repeat-rich genome region is less likely to have a severe fitness cost than a
430 transposition into repeat-poor regions. Furthermore, the co-occurrence of high LTR-RTs
431 genome contents and the switch in mating-type organization from tetra- to bipolar may
432 indicate that mating-type polarity impacts LTR-RT proliferation and/or retention (52). In case
433 of biallelic *a* and *b* loci, the switch from a tetra-to bipolarity results in a basidiospore
434 compatibility change from 25 to 50%. Consequently, it takes a tetrapolar smut on average
435 longer to find a mating type than a bipolar smut. This longer time might increase the
436 opportunity to mate with spores from a different offspring and, thus, increase outcrossing.
437 The higher outcrossing rate for tetra- compared to bipolar smuts is even more pronounced
438 when multiallelism exists for the *a* and *b* loci (53). Multiallelism increases the compatibility
439 on population level, whereas compatibility within the same offspring remains 25 and 50% for
440 tetra- and bipolar smuts, respectively. Lower levels of outcrossing reduce the purifying
441 recombination ability of smuts, which may be the reason why LTRs could be retained for
442 longer and proliferate to a further extent in bipolar smuts (28, 52).

443 TEs are import drivers of genome evolution as they directly cause mutagenesis
444 through their transpositions and indirectly increase the change of non-homologous
445 recombination due to their repetitive nature (12). LTR-RT proliferation in *U. hordei* indicates

446 that TE activity may also influence local nucleotide substitution regimes and increase the
447 substitution levels in the genome regions where they insert. Consequently, genes in the
448 proximity of these insertion sites encounter more nonsynonymous substitutions and thus
449 evolve faster (Figure 5C). Fast gene evolution may be advantageous under stressful
450 condition, when TEs are typically more active or change their activity (54, 55).

Material and Methods

Genome sequencing and assembly

Genomic DNA from all smut species was isolated using a MasterPure™ Complete DNA&RNA Purification Kit (Epicentre®, Illumina®, Madison, Wisconsin, USA) according to the manufacturer's instructions. Long *U. hordei* reads were obtained with the Oxford Nanopore MinION device. The genomes of six *U. hordei* strains were sequenced: Uh359, Uh805, Uh811, Uh818, Uh121 and Uh122 (10). The library was prepared according the Oxford Nanopore Technology (ONT) protocol for native barcoding genomic DNA (EXP-NBD104 and SQK-LSK109). Three *U. hordei* strains were multiplexed for every run. The prepared library was loaded on an R9.4.1 Flow Cell. ONT reads were base-called, filtered (default value) and barcodes were trimmed with the Guppy Basecalling Software v3.5.1 of ONT. Paired-end *U. hordei* 150 bp reads were obtained with the Illumina HiSeq 4000 device. Library preparation (500bp insert size) and sequencing were performed by the BGI Group (Beijing, China). Paired-end *U. hordei* reads were filtered using Trimmomatic v0.39 with the settings "LEADING:3 TRAILING:3 SLIDINGWINDOW:4:15 MINLEN:100", only reads that remained paired after filtering were used in the assembly (56). In total, 3.2-4.5 Gb of filtered paired-end reads and 1.5-6.5 Gb of filtered Nanopore reads were used for assembly. An initial assembly was obtained by using the "ONT assembly and Illumina polishing pipeline" (https://github.com/nanoporetech/ont-assembly-polish). The assembly was further upgraded using the FinisherSC script (57) Mitochondrial contigs were removed from the assembly and were not used for any analysis. Additionally, small contigs were removed that contained a paired-end read coverage lower than 50% of the genome-wide average.

Long *U. nuda*, *U. loliicola* and *U. tritici* reads were obtained through Single Molecular Real-Time (SMRT) Sequencing using the PacBio Sequel system. A total of 6.3-9.7 Gb of raw long reads were obtained for the different species. The initial assembly was

obtained using the Canu assembler and was further upgraded with the FinisherSC script (57, 58). Mitochondrial contigs were remove from the assembly and were not used for further analysis.

The quality of genome assemblies was assessed by screening the presences of BUSCOs using the BUSCO software version 5.0.0 with the database "basidiomycota_odb10" (59).

Transposable element annotation and classification

The smut genome assemblies were scavenged for repetitive sequences in order to construct a repeat library for repeat annotation. Helitron TEs were identified using the EAHelitron script (60). LTR-RTs were identified using LTRharvest (61). Miniature inverted-repeat TEs were identified with MITE Tracker (62). Short interspersed nuclear elements were identified with the SINE-scan tool (63). Finally, RepeatModeler (v1.0.11) was also used for *de novo* repeat identification. These repeats were than combined with the repeat library from RepBase (release20170127) (64). The CD-HIT-EST tool under default settings was used to remove redundancy in the constructed library (65). RepeatMasker (v4.0.9) was then used to annotate the repeats to specific genome locations. The annotated repeat sequences were filtered on size and only sequences larger than 500bp were retained. Furthermore, repeats that were nested or had more than 50% overlap with other repeats were removed from the library. In case two repeats had reciprocally 50% overlap was the longest repeat retained. Repeats were classified into different TE orders using the PASTEC tool using PiRATE-Galaxy (66, 67).

***U. hordei* RNA sequencing and expression analysis**

Total RNA from *U. hordei* strain 4857-4 strain grown axenically and *in planta* was extracted for three biological replicates. For the axenic samples, *U. hordei* was grown in YEPS light

(0.4% yeast extract, 0.4% peptone, and 2% saccharose) liquid medium at 22°C with 200 rpm shaking till OD:1.0. For the *in planta* samples, Golden Promise barley cultivar was grown in a greenhouse at 70% relative humidity, at 22°C during the day and the night; with a light/dark regime of 15/9 hrs and 100 Watt/m^2 supplemental light when the sunlight influx intensity was less than 150 Watt/m^2. Barley plants were infected with *U. hordei* through needle injection as previously described (68) and samples were harvested 3dpi. Here, the third leaves of the *U. hordei* infected barley plants were collected by cutting 1 cm below the injection needle sites. Leaf samples were then frozen in liquid nitrogen and grinded using a mortar and pestle under constant liquid nitrogen. The total RNA was isolated by using the TRIzol® extraction method (Invitrogen; Karlsruhe, Germany) according to the manufacturer's instructions. Subsequently, total RNA samples were treated with Turbo DNA-Free™ Kit (Ambion/Applied Biosystems; Darmstadt, Germany) to remove any DNA contamination according to the manufacturer's instructions. Total RNA was then sent to for library preparation and sequencing to Novogene (Beijing, China). Libraries (250-300 bp insert size) were loaded on Illumina NovaSeq6000 System for 150bp paired-end sequencing using a S4 flowcell.

In total, 5.1-8.4 and 36.0-45.2 Gb of raw reads were obtained for the samples grown in liquid medium and *in planta*, respectively. The reads were filtered using the Trinity software (v2.9.1) option trimmomatic under the standard settings (69). The reads were then mapped to the reference genome using Bowtie 2 (v2.3.5.1) with the first 15 nucleotides on the 5'-end of the reads being trimmed due to inferior quality (70). The reads were mapped onto a combined file of the *U. hordei* strain Uh114 genome assembly and the *Hordeum vulgare* (IBSC_v2) (71) genome assembly. Reads were counted to the *U. hordei* loci using the R package Rsubread (v1.34.7) (72). Here the default minimum mapping quality score of 0 was used, to include reads that would have multiple best mapping locations. For the gene loci, reads were counted that were mapped to the predicted coding regions. For the LTR-RT

loci, reads were only counted that mapped within LTR-RT loci, excluding the reads that mapped onto the 10% of either edge of the locus. Loci were considered expressed if they had more than one count per million in at least two of the six samples (three replicates of two treatments). Significant differential expression of a locus was calculated using the R package edgeR (v3.26.8), using the function "decideTestsDGE" (73). Here, a threshold of log2 fold change of 1 was used and differential expression was determent using a p-value < 0.01 with Benjamini-Hochberg correction

Gene annotation

U. hordei genomes were annotated using the BRAKER v2.1.4 pipeline with RNA-Seq and protein supported training with the options "--softmasking" and "--fungus" enabled (74). RNA-seq reads from *U. hordei* grown in axenic culture and *in planta* (all replicates) were mapped to the assemblies using TopHat v2.1.1 (75). Protein predictions from numerous Ustilaginales species were used to guide the annotation, i.e. *Anthracocystis flocculosa*, *Melanopsichium pennsylvanicum*, *Moesziomyces antarcticus*, *S. reilianum*, *U. brachipodii-distachyi*, *U. hordei*, *U. maydis* (25–28, 76–78). *U. nuda* and *U. tritici* genomes were also annotated with the BRAKER v2.1.4 pipeline, but no RNA-seq data was used to guide the annotation. The option "--fungus" was enabled and the previously published protein files of the following species were used for protein supported training: *M. pennsylvanicum*, *S. reilianum*, *U. brachipodii-distachyi* and *U. maydis* (25–27, 76). Our annotation of *U. hordei* Uh805 was also included to train the annotation software. The *U. brachipodii-distachyi* and *U. maydis* genomes were previously annotated and this annotation was used for analysis (25, 27). Predicted genes that included an internal stop codon or did not start with a methionine were removed.

Secreted proteins are proteins with a predicted signal peptide using SignalP version 5.0 (79) and the absence of a transmembrane domain predicted with TMHMM2.0c in the protein sequence excluding the signal peptide (80). Gene Ontology (GO) terms were annotated to the *U. hordei* strain Uh114 protein prediction using InterProScan (v5.42-78.0) (81). Significance of GO term enrichments in a subset of genes were calculated with a Fisher exact test with the alternative hypothesis being one-sided (greater). The significance values of the multiple enrichments were corrected according Benjamini and Hochberg (82). Carbohydrate-Active enzymes (CAZymes) were annotated using the dbCAN2 meta server (83, 84). A protein was considered a CAZyme if at least two of the three tools (HMMER,DIAMOND and Hotpep) predicted a CAZyme function.

Comparative genomic analyses.

Phylogenetic trees were constructed based on BUSCOs from the database "basidiomycota_odb10" that are present without paralog in all members of the tree (59). For every gene, the encoded protein sequences were aligned using MAFFT (v7.464) option "--auto" (85). These aligned protein sequences were then concatenated for every species and used for tree construction using RAxML (v8.2.11) with substitution model "PROTGAMMAWAG" and 100 bootstraps (86). Here, protein sequences that were present in at least 60% of the tree members were excluded for tree construction.

Synteny block between the smut genome assemblies of were identified with SynChro with DeltaRBH = 3 (87, 88). The genome assembly of the epiphytic yeast *Moesziomyces bullatus* ex *Albugo* was included in this analysis to use as an outgroup (89). The ancestral chromosome gene order was constructed with AnChro with Delta´ = 3 and Delta´´ = 3 (88, 90). Inter-chromosomal rearrangements, i.e. translocations of two blocks, were identified with ReChro Delta = 10 (88, 90). No inter-chromosomal rearrangements in *U. nuda* could be

575 automatically detected by ReChro. Here, the inter-chromosomal rearrangement that lead to a
576 mating-type polarity switch was manually determined.

577 To determine the specificity of *MAT* locus sequences, absent/present polymorphisms
578 between *U. hordei* strains were determined with NUCmer (version 3.1) from the MUMmer
579 package with the option "--maxmatch" (91). From the same package, delta-filter with the
580 option "-1" was used to find the one-to-one alignments.

581

582 **LTR-RT evolution**

583 To know the sequence identity distribution, the best orthologous and paralogous LTR-RTs
584 were identified using blastn (v2.2.31+) (92). LTR-RTs that did not belong to an LTR-RT
585 family of multiple members, were excluded from the analysis. Members of the same LTR-RT
586 family share at least 80% sequence identity in at least 80% of their sequence with at least one
587 other member (19). Orthologous or paralogous LTR-RTs that have reciprocally the highest
588 bit-score were used for analysis. The nucleotide identity distribution of these orthologous and
589 paralogous LTR-RTs was constructed using Gaussian Kernel Density Estimation with a
590 kernel bandwidth of 1.5.

591 To reconstruct the ancestor LTR-RTs, a subset LTR-RTs were used. LTR-RTs were
592 included that were larger than 3 kb and smaller than 15 kb. Furthermore, repetitive sequences
593 within the LTR-RT (>50 bp) were indicated using blastn (v2.2.31+) and removed from the
594 sequence (92). The region between the repeats were then used for ancestor construction if this
595 region was larger than 500 bp. Here, bedtools (v.2.29.2) function "getfasta" was used (93).
596 Open reading frames (ORFs) and there encoding amino acid sequence of were determined
597 with esl-translate (-l 50) as part of the Easel (v0.46) package. Functional domain within these
598 amino acid sequences were determined with pfam_scan.pl (-e_seq 0.01) using the Pfam
599 database version 32.0 (94). Only sequences were included in the ancestor construction if they

had at least 3 different Pfam domains from the following domains: PF00078, PF00665, PF03732, PF07727, PF08284, PF13975, PF13976, PF14223, PF17917, PF17919 and PF17921. All of these predicted Pfam domain had to located on the same nucleotide strand in order to be used for ancestor construction. These sequences were then grouped in families, according to the definition that family members share at least 80% sequence identity in at least 80% of their sequence with at least one other member (19). Families were classified in *Copia* or *Gypsy* using the tool LTRclassifier (95). Ancestors were constructed using prank (v.170427) with the options "-showall" and "-F". Nucleotide substitutions between LTR-RTs and their constructed ancestor were then determined after they were aligned using MAFFT (v7.464) options "--auto" (85).

Gene divergence

One-to-one orthologs and homologs between *U. hordei* strains were found using the SiLiX (v.1.2.10-p1) software with the setting of at least 35% identity and 40% overlap (96). Homolog groups consisting of two members, each one of a different strain/species, were considered one-to-one homologs. Nucleotide substitutions for orthologs were identified after orthologs were aligned using MAFFT (v7.464) options "--auto" (85). Synonymous and nonsynonymous substitutions between orthologs were identified using SNAP (97). The nucleotide substitution level distributions were constructed using Gaussian Kernel Density Estimation with a kernel bandwidth of 0.5.

Data accession

Raw RNAseq reads and genome assemblies are deposited at NCBI under the BioProject PRJNA698760.

Acknowledgements

This work has been supported by the Alexander von Humboldt Foundation, the European Research Council (ERC 2017 COG 771035, conVIRgens), the Cluster of Excellence on Plant Sciences (CEPLAS; Germany's Excellence Strategy–EXC 2048/1 – Project ID: 390686111) and the University of Cologne. We also thank the Regional Computing Centre of Cologne (RRZK) for access to the Cologne High Efficient Operating Platform for Science (CHEOPS). We thank Guus Bakkeren for sharing the *U. hordei* strains with us and critically reading the manuscript. We thank Karl-Josef Müller for the isolation and sharing of the *U. nuda* and *U. tritici* strain with us.

635 **SUPPLEMENTAL MATERIAL**

636 **Table S1. Characteristics of the mating-type loci of smut species with a bipolar mating-**
637 **type system**.

638

639 **Table S2. Transposable elements annotation in various smut genome assemblies.**

640

641 **Figure S1. Phylogenetic relationship between *Ustilago hordei* lineages based on**
642 **Benchmarking Universal Single-Copy Orthologs (BUSCOs).** In total, 1,692 BUSCOs
643 were used for tree construction. Homologous BUSCO protein sequences were aligned using
644 MAFFT and then concatenated for tree construction using RAxML with substitution model
645 "PROTGAMMAWAG". *U. nuda* was used as on outgroup species to root the tree. The
646 robustness of the trees was assessed using 100 bootstrap replicates.

647

648 **Figure S2. Comparison of repeat content of gene flanking regions between expressed**
649 **secretome genes and other expressed genes.** Upregulated means a significantly higher
650 expression *in planta* compared to growth in axenic culture. In total, 20 kb sequences on each
651 side of the genes (40 kb in total) were considered as flanking regions. Significant differences
652 were calculated with a two-sided T-test. No significant differences with p-value < 0.01 were
653 found.

654

655 **Figure S3. Correlation between median nucleotide substitution level and fraction *U.***
656 ***hordei* (UH) specific genes for ortholog windows.** Ortholog windows of 75 UH genes with
657 a sliding step of 10 were used to determine the number of substitutions with *U. nuda*. UH
658 specific genes do not have an ortholog in *U. maydis*.

659

Figure S4. Comparison of nucleotide substitution type fractions for *U. brachipodii-distachyi*/*U. tritici* and *U. hordei*/*U. nuda* ortholog windows. The nucleotide substitutions were calculated for windows of 75 genes with a sliding step of 10. The x-axis consistently displays the total substitutions per 1 kb for these windows. (**A**) The y-axis depicts the fraction of every substation type (CT, AG, CG, AC, GT, AT) of ortholog windows. (**B**) The y-axis depicts the fraction of every substation type for the intergenetic regions of ortholog windows.

Figure S5. Correlations between the nucleotide substitution levels and the encoded amino acid composition of genes. Correlations were calculated for windows of 75 *U. hordei* genes with a sliding step of 10. Significant correlations, with p-value < 0.01, are indicated by a black edge around the square. (**A**) Correlations between amino acid compositions of encoded *U. hordei* proteins and the number of nucleotide substitutions with *U. nuda*. (**B**) Correlations between amino acid alternations for encoded *U. hordei* proteins and the number of nucleotide substitutions using *U. nuda* and *U. brachipodii-distachyi* orthologs as comparison.

Figure S6. Fractions of transitions and transversion of recently proliferated *U. hordei* long terminal repeat retrotransposons (LTR-RTs) and their flanking regions. The fraction of transitions and transversions between recently proliferated *U. hordei* LTR-RTs and their ancestors were determined. Fractions of the 20 kb flanking regions (40 kb in total) of the LTR-RT with *U. nuda* (UN) were also determined. Significant differences between LTR and flanking regions were determined for transitions and transversions separately with an unequal variance t-test. ***: p-value < 0.001.

References

1. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, Mager DL, Feschotte C. 2018. Ten things you should know about transposable elements. Genome Biol 19:199.
2. Stajich JE. 2017. Fungal genomes and instights into the evolution of the kingdom, p. 619–633. *In* The Fungal Kingdom.
3. Stukenbrock EH, Croll D. 2014. The evolving fungal genome. Fungal Biol Rev 28:1-12.
4. Katinka MD, Duprat S, Cornillott E, Méténler G, Thomarat F, Prensier G, Barbe V, Peyretaillade E, Brottier P, Wincker P, Delbac F, El Alaoul H, Peyret P, Saurin W, Gouy M, Weissenbach J. 2001. Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. Nature 414:450-453.
5. Corradi N, Pombert JF, Farinelli L, Didier ES, Keeling PJ. 2010. The complete sequence of the smallest known nuclear genome from the microsporidian *Encephalitozoon intestinalis*. Nat Commun 1:77.
6. Ramos AP, Tavares S, Tavares D, Silva MDC, Loureiro J, Talhinhas P. 2015. Flow cytometry reveals that the rust fungus, *Uromyces bidentis* (Pucciniales), possesses the largest fungal genome reported-2489Mbp. Mol Plant Pathol 16:1006-1010.
7. Tavares S, Ramos AP, Pires AS, Azinheira HG, Caldeirinha P, Link T, Abranches R, Silva M do C, Voegele RT, Loureiro J, Talhinhas P. 2014. Genome size analyses of Pucciniales reveal the largest fungal genomes. Front Plant Sci 5:422.
8. Cuomo CA, Bakkeren G, Khalil HB, Panwar V, Joly D, Linning R, Sakthikumar S, Song X, Adiconis X, Fan L, Goldberg JM, Levin JZ, Young S, Zeng Q, Anikster Y, Bruce M, Wang M, Yin C, McCallum B, Szabo LJ, Hulbert S, Chen X, Fellers JP. 2017. Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. G3 7:361–376.
9. Oggenfuss U, Badet T, Wicker T, Hartmann FE, Singh NK, Abraham L, Karisto P, Vonlanthen T, Mundt C, McDonald BA, Croll D. 2021. A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. Elife 10:e69249.
10. Ali S, Laurie JD, Linning R, Cervantes-Chávez JA, Gaudet D, Bakkeren G. 2014. An immunity-triggering effector from the barley smut fungus *Ustilago hordei* resides in an Ustilaginaceae-specific cluster bearing signs of transposable element-assisted evolution. PLoS Pathog 10:e1004223.
11. Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den Berg GCM, Wittenberg AHJ, Thomma BPHJ. 2016. Transposons passively and actively contribute to evolution of the two-speed genome of a fungal pathogen. Genome Res 26:1091–1100.
12. Seidl MF, Thomma BPHJ. 2014. Sex or no sex: Evolutionary adaptation occurs regardless. BioEssays 36:335–345.
13. Thines M. 2019. An evolutionary framework for host shifts – jumping ships for survival. New Phytol 224:605-617.
14. Cook DE, Mesarich CH, Thomma BPHJ. 2015. Understanding plant immunity as a surveillance system to detect invasion. Annu Rev Phytopathol 53:541–563.

15. Rouxel T, Balesdent MH. 2017. Life, death and rebirth of avirulence effectors in a fungal pathogen of Brassica crops, *Leptosphaeria maculans*. New Phytol 214:526-532.
16. Depotter JRL, Doehlemann G. 2020. Target the core: durable plant resistance against filamentous plant pathogens through effector recognition. Pest Manag Sci 76:426-431.
17. Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev 35:57–65.
18. Croll D, McDonald BA. 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog 8:e1002608.
19. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nat Rev Genet 8:973-982.
20. Finnegan DJ. 2012. Retrotransposons. Curr Biol 22:R432-R437.
21. Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. Genome Biol 5:225.
22. Zuo W, Ökmen B, Depotter JRL, Ebert MK, Redkar A, Misas Villamil J, Doehlemann G. 2019. Molecular interactions between smut fungi and their host plants. Annu Rev Phytopathol 57:411–430.
23. Möller M, Stukenbrock EH. 2017. Evolution and genome architecture in fungal plant pathogens. Nat Rev Microbiol 15:756–771.
24. Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol 10:417–430.
25. Kämper J, Kahmann R, Bölker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Müller O, Perlin MH, Wösten HAB, De Vries R, Ruiz-Herrera J, Reynaga-Peña CG, Snetselaar K, McCann M, Pérez-Martín J, Feldbrügge M, Basse CW, Steinberg G, Ibeas JI, Holloman W, Guzman P, Farman M, Stajich JE, Sentandreu R, González-Prieto JM, Kennell JC, Molina L, Schirawski J, Mendoza-Mendoza A, Greilinger D, Münch K, Rössel N, Scherer M, Vraněs M, Ladendorf O, Vincon V, Fuchs U, Sandrock B, Meng S, Ho ECH, Cahill MJ, Boyce KJ, Klose J, Klosterman SJ, Deelstra HJ, Ortiz-Castellanos L, Li W, Sanchez-Alonso P, Schreier PH, Häuser-Hahn I, Vaupel M, Koopmann E, Friedrich G, Voss H, Schlüter T, Margolis J, Platt D, Swimmer C, Gnirke A, Chen F, Vysotskaia V, Mannhaupt G, Güldener U, Münsterkötter M, Haase D, Oesterheld M, Mewes HW, Mauceli EW, DeCaprio D, Wade CM, Butler J, Young S, Jaffe DB, Calvo S, Nusbaum C, Galagan J, Birren BW. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444:97–101.
26. Schirawski J, Mannhaupt G, Münch K, Brefort T, Schipper K, Doehlemann G, Di Stasio M, Rössel N, Mendoza-Mendoza A, Pester D, Müller O, Winterberg B, Meyer E, Ghareeb H, Wollenberg T, Münsterkötter M, Wong P, Walter M, Stukenbrock E, Güldener U, Kahmann R. 2010. Pathogenicity determinants in smut fungi revealed by genome comparison. Science 330:1546-1548.
27. Rabe F, Bosch J, Stirnberg A, Guse T, Bauer L, Seitner D, Rabanal FA, Czedik-eysenberg A, Uhse S, Bindics J, Genenncher B, Navarrete F, Kellner R, Ekker H, Kumlehn J, Vogel JP, Gordon SP, Walter MC, Marcel TC, Mu M, Sieber CMK, Mannhaupt G, Gu U, Kahmann R, Djamei A. 2016. A complete toolset for the study of

Ustilago bromivora and *Brachypodium* sp. as a fungal-temperate grass pathosystem. Elife 5:e20522.

28. Laurie JD, Ali S, Linning R, Mannhaupt G, Wong P, Guldener U, Munsterkotter M, Moore R, Kahmann R, Bakkeren G, Schirawski J. 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.
29. Kruse J, Dietrich W, Zimmermann H, Klenke F, Richter U, Richter H, Thines M. 2018. *Ustilago* species causing leaf-stripe smut revisited. IMA Fungus 9:49–73.
30. Maire RCJ. 1919. Une ustilaginale nouvelle de la flore nord-Africaine. Bull la Société d'Histoire Nat l'Afrique du Nord 10:46–47.
31. Bakkeren G, Kronstad JW. 1994. Linkage of mating-type loci distinguishes bipolar from tetrapolar mating in basidiomycetous smut fungi. Proc Natl Acad Sci U S A 91:7085-7089.
32. Lee N, Bakkeren G, Wong K, Sherwood JE, Kronstad JW. 1999. The mating-type and pathogenicity locus of the fungus *Ustilago hordei* spans a 500-kb region. Proc Natl Acad Sci U S A 96:15026-15031.
33. Gillissen B, Bergemann J, Sandmann C, Schroeer B, Bölker M, Kahmann R. 1992. A two-component regulatory system for self/non-self recognition in *Ustilago maydis*. Cell 68:647-657.
34. Raudaskoski M, Kothe E. 2010. Basidiomycete mating type genes and pheromone signaling. Eukaryot Cell 9:847-859.
35. Bakkeren G, Jiang G, Warren RL, Butterfield Y, Shin H, Chiu R, Linning R, Schein J, Lee N, Hu G, Kupfer DM, Tang Y, Roe BA, Jones S, Marra M, Kronstad JW. 2006. Mating factor linkage and genome evolution in basidiomycetous pathogens of cereals. Fungal Genet Biol 43:655-666.
36. Yadav V, Sun S, Billmyre RB, Thimmappa BC, Shea T, Lintner R, Bakkeren G, Cuomo CA, Heitman J, Sanyal K. 2018. RNAi is a critical determinant of centromere evolution in closely related fungi. Proc Natl Acad Sci U S A 115:3108-3113.
37. Schweizer G, Munch K, Mannhaupt G, Schirawski J, Kahmann R, Dutheil J. 2018. Positively selected effector genes and their contribution to virulence in the smut fungus *Sporisorium reilianum*. Genome Biol Evol 10:629–645.
38. Zuther K, Kahnt J, Utermark J, Imkampe J, Uhse S, Schirawski J. 2012. Host specificity of *Sporisorium reilianum* is tightly linked to generation of the phytoalexin luteolinidin by *Sorghum bicolor*. Mol Plant-Microbe Interact 25:1230–1237.
39. Wang Q-M, Begerow D, Groenewald M, Liu X-Z, Theelen B, Bai F-Y, Boekhout T. 2015. Multigene phylogeny and taxonomic revision of yeasts and related fungi in the *Ustilaginomycotina*. Stud Mycol 81:55–83.
40. Depotter JRL, Shi-Kunne X, Missonnier H, Liu T, Faino L, van den Berg GCM, Wood TA, Zhang B, Jacques A, Seidl MF, Thomma BPHJ. 2019. Dynamic virulence-related regions of the plant pathogenic fungus *Verticillium dahliae* display enhanced sequence conservation. Mol Ecol 28:3482-3495.
41. Wicker T, Yu Y, Haberer G, Mayer KFX, Marri PR, Rounsley S, Chen M, Zuccolo A, Panaud O, Wing RA, Roffler S. 2016. DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nat Commun 7:12790.

42. Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18:337.340.
43. Hwang DG, Green P. 2004. Bayesian Markov chain Monte Carlo sequence analysis reveals varying neutral substitution patterns in mammalian evolution. Proc Natl Acad Sci U S A 101:13994-14001.
44. Habig M, Lorrain C, Feurtey A, Komluski J, Stukenbrock EH. 2021. Epigenetic modifications affect the rate of spontaneous mutations in a pathogenic fungus. Nat Commun 12:5869.
45. Lee GE, Mauro E, Parissi V, Shin CG, Lesbats P. 2019. Structural insights on retroviral DNA integration: learning from foamy viruses. Viruses 21:249-256.
46. Lesbats P, Engelman AN, Cherepanov P. 2016. Retroviral DNA integration. Chem Rev 116:12730-12757.
47. Torres DE, Oggenfuss U, Croll D, Seidl MF. 2020. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biol Rev 3:136-143.
48. Frantzeskakis L, Kusch S, Panstruga R. 2019. The need for speed□: compartmentalized genome evolution in filamentous phytopathogens. Mol Plant Pathol 20:3–7.
49. Stam R, Münsterkötter M, Pophaly SD, Fokkens L, Sghyer H, Güldener U, Hückelhoven R, Hess M. 2018. A new reference genome shows the one-speed genome structure of the barley pathogen *Ramularia collo-cygni*. Genome Biol Evol 10:3243-3249.
50. Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, Spanu PD, Maekawa T, Schulze-Lefert P, Panstruga R. 2018. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics 19:381.
51. Schwessinger B, Sperschneider J, Cuddy WS, Garnica DP, Miller ME, Taylor JM, Dodds PN, Figueroa M, Park RF, Rathjen P. 2018. A near-complete haplotype-phased genome of the dikaryotic. MBio 9:e02275-17.
52. Laurie JD, Linning R, Wong P, Bakkeren G. 2013. Do TE activity and counteracting genome defenses, RNAi and methylation, shape the sex lives of smut fungi? Plant Signal Behav 8:e23853.
53. Coelho MA, Bakkeren G, Sun S, Hood ME, Giraud T. 2018. Fungal Sex: The Basidiomycota, p. 147–175. *In* Heitman, J, Howlett, BJ, Crous, PW, Stukenbrock, EH, James, TY, Gow, NAR (eds.), The Fungal Kingdom. American Society of Microbiology, Washington DC.
54. Horváth V, Merenciano M, González J. 2017. Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends Genet 33:832-841.
55. Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D. 2020. Stress-driven transposable element de-repression dynamics and virulence evolution in a fungal pathogen. Mol Biol Evol 37:221-239.
56. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.
57. Lam KK, Labutti K, Khalak A, Lam K, Tse D. 2015. FinisherSC: a repeat-aware tool for upgrading de-novo assembly using long reads. Bioinformatics 31:3207–3209.

58. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu : scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res 27:722–736.
59. Seppey M, Manni M, Zdobnov EM. 2019. BUSCO: Assessing genome assembly and annotation completeness, p. 227-245. *In* Methods in Molecular Biology. Humana Press Inc.
60. Hu K, Xu K, Wen J, Yi B, Shen J, Ma C, Fu T, Ouyang Y, Tu J. 2019. Helitron distribution in Brassicaceae and whole genome Helitron density as a character for distinguishing plant species. BMC Bioinformatics 20:354.
61. Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.
62. Crescente JM, Zavallo D, Helguera M, Vanzetti LS. 2018. MITE Tracker: an accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics 19:348.
63. Mao H, Wang H. 2017. SINE-scan: An efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics 33:743-745.
64. Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.
65. Li W, Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659.
66. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, Quesneville H. 2014. PASTEC: An automatic transposable element classification tool. PLoS One 9:e91929.
67. Berthelier J, Casse N, Daccord N, Jamilloux V, Saint-Jean B, Carrier G. 2018. A transposable element annotation pipeline and expression analysis reveal potentially active elements in the microalga *Tisochrysis lutea*. BMC Genomics 19:378.
68. Ökmen B, Mathow D, Hof A, Lahrmann U, Aßmann D, Doehlemann G. 2018. Mining the effector repertoire of the biotrophic fungal pathogen *Ustilago hordei* during host and non-host infection. Mol Plant Pathol 19:2603–2622.
69. Grabherr MG., Brian J. Haas, Moran Yassour Joshua Z. Levin, Dawn A. Thompson, Ido Amit, Xian Adiconis, Lin Fan, Raktima Raychowdhury, Qiandong Zeng, Zehua Chen, Evan Mauceli, Nir Hacohen, Andreas Gnirke, Nicholas Rhind, Federica di Palma, Bruce W. N, Friedman and AR. 2013. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 29:644–652.
70. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
71. Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, Bayer M, Ramsay L, Liu H, Haberer G, Zhang XQ, Zhang Q, Barrero RA, Li L, Taudien S, Groth M, Felder M, Hastie A, Šimková H, Stanková H, Vrána J, Chan S, Munõz-Amatriaín M, Ounit R, Wanamaker S, Bolser D, Colmsee C, Schmutzer T, Aliyeva-Schnorr L, Grasso S, Tanskanen J, Chailyan A, Sampath D, Heavens D, Clissold L, Cao S, Chapman B, Dai F, Han Y, Li H, Li X, Lin C, McCooke JK, Tan C, Wang P, Wang S, Yin S, Zhou G, Poland JA, Bellgard MI, Borisjuk L, Houben A, Doleael J, Ayling S, Lonardi S, Kersey P, Langridge P,

902 Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close
903 TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein
904 N. 2017. A chromosome conformation capture ordered sequence of the barley genome.
905 Nature 544:427-433.
906 72. Liao Y, Smyth GK, Shi W. 2019. The R package Rsubread is easier, faster, cheaper
907 and better for alignment and quantification of RNA sequencing reads. Nucleic Acids
908 Res 47:e47.
909 73. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for
910 differential expression analysis of digital gene expression data. Bioinformatics 26:139-
911 140.
912 74. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation
913 with BRAKER, p. 65–95. *In* Methods in Molecular Biology. Humana Press Inc.
914 75. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2□:
915 accurate alignment of transcriptomes in the presence of insertions , deletions and gene
916 fusions. Genome Biol 14:R36.
917 76. Sharma R, Mishra B, Runge F, Thines M. 2014. Gene loss rather than gene gain is
918 associated with a host jump from monocots to dicots in the smut fungus
919 *Melanopsichium pennsylvanicum*. Genome Biol Evol 6:2034–2049.
920 77. Lefebvre F, Joly DL, Labbe C, Teichmann B, Linning R, Belzile F, Bakkeren G,
921 Belanger RR. 2013. The transition from a phytopathogenic smut ancestor to an
922 anamorphic biocontrol agent deciphered by comparative whole-genome analysis. Plant
923 Cell 25:1946–1959.
924 78. Morita T, Koike H, Koyama Y, Hagiwara H, Ito E, Fukuoka T, Imura T, Machida M,
925 Kitamoto D. 2013. Genome sequence of the basidiomycetous yeast *Pseudozyma*
926 *antarctica* T-34 , a producer of the glycolipid biosurfactants. Genome Announc
927 1:e00064-13.
928 79. Juan J, Armenteros A, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S,
929 Heijne G Von, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using
930 deep neural networks. Nat Biotechnol 37:420–423.
931 80. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting
932 transmembrane protein topology with a hidden Markov model: application to complete
933 genomes. J Mol Biol 305:567-580.
934 81. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J,
935 Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M,
936 Yong SY, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function
937 classification. Bioinformatics 30:1236–1240.
938 82. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate□: a practical and
939 powerful approach to multiple testing. J R Stat Soc B 57:289–300.
940 83. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. DbCAN: a web resource for
941 automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445-451.
942 84. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y.
943 2018. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation.
944 Nucleic Acids Res 46:W95-W101.
945 85. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version

7: Improvements in performance and usability. Mol Biol Evol 30:772–780.

86. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.

87. Drillon G, Carbone A, Fischer G. 2014. SynChro: a fast and easy tool to reconstruct and visualize synteny blocks along eukaryotic chromosomes. PLoS One 9:e92621.

88. Drillon G, Carbone A, Fischer G. 2013. Combinatorics of chromosomal rearrangements based on synteny blocks and synteny packs. J Log Comput 23:815-838.

89. Eitzen K, Sengupta P, Kroll S, Kemen E, Doehlemann G. 2021. A fungal member of the *Arabidopsis thaliana* phyllosphere antagonizes *Albugo laibachii* via a GH25 lysozyme. Elife 10:e65306.

90. Vakirlis N, Sarilar V, Drillon G, Fleiss A, Agier N, Meyniel JP, Blanpain L, Carbone A, Devillers H, Dubois K, Gillet-Markowska A, Graziani S, Huu-Vang N, Poirel M, Reisser C, Schott J, Schacherer J, Lafontaine I, Llorente B, Neuvéglise C, Fischer G. 2016. Reconstruction of ancestral chromosome architecture and gene repertoire reveals principles of genome evolution in a model yeast genus. Genome Res 26:918-932.

91. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL. 2004. Versatile and open software for comparing large genomes. Genome Biol 5:R12.

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

93. Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

94. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427-D432.

95. Monat C, Tando N, Tranchant-Dubreuil C, Sabot F. 2016. LTRclassifier: a website for fast structural LTR retrotransposons classification in plants. Mob Genet Elements 6:e1241050.

96. Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

97. Korber B. 2000. HIV signature and sequence variation analysis., p. 55–72. *In* Rodrigo, AG, Learn, GH (eds.), Computational analysis of HIV molecular sequences. Kluwer Academic Publishers, Dordrecht.