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)

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

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 long terminal repeat retrotransposons (LTR-RTs) of the *Copia* and *Gypsy* superfamilies. From the studied smuts, LTR-RTs proliferated the most recently and to the furthest extent in the *U. hordei* genome, in which they make up for 19.5% of the genome. Interestingly, the extent of TE proliferation in different smut species is positively correlated to the mating-type locus size, which is largest in *U. hordei* with up to ~560 kb. TE transposition within the mating-type loci and their flanking regions are mating-type specific, which is likely due to the very low recombination activity in this region. Furthermore, LTR-RT proliferation was found to be associated with higher nucleotide substitution levels, as genes in genome regions that are rich in dynamic LTR-RTs display higher nucleotide substitution levels. The high nucleotide substitution rate particularly affected the evolution of genes encoding secreted proteins as substitutions more frequently led to amino acid alterations. The mechanism behind this increase in nucleotide substitution rate remains elusive, but seems not to be a consequence of the repeat-induced point mutation (RIP) mechanism, as genes and LTR-RTs did not display typical RIP substitutions.

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 of TE proliferation (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). For plant pathogenic fungi, TE transposition can be a source of mutagenesis to evade host immunity and/or lead to an optimized host interaction (9). TEs can also passively contribute to mutagenesis, as their transpositions increase homologous genome sequences that are prone to ectopic recombination (10, 11). Pathogens evolve by host jumps, radiation and subsequent arms races with their hosts (12), in which the latter attempts to detect pathogen ingress through the recognition of so-called invasion patterns (13). One invasion pattern that is typically detected are effectors, i.e. secreted proteins that facilitate host colonization (14). To quickly adapt to effector-triggered immunity and yet continue host symbiosis, effector genes often reside in genome regions that facilitate mutagenesis (12, 15), such as those rich in TEs (16). 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 (17).

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 (18). Retrotransposons with direct repeats at each end of their sequence are long terminal repeat retrotransposons (LTR-RTs) (19). LTR-RTs can encode the structural and enzymatic machinery for autonomous transposition. However, LTR-RT may lose this ability through mutagenesis, but still be able to transpose using proteins of other LTR-RTs (20). These non-autonomous LTR-RT derivates can be classified into large retrotransposon derivative (LARD) and terminal-repeat retrotransposon in miniature (TRIM) depending on their size. 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 (18). A major type of DNA transposons are terminal inverted repeats (TIRs). Terminal repeats are recognized by the transposase enzyme for excision. Non-autonomous TIR derivatives that cannot encode their own transposase are classified as miniature inverted TEs (MITEs) (21).

TE activity can have detrimental outcomes for the hosting genome. Consequently, organisms have evolved several protection mechanisms to prevent uncontrolled TE proliferation (22). TE activity can be epigenetically suppressed through DNA methylation, histone modification or RNA interference (RNAi). Epigenetic TE suppression is reversible and, thus, TEs can be re-activated, which typically occurs under particular stress conditions (23). Alternatively, TEs can be irreversibly de-activated through mutagenesis (24). In fungi, repeat-induced point mutation (RIP) is a meiosis-associated defence mechanism that generates G to A and C to T transitions in duplicated sequences (25). RIP was first observed in the ascomycete *Neurospora crassa* and has hitherto been experimentally verified in a limited number of other ascomycetes (26–31). In basidiomycetes, RIP has never been experimentally verified, but biases towards RIP substitution in repetitive DNA have been reported (32–34).

Smut fungi are a diverse group of plant pathogenic basidiomycetes of which many infect monocot plants, in particular grasses (35). Generally, these pathogens have small genomes compared to other plant pathogens (36, 37). 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 (38, 39). 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 (40). *Ustilago hordei* and *U. brachipodii-distachyi* are two related smut fungi and have assembled genomes of 21.15 Mb and 20.5 Mb, respectively (33, 40). 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 (40). The assembled genome of *U. brachipodii-distachyi* is originally published under the species name *U. bromivora* (Rabe et al. 2016). *U. brachipodii-distachyi* infect members from the tribe Brachipodieae, whereas *U. bromivora* affects bromes from the supertribe Triticodae (41, 42). 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 (40, 41).

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 (40, 43, 44). 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 (45, 46). 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 (44). A large fraction of the mating-type loci consists of repetitive sequences, i.e. ~45% repeats for *U. hordei* (33, 47). In contrast, the

tetrapolar smuts, *U. maydis* and *S. reilianum*, have there *a* and *b* loci on different chromosomes causing them to segregate independently during meiosis (39, 43).

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 (33), a 27.1 Mb assembly was obtained (48). Thus, the *U. hordei* genome underwent a recent genome expansion as it is more than 30% larger than that of its relative *U. brachipodii-distachyi*, which genome is assembled on chromosome level and is 20.5 Mb in size (40). This finding triggered us to study the *U. hordei* genome more in depth and use recently developed long-read sequencing technology to unravel how its genome recently expanded.

Results

LTR-RT proliferation drives *U. hordei* genome size

To study the expansion in genome size of *U. hordei*, we sequenced 6 *U. hordei* strains of different geographic origin (Figure S1). To obtain a high-quality assembly with large contigs and high base calling confidence, we used paired-end Illumina reads in combination with long reads obtained with the Oxford Nanopore MinION device. 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. The resulting assemblies were composed of 23-46 contigs and ranged from 25.8-27.2 Mb in size (Table 1). In total, 98.6-99.0% of the basidiomycete Benchmarking Universal Single-Copy Orthologs (BUSCOs) were present in the *U. hordei* assemblies (Table 1) (49) which is similar to the quality of the *U. maydis* reference genome assembly (98,8% of the BUSCOs) (38). 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 (40). Repetitive sequences in the assemblies were identified and classified into TE orders (Table 1). In general, class I TE sequences (retrotransposons) are over 6 times more abundant than class II TE sequences (DNA transposons) in *U. hordei*. 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 (Figure 1A, Table 1). For the class II TE sequences, more than 75% are TIRs with a total sequence amount ranging from 589 to 667 kb (Table 1). The number of LTR-RT sequences is positively correlated to the assembly size (r = 0.94, p-value = 0.0051), whereas the number of TIR TEs is not (r = -0.25, p-value = 0.63). 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. In conclusion, the variation in genome size between *U. hordei* strains can largely be attributed to intraspecific differences in LTR-RT proliferation and retention.

Table 1. Transposable elements annotation in various smut genome assemblies.

Species	*U. hordei*						*U. nuda*	*U. brachipodii-distachyi*[40]	*U. tritici*	*U. loliicola*	*U. maydis*[38]	*S. reilianum*[50, 51]
Strain	Uh359	Uh805	Uh811	Uh818	Uh1273	Uh1278	DE_29490	UB2112	Ut_3	K530	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
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
LARD (kb)	126	54	213	136	233	104	106	39	43	65	11	0
TRIM (kb)	4	16	11	11	11	12	4	1	4	0	0	0
DIRS (kb)	163	117	59	117	97	122	98	135	119	0	0	0
LINE (kb)	8	0	0	5	22	7	9	4	71	27	3	0
SINE (kb)	0	1	0	0	1	2	1	0	1	1	0	0
Class II TEs (kb)[§]	781	746	708	791	731	692	395	473	285	482	5	103
TIR (kb)	604	619	589	667	612	622	342	418	242	322	5	100
MITE (kb)	12	13	11	18	13	12	11	9	5	88	0	2
Crypton (kb)	31	30	34	34	23	30	42	45	38	0	0	0
Maverick (kb)	38	11	22	21	31	15	0	0	0	0	0	0
Helitron (kb)	92	73	47	45	51	3	0	0	0	72	0	0

[§] Only repetitive sequences that were larger than 500 bp were classified

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

To guide gene annotation and to study gene/LTR-RT expression, RNA was extracted and sequenced from *U. hordei* grown in liquid medium and in barley leaves at 3 days post infection (dpi). RNA 150bp paired-end reads were obtained through the Illumina NovaSeq6000 System. 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. Genes were annotated with BRAKER with the strain Uh803 assembly as reference using RNA-Seq data and protein homology (53). In total, 7704 genes were predicted of which 396 were predicted to encode for proteins with a signal peptide and without a transmembrane domain, hereafter referred to as secreted proteins. RNA-seq reads were mapped to the Uh803 genome assembly and quantified for gene and LTR-RT loci. In total, 6229 of the 7704 (81%) predicted gene loci were expressed in either of the two *U. hordei* growth conditions, whereas only 19 of the 904 (2%) LTR-RTs displayed expression (more than one count per million of expressed sequence in at least two of the six samples) (Figure 1B). Moreover, only 7 of the expressed LTR-RTs displayed expression in more than half of their sequence in at least one of the samples. 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-RTs were inactive in the two tested environmental conditions. For the genes, 978 (16%) were upregulated *in planta*, whereas 1178 (19%) were downregulated (Figure 1B). Up- and downregulated genes were screened for Gene Ontology (GO) term enrichments, to see which biological processes are affected upon plant colonization. In total, 33 and 23 GO terms were enriched in *in planta* up- and down-regulated genes, respectively (Table S1). Generally, processes associated with the fungal membrane, including transmembrane transport were upregulated *in planta*, whereas cytoplasmic

processes associated with ribosome, mitochondrial membrane and proteasome were downregulated (Table S1). In correspondence with these results, 214 of the 978 genes upregulated *in planta* encode secreted proteins. Thus, 59% (214/360) of the expressed genes that encode secreted proteins were upregulated *in planta*, which is a significant enrichment (Fisher exact test, *p*-value = 1.26e-86) (Figure 1B). In contrast, only 10% of the genes encoding a secreted protein were downregulated. Of these downregulated genes, 35% (13/37) were predicted to have a carbohydrate-active (CAZyme) function, whereas this is 17% (36/214) for *in planta* upregulated secretome genes, which is significantly less (Fisher exact test, *p*-value = 0.014). In conclusion, the *U. hordei* transmembrane transport system and secreted proteins are strongly activated upon plant colonization, whereas hardly any LTR-RTs display expression.

The majority of the mating-type locus sequences are mating-type specific

Several plant pathogens have genomes compartmentalized in in a way that facilitate fast effector repertoire evolution (12), such as repeat-rich regions (16). To test if TE-rich regions are the cradle for the evolution of new pathogenicity effectors in *U. hordei*, the location of secreted proteins and abundance of repeats and presence/absence polymorphisms were plotted across the *U. hordei* genome (Figure 2). Strain Uh805 was used as a reference as it contains 23 contigs that represent separate chromosomes. In these analyses, we considered windows of 10 kb upstream of 5'-ends and 10kb downstream of 3'-ends of *U. hordei* genes and determined the fraction of repeats and presence/absence polymorphisms within these 20 kb windows surrounding genes. Upregulated genes encoding secreted proteins, which are promising effector candidates (15, 54), have a median of 13% of repetitive sequences in their flanking region. This is not significantly different from other expressed genes that have a median of 17% (t-test, *p*-value = 0.32). Thus, in contrast to some other filamentous plant

pathogens, *in planta* upregulated genes encoding secreted proteins are not especially associated with repeat-rich genome regions in *U. hordei*. For presence/absence polymorphisms, we calculated the fraction of the window that is absent in at least one of the five other *U. hordei* strains. Here, upregulated genes encoding secreted proteins and other expressed genes were also not significantly different from each other (t-test, *p*-value = 0.36, both median = 0). In conclusion, *U. hordei* effector candidates, i.e. *in planta* upregulated genes encoding secreted proteins, are not enriched in repeat-rich genome regions that encounter frequent presence/absence polymorphisms.

When we plotted the *U. hordei* chromosomes, our attention was drawn to a ±816 kb genome region on chromosome 1 between genome coordinates roughly at 950-1,766 kb, which in only 27% of the region had a one-to-one best alignment to strain Uh1278, whereas this ranged from 95% to 100% with other strains (Figure 2-3A). Moreover, this region is repeat rich as it consists of 76% repeats and 36% of classified LTR-RTs, which is around double the genome wide average of 35% and 17%, respectively (Table1). The *MAT-1* locus, the region between and including the *a* and *b* loci, is embedded in this region and is 560 kb in size, consisting of 77% of repetitive sequences (Table S2). Uh1278 is the only sequenced *MAT-2 U. hordei* strain in this study (Tables S2). The low number of one-to-one best alignment to strain Uh1278, indicates that most sequences within the mating-type locus and its flanking regions are mating-type specific (Figure 3A). Similarly to *MAT-1*, *MAT-2* resides in a genome region of ±829 kb that has a high repeat (77%) and LTR-RT (35%) content (Figure 3a). The sequenced *MAT-2* locus is 472 kb, which is smaller than any of the sequenced *MAT-1* loci that ranged from 536 to 564 kb (Table S2). Homologous recombination is suppressed in the mating-type region (44). Consequently, any mutation within these regions is by definition mating-type specific. The repetitive nature and the large

size of the mating-type loci indicates that many mating-type specific sequences likely originate from TE transpositions (Figure 3B).

LTR-RT proliferation in smuts with a bipolar mating-type system

As LTR-RT played a predominant role in the genome expansion of *U. hordei*, we also studied the impact of TE dynamics on the genome evolution of *U. hordei* sister species. We sequenced genomes of *Ustilago loliicola*, *Ustilago nuda* and *Ustilago tritici*, which are smut species that are close relatives of *U. hordei* and *U. brachipodii-distachyi* (41, 55). Using the PacBio Sequel system we obtained a total of 6.3-9.7 Gb of raw long reads for the strains listed in Table 1. Assemblies of 20.8, 21.4 and 20.4 Mb were obtained in 41, 31 and 32 contigs for *U. loliicola*, *U. nuda* and *U. tritici*, respectively. In total, 98.8-98.9% of the BUSCOs were present in the assemblies (Table 1) (49). A phylogenetic tree was constructed based on 1,667 BUSCO proteins including the newly sequenced species as well as *U. brachipodii-distachyi*, *U. maydis* and *S. reilianum* (Figure 4A) (38, 40, 50, 51). *U. hordei*, *U. nuda*, *U. brachipodii-distachyi* and *U. tritici* cluster together (HNBT-cluster hereafter) with *U. loliicola* being the closest outgroup species. Within the cluster, *U. hordei* diverged the most recently from *U. nuda*, which also infects *Hordeum* species (Figure 4A). Synteny between the different contigs was also investigated and the ancestral gene order reconstructed. Intriguingly, *S. reilianum* and *U. loliicola* do not have inter-chromosomal rearrangement in comparison to their reconstructed last common ancestor (Figure 4A). The *U. maydis* genome has one inter-chromosomal rearrangement with respect to its last common ancestor with *S. reilianum.* HNBT-cluster species share one inter-chromosomal rearrangement that occurred after their divergence from *U. loliicola* (Figure 4A). As a consequence of this rearrangement, the *a* and *b* mating-type loci came to reside on the same chromosome in linked to each other, resulting in the mating-type polarity switch from

tetrapolar to bipolar (43). This inter-chromosomal rearrangement is the only one observed in the assemblies of *U. brachipodii-distachyi*, *U. nuda* and *U. hordei*, whereas the *U. tritici* assembly has one additional inter-chromosomal rearrangement. In conclusion, HNBT-cluster species have a bipolar mating system due to an inter-chromosomal rearrangement that occurred after their divergence from *U. loliicola*.

To see if mating-type polarity relates to TE evolutionary dynamics, repeats were annotated and classified in the TE orders for the different genomes in a similar fashion to the *U. hordei* strains. In total, 22.6% of the *U. nuda* assembly consists of repetitive sequences, which is 13-16% less than *U. hordei* from which *U. nuda* recently diverged (Table 1, Figure 4A). *Ustilago brachipodii-distachyi* (17.0 %) and *U. tritici* (16.4%) have similar levels of repetitive sequences in their genomes (Table 1). The *U. loliicola*, *U. maydis* and *S. reilianum* assemblies, which outgroup the HNBT-cluster, consist 8.9, 4.6 and 2.7% of repeats, respectively. Thus, species of the HNBT-cluster underwent an expansion of repeat content in their genome (Figure 4A). Interestingly, the repeat and the LTR-RT content of the HNBT-cluster species genomes positively correlates to the size of their mating-type loci (r = 0.98, p-value = 0.02, using strain Uh805 for *U. hordei*), which ranges from 190 kb for *U. brachipodii-distachyi* to 560 kb for *U. hordei* (Figure 4A, Table S2). This confirms our hypothesis that the size of the *MAT* locus is determined by TE proliferation and the insertion of TEs in the *MAT* locus and its flanking regions (Figure 3B).This increase and interspecific differences in repeat content between HNBT-cluster species can largely be attributed to differences in LTR-RT content that ranged from 462 kb in *U. tritici* to 4,326 kb in *U. hordei* (Uh805) (Table 1). In contrast, *U. loliicola*, *U. maydis* and *S. reilianum* contain 9, 185 and 5 kb of LTR-RT sequences, respectively (Table 1). Thus, LTR-RTs proliferated after the divergence of *U. loliicola* and the HNBT-cluster, which is also the moment when the mating-type polarity switched from tetrapolar to bipolar.

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

Although HNBT-cluster species collectively encountered a LTR-RT proliferation, there are large differences within the cluster as *U. hordei* has more than 9 times the amount of LTR-RT sequences than *U. tritici* (Table 1). To study intraspecific differences, the nucleotide sequence identity distribution of the best reciprocal paralogous and orthologous LTR-RT sequences was calculated (Figure 4B). This was on the one hand done for the recently diverged species with the highest LTR-RT content, *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 at 82-83% and at 88-90% (Figure 4B). The maximum of the orthologous LTR-RTs between *U. brachipodii-distachyi* and *U. tritici* was at 93%. Thus, LTR-RTs in *U. brachipodii-distachyi* and *U. tritici* mainly proliferated before the last common ancestor of the HNBT species (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 4B). Thus, in contrast to *U. brachipodii-distachyi* and *U. tritici*, LTR-RTs continued to proliferated after the HNBT last common ancestor.

As TE-active genome regions have been associated with distinct nucleotide substitution regimes in fungi (10, 56), we studied if different extents of LTR-RT proliferation is associated with different nucleotide substitution regimes. Gene nucleotide divergence in *U. brachipodii-distachyi*/*U. tritici* and *U. hordei*/*U. nuda* was tested for clustering, where a cluster was defined to comprise at least 10 of a fraction of the most diverged orthologs in a window of 20 genes. Orthologs are defined to encode proteins with at least 80% identity and 80% overlap. Using this definition, the fraction of clustering increases with increasing fraction of the most divergent orthologs, until around 25% where the clustering plateaus

(Figure 4C). We further compared the clustering between species by using 20% of the most divergent genes as its central within the 15 and 25% range where the clustering progresses in a more or less constant linear relationship with the percentage of the most diverged orthologs (Figure 4C). These clusters using 20% most diverged genes are hereafter referred to as Diverged Gene Clusters (DGCs). For *U. brachipodii-distachyi*/*U. tritici* 24% of the 20% most diverged orthologs reside in DGCs, which is at least two times less than for ortholog comparison involving *U. hordei* and/or *U. nuda* (48-65%) (Figure 4C). Thus, gene nucleotide sequence divergence occurs more clustered in *U. hordei* and *U. nuda* than in *U. brachipodii-distachyi* and *U. tritici*, which might be linked to the more recent LTR-RT proliferation of *U. hordei* and *U. nuda* (Figure 4A-B). In total, 89% and 78% of the *U. brachipodii-distachyi* DGC genes are the same using *U. hordei* (1,346/1,507) and *U. nuda* (1,346/1,718) orthologs, respectively (Figure S2). Thus, this indicates that clustered higher nucleotide substitution rates occurred to a similar extent in *U. hordei* and *U. nuda* since their last common ancestor and mainly affect the same genome regions.

We then looked if DGCs are distinct in gene density, intraspecific presence/absence polymorphisms or LTR-RT content from other genome regions (Figure S3). For this, genome windows of 20 kb were used and the DGCs were constructed based on the 20% most diverged *U. hordei* genes with *U. brachipodii-distachyi* (Figure 2). We compared these two species as (1) LTR-RT proliferated to the furthest extent in *U. hordei* and (2) *U. brachipodii-distachyi* is its closest relative where LTR-RT did not extensively proliferate after the HNBT last common ancestor (Figure 4). The median 20 kb window in *U. hordei* DGCs consist 46% of expressed gene sequence. This is significantly less than the median of 51% for genome regions outside DGCs (t-test, *p*-value = 0.025) (Figure S3). The fraction of Uh803 sequences that is absent in at least one of the other sequenced *U. hordei* strains is not significantly different within and outside DGCs (t-test, *p*-value = 0.12) (Figure S3). The fraction of

sequences within DGCs that is annotated as LTR-RT has a median of 21%, whereas this is 0% and significantly less for other genome regions (t-test, *p*-value = 2.95e-5) (Figure S3). In conclusion, DGCs are generally gene poorer and LTR-RT richer than other genome regions.

To test if higher substitution levels are associated with LTR-RT proliferation, we looked if interspecific LTR-RT variation especially occurs in DGCs. For genes that have a single ortholog in every species of the HNBT-cluster, a window of 10 kb upstream of its 5'-end and 10 kb downstream of its 3'-end was constructed, and the fraction of this 20 kb window that consists of LTR-RT was determined. Then, the absolute difference of this fraction for every *U. hordei* gene was calculated with its corresponding ortholog of one of the HNBT-cluster species. These differences were then compared for genes that reside in DGCs and genes that do not. The median differences in LTR-RT content of *U. hordei* genes in DGCs were 8, 9 and 6% using the orthologs of *U. nuda*, *U. brachipodii-distachyi* and *U. tritici*, respectively (Figure 5). In contrast, genes that do not reside in DGCs have a median change of 0% independent of the species that was used for orthologs (Figure 5). This difference in LTR-RT alterations is significant (Welch's t-test; *U. nuda*: *p*-value = 5.63e-13, *U. brachipodii-distachyi*: *p*-value = 8.52e-23 and *U. tritici*: *p*-value = 4.92e-15). The same phenomenon was also observed for *U. brachipodii-distachyi* DGC genes when compared to *U. hordei* orthologs (Figure 5). Genes inside DGCs had a median LTR-RT alteration of 8%, whereas this was 0% for genes outside DGCs (Welch's t-test, *p*-value = 1.17e-20). For the *U. nuda* and *U. tritici* orthologs, medians of 0% occurred inside and outside *U. brachipodii-distachyi* DGCs (Figure 5). Conceivably, the difference in LTR-RT content of *U. brachipodii-distachyi* with *U. nuda* and *U. tritici* were too small to find significance (Table 1). In conclusion, differences in LTR-RT content between *U. hordei* and the other species occur more frequently in genome regions were genes encountered a higher nucleotide substitution level.

LTR-RTs do not display typical RIP substitutions

RIP is an genome defence mechanism that induces substitution levels in repetitive nucleotide sequences. However, RIP leaks to genome regions neighbouring these repeats, which can affect coding regions flanking TEs (57). To test if LTR-RTs are affected by RIP in HNBT-cluster species, 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 (18). 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 TE-LTR families were reconstructed using 252 LTR-RT sequences (Figure 6A). The largest three families contained 65, 56 and 31 LTR-RTs respectively and contained only *U. hordei* and *U. nuda* members. The ancestor of the two largest families were from the *Copia* superfamily, whereas the third largest belong to *Gypsy* and contained a chromodomain at the C-terminal region of the integrase, so-called chromovirus (58). The substitutions that occurred between this subset of LTR-RTs and their respective ancestor are on average 84% transitions (Figure 6B). RIP targets repeats through C→T and G→A substitutions. On average, 19% of the encountered LTR-RT substitutions since their ancestor are T→C, which is significantly smaller than the fraction of 24% for the typical RIP substitution C→T (t-test, *p*-value = 1.10e-23). Conversely, the typical RIP substitution G→A occurred with an average of 19% less frequently, than A→G substitutions (average 22%) (Figure 6B) (t-test, *p*-value = 5.27e-7). In conclusion, there is no clear enrichment for C→T and G→A substitutions between the LTR-RTs and their reconstructed ancestor.

High local nucleotide substitution levels especially impact secretome evolution

To see if higher nucleotide substitution levels affect a particular group of genes more than others, we tested if genes residing in DGCs are enriched for particular GO terms. Also here we opted to choose *U. hordei* and *U. brachipodii-distachyi* for comparison due to their respective extensive and lack of LTR-RT proliferation after the HNBT last common ancestor (Figure 4). No single GO term was enriched in *U. hordei* DGC genes indicating that a higher local nucleotide substitution rate does not affect particular gene types more than others. Also, DGC genes were not more likely to encode secreted proteins (Fisher exact test, *p*-value = 0.80 Although high nucleotide substitution levels affect all type of genes, the consequences for the evolution of particular genes might be different. Presence/absence polymorphisms were compared within and outside DGCs. In DGCs, 27% of the *U. hordei* genes do not have an *U. brachipodii-distachyi* ortholog that encode for a protein with at least 80% identity and 80% overlap. Outside DGCs, 23 % of the *U. hordei* genes do not have such orthologs, which is significantly less than inside DGCs (Fisher exact test, *p*-value = 0.00047). Intriguingly, this difference is pronounced for genes that encode secreted proteins. Inside DGCs, 44% of the secretome genes are *U. hordei* specific, whereas this is 27% outside DGCs (Fisher exact test, *p*-value = 0.0038). Thus, the increased level of presence/absence polymorphisms in DGCs affects especially genes encoding secreted proteins.

We also studied the nucleotide substitution types. Substitutions were determined for *U. hordei* and *U. brachipodii-distachyi* orthologs (at least 35% identity and 40% overlap) that are single copy genes (using the same 35% and 40% cut-offs), as paralogs can impact the nucleotide substitution regime (59). In general, transitions had medians ranging from 0.016 to 0.019 transitions per codon outside DGCs, whereas transversions ranged from 0.0018 to 0.0056 transitions per codon (Figure 7A). Thus, similar to the LTR-RT substitutions, transitions occur more frequently than transversion in gene coding regions. Within DGCs,

median transitions and transversion ranged from 0.021 to 0.025 and from 0.0037 to 0.024 substitutions per codon, respectively (Figure 7A). All type of substitutions were significantly higher for DGCs genes compared to genes outside DGCs (Welch's t-test, *p*-value < 0.001) (Figure 7A). Intriguingly, the number of CG substitutions for DGC genes had a median that was 4.26 times higher than for genes outside DGCs, whereas this ranged from 1.29 to 2.03 for other substitution types (Figure 7A). Although *U. hordei* encountered more recent and a higher extant of LTR-RT proliferation than *U. brachipodii-distachyi* (Figure 4B, Table 1), C→T, T→C, A→G and G→A substitutions between these two species were all higher inside DGCs compared to outside (Figure 7B). Thus, the higher nucleotide substitution rate in DGCs does not display typical RIP substitutions patterns, rather all substitution types are increased.

Nucleotide substitutions in coding regions can be synonymous or nonsynonymous in nature. Outside *U. hordei* DGCs, genes contain a median of 0.04 ds, whereas this is 0.06 within DGCs, which is a significant increase (Welch's t-test, *p*-value = 2.60e-78) (Figure 7C). Similarly, the median dn is 0.01 outside and 0.02 inside DGCs, respectively (Welch's t-test, *p*-value = 5.85e-35) (Figure 7D). Within DGCs, the median fraction of dn and ds (dn/ds) is 0.35 for genes that encode a secreted protein, whereas this is significantly less for other genes that have a median of 0.30 (Welch's t-test, *p*-value = 0.031) (Figure 7E). In conclusion, the higher nucleotide substitution levels in DGCs do more frequently lead to amino acid changes in the secreted proteins than in proteins that are not secreted.

Discussion

Nucleotide substitution rates are unevenly distributed across genomes and can be influenced by numerous factors, including neighbouring nucleotides, recombination frequencies and TE activity (60–62). Nucleotide divergence in *U. hordei* and *U. nuda* occurs more clustered in

their genomes compared to *U. brachipodii-distachyi* and *U. tritici* (Figure 4C). These former two smut species encountered a more recent and higher extent of LTR-RT proliferation compared to the latter two (Figure 4B). LTR-RT proliferation appears to be associated with higher nucleotide substitution rates, as *U. hordei* genome regions with high nucleotide substitution levels are enriched for dynamic LTR-RTs (Figure 5). However, the exact mechanism by which this occurs remains uncertain. The occurrence of genome regions with higher nucleotide substitution rates is likely not a consequence of RIP, as LTR-RTs do not more frequently encounter the typical RIP substitutions (Figure 6). Moreover, the higher local nucleotide substitution levels can be observed for all substitution types (Figure 7A). The associated higher nucleotide substitution levels in repeat-rich genome regions has been previously found in other eukaryotes (63). It has been proposed that repeat-rich genome regions are replicated by more error-prone polymerases. More specifically, it has been suggested that DNA polymerase errors may be introduced during TE insertion. In rice, higher nucleotide substitutions levels occur close to LTR-RT insertion sites (62). LTR-RT insertion causes cuts in the host DNA that is then ligated by the host (64, 65). Wicker et al. (2016) suggest that cut host DNA might become a target for 3'->5' exonuclease resulting in a segment of single-stranded DNA (62). The complementary strand of this stretch of DNA would then be synthesized by a replication complex with lower DNA polymerase fidelity and mismatch repair. Conversely, higher substitution rates can also be the cause rather the consequence why LTR-RT proliferation is observed in genome regions with higher nucleotide substitution levels. Genome regions with pre-existing higher substitution regimes may be more prone to TE fixation, as the more frequent substitutions increase the chance of TE inactivation. Inactive TEs are generally less harmful for the host and may therefore have lower chances to be selected against. However, the less clustered occurrence of high nucleotide substitution levels in the LTR-RT poorer species *U. brachipodii-distachyi* and *U.*

tritici indicates that LTR-RT proliferation is the driver of clustered high nucleotide substitutions rather than the consequence (Figure 4C).

Local higher levels of nucleotide substitutions impact the evolution of the genes that reside in the affected genome regions. Generally, more nonsynonymous substitutions can be observed (Figure 7D), but it is especially the genes that are encoding secreted proteins that encounter amino acid alterations (Figure 7E). Conceivably, the generally more accessory nature of secretome genes can make them more prone for amino acid changes (66). Although the effect of nucleotide substitutions affected a particular gene type more than others, enrichments for particular gene types could not be found. Also, genome regions with higher nucleotide substitution levels, which are gene poorer and LTR-RT richer (Figure S3), are not enriched for secretome genes, which include effector candidates. Thus, the model that effector candidates often reside in accessory, repeat-rich genome regions that facilitates fast adaptation, so-called two speed genome model (17, 37), does not hold for the DGCs. More generally, repeat content and intraspecific presence/absence polymorphisms, features that have been associated with the two-speed genome model, were not more frequently found in the proximity of effector candidate genes compared to other genes (Figure 2) (56, 67). The specificity and the universality of the two-speed genome model for filamentous plant pathogens has recently been contested (68, 69). 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 mentioned *P. striiformis* f. sp. *tritici* and the barley powdery mildew pathogen *Blumeria grami*nis f. sp. *hordei* (70–72).

LTR-RT proliferation is the main factor that led to the *U. hordei* genome expansion (Table 1). The expansion occurred especially in the mating-type locus that increased almost three times in size in comparison to *U. brachipodii-distachyi* (Table S2). The lack of the purifying recombination ability in this genome region can be the reason why LTR-RTs

especially accumulated in the mating-type and flanking genome regions (33, 44). Conceivably, this process is reinforced by the increasing presence of repetitive sequences as the transposition into a repeat-rich genome region is less likely to have a severe fitness cost than a transposition in repeat-poor region. Furthermore, the co-occurrence of the LTR-RTs proliferation in the HNBT branch and the switch in mating-type organization from tetra- to bipolar may indicate that mating-type polarity switch impacts LTR-RT proliferation (73). In case of biallelic *a* and *b* loci, the switch from a tetra-to bipolarity results in a basidiospore compatibility change from 25 to 50%. Consequently, it takes a tetrapolar smut on average longer to find a mating type than a bipolar smut. This longer time might increase the opportunity to mate with spores from a different offspring and, thus, increase outcrossing. The higher outcrossing rate for tetra- compared to bipolar smuts is even more pronounced when multiallelism exists for the *a* and *b* loci (74). Multiallelism increased the compatibility on population level, whereas compatibility within the same offspring remains 25 and 50% for tetra- and bipolar smuts, respectively. Lower levels of outcrossing reduce the purifying recombination ability of smuts, which may be the reason why LTRs could proliferate to a further extent in bipolar smuts (33, 73).

TEs are import drivers of genome evolution as they directly cause mutagenesis through their transpositions and indirectly increase the change of non-homologous recombination due to their repetitive nature (11). LTR-RT proliferation in *U. hordei* indicates that TE activity also influence local nucleotide substitution regimes and increase the substitution levels in the genome regions where they insert. Consequently, genes in the proximity of these insertion sites evolve faster (Figure 7D). Fast gene evolution may be advantageous under stressful condition, when TEs are typically more active or change their activity (23, 75).

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 (9). 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 Baseballing 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 (76). 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 (77) Mitochondrial contigs were remove from the assembly and was 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. loliicola*, *U. nuda* and *U. tritici* reads were obtained through Single Molecular Real-Time (SMRT) Sequencing using the PacBio Sequel system. *U. loliicola* K530 was isolated by Julia Kruse from *Lolium perenne* in Untermerzbach, Germany. *U. nuda* DE_29490 and *U. tritici* Ut_3 were both isolated by Karl-Josef Müller in Darzau, Germany.

The initial assembly was obtained using the Canu assembler and was further upgraded with the FinisherSC script (77, 78). 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” (49).

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 (79). LTR-RTs were identified using LTRharvest (80). Miniature inverted-repeat TEs were identified with MITE Tracker (81). Short interspersed nuclear elements were identified with the SINE-scan tool (82). 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) (83). The CD-HIT-EST tool under default settings was used to remove redundancy in the constructed library (84). 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 (85, 86).

***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 (87) 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.

The reads were filtered using the Trinity software (v2.9.1) option trimmomatic under the standard settings (88). 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 (89). The reads were mapped onto a combined file of the *U. hordei* strain Uh114 genome assembly and the *Hordeum vulgare* (IBSC_v2) (90) genome assembly. Reads were counted to the *U. hordei* loci using the R package Rsubread (v1.34.7) (91). 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. Significant differential expression of a locus was calculated using the R package edgeR (v3.26.8) (92). Significance of differential expression was calculated using t-tests relative to a threshold of log2 fold change of 1 with Benjamini-Hochberg correction. The cut-off of 0.05 was used.

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 (53). 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 (93). 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* (33, 38–40, 94–96). *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* (38–40, 94). Our annotation of *U. hordei* Uh805 was also included to train the annotation software. The *U. brachipodii-distachyi* genome was previously annotated and this annotation was used for analysis (40). 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 (97) and the absence of a transmembrane domain predicted with TMHMM2.0c in the protein sequence excluding the signal peptide (98). Gene Ontology (GO) terms were annotated to the *U. hordei* strain Uh114 protein prediction using InterProScan (v5.42-78.0) (99). 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 (100). Carbohydrate-Active enzymes (CAZymes) were annotated using the dbCAN2 meta server (101, 102). 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 (49). For every gene, the encoded protein sequences were aligned using MAFFT (v7.464) option "--auto" (103). 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 (104). Here, protein sequences that were present in least than 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 (105, 106). The genome assembly of the epiphytic yeast *Moesziomyces bullatus* ex *Albugo* was included in this analysis to use as an outgroup (107). The ancestral chromosome gene order was constructed with AnChro with Delta´ = 3 and Delta´´ = 3 (106, 108). Inter-chromosomal rearrangements, i.e. translocations of two blocks, were identified with ReChro Delta = 10 (106, 108). No inter-chromosomal rearrangements in *U. nuda* could be automatically detected by ReChro. Here, the inter-chromosomal rearrangement that lead to a mating-type polarity switch was manually determined.

The absent/present polymorphisms between the *U. hordei* strains were determined with NUCmer (version 3.1) from the MUMmer package with the option "--maxmatch" (109). From the same package, delta-filter with the option "-1" was used to find the one-to-one

alignments. A BED format file was created for the genome regions of the reference genome that had a one-to-one alignment to the query genome. The BED format files of the different query genomes with the same reference genome were then combined and the coverage for every genome region was determined with the software bedtools (v.2.29.2) to determine the percentage of presence a particular genome region in the other *U. hordei* strains (110). Bedtools was also used to divide the genome in windows for which particular genome feature were calculated, such as gene density, repeat content and presence/absence polymorphisms.

LTR-RT evolution

To know the sequence identity distribution, the best orthologous and paralogous LTR-RTs were identified using blastn (v2.2.31+) (111). LTR-RTs that did not belong to an LTR-RT family of multiple members, were excluded from the analysis. Members of the same LTR-RT family share at least 80% sequence identity in at least 80% of their sequence with at least one other member (18). Orthologous or paralogous LTR-RTs that have reciprocally the highest bit-score were used for analysis. The nucleotide identity distribution of these orthologous and paralogous LTR-RTs was constructed using Gaussian Kernel Density Estimation with a kernel bandwidth of 1.5.

To reconstruct the ancestor LTR-RTs, a subset of the HNBT-cluster LTR-RTs were used. LTR-RTs were included that were larger than 3 kb and smaller than 15 kb. Furthermore, repetitive sequences within the LTR-RT (>50 bp) were indicated using blastn (v2.2.31+) and removed from the sequence (111). The region between the repeats were then used for ancestor construction if this region was larger than 500 bp. Here, bedtools (v.2.29.2) function “getfasta” was used (110). Open reading frames (ORFs) and there encoding amino acid sequence of were determined with esl-translate (-l 50) as part of the Easel (v0.46) package. Functional domain within these amino acid sequences were determined with

pfam_scan.pl (-e_seq 0.01) using the Pfam database version 32.0 (112). 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 (18). Ancestors were constructed for families with at least four members, using prank (v.170427) with the options "-showall" and "-F". For every family was the last common ancestor used that could be classified in *Copia* or *Gypsy* using the tool LTRclassifier (113). If LTRclassifier was unable to classify the LTR-RT, a more recent ancestor was used as a reference. Nucleotide substitutions between LTR-RTs and their constructed ancestor were then determined after they were aligned using MAFFT (v7.464) options "--auto" (103).

Gene divergence

One-to-one orthologs and homologs between *U. hordei* strains were found using the SiLiX (v.1.2.10-p1) software (114). Homolog groups consisting of two members, each one of a different strain/species, were considered one-to-one homologs. Fractions of the most divergent orthologs were determined using the sequence identity of the hit with the best bit score using blastn (v2.2.31+) (111). DGCs were identified with CROC with a minimum number of 10 genes that can be found for a cluster using a gene window of 20 (115). Here, the Benjamini and Hochberg method was used to correct for multiple testing. To extract the genes within DGCs the software bedtools (v.2.29.2) was used (110). Also genes that did not have an orthologs were considered in the DGC analyses. To overlap of DGC genes was presented in venn diagrams that were constructed using InteractiVenn (116). Nucleotide

substitutions for orthologs were identified after orthologs were aligned using MAFFT (v7.464) options "--auto" (103). Synonymous and nonsynonymous substitutions between orthologs were identified using SNAP (117).

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), and 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.

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. 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.
10. 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.
11. Seidl MF, Thomma BPHJ. 2014. Sex or no sex: Evolutionary adaptation occurs regardless. BioEssays 36:335–345.
12. Thines M. 2019. An evolutionary framework for host shifts – jumping ships for survival. New Phytol 224:605-617.
13. Cook DE, Mesarich CH, Thomma BPHJ. 2015. Understanding plant immunity as a surveillance system to detect invasion. Annu Rev Phytopathol 53:541–563.
14. 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.
15. Depotter JRL, Doehlemann G. 2020. Target the core: durable plant resistance against filamentous plant pathogens through effector recognition. Pest Manag Sci 76:426-431.
16. Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev 35:57–65.
17. Croll D, McDonald BA. 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog 8:e1002608.
18. 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.
19. Finnegan DJ. 2012. Retrotransposons. Curr Biol 22:R432-R437.
20. Havecker ER, Gao X, Voytas DF. 2004. The diversity of LTR retrotransposons. Genome Biol 5:225.
21. Jiang N, Feschotte C, Zhang X, Wessler SR. 2004. Using rice to understand the origin

and amplification of miniature inverted repeat transposable elements (MITEs). Curr Opin Plant Biol 7:115-119.

22. Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nat Rev Genet 8:272–285.

23. 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.

24. Kinsey JA, Garrett-Engele PW, Cambareri EB, Selker EU. 1994. The *Neurospora* transposon Tad is sensitive to repeat-induced point mutation (RIP). Genetics 138:657-664.

25. Selker EU, Stevens JN. 1985. DNA methylation at asymmetric sites is associated with numerous transition mutations. Proc Natl Acad Sci U S A 82:8114-8118.

26. Selker EU, Cambareri EB, Jensen BC, Haack KR. 1987. Rearrangement of duplicated DNA in specialized cells of *Neurospora*. Cell 51:741-752.

27. Graïa F, Lespinet O, Rimbault B, Dequard-Chablat M, Coppin E, Picard M. 2001. Genome quality control: RIP (repeat-induced point mutation) comes to *Podospora*. Mol Microbiol 40:586-595.

28. Ikeda KI, Nakayashiki H, Kataoka T, Tamba H, Hashimoto Y, Tosa Y, Mayama S. 2002. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. Mol Microbiol 45:1355-1364.

29. Cuomo CA, Güldener U, Xu J-R, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma L-J, Baker SE, Rep M, Adam G, Antoniw J, Baldwin T, Calvo S, Chang Y-L, Harris LJ, Hilburn K, Kennell JC, Kroken S, Magnuson JK, Muehlbauer G, Münsterkötter M, Nelson D, Donnell KO, Ouellet T. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science 317:1400–1402.

30. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, Schmutz J, Taga M, White GJ, Zhou S, Schwartz DC, Freitag M, Ma LJ, Danchin EGJ, Henrissat B, Coutinho PM, Nelson DR, Straney D, Napoli CA, Barker BM, Gribskov M, Rep M, Kroken S, Molnár I, Rensing C, Kennell JC, Zamora J, Farman ML, Selker EU, Salamov A, Shapiro H, Pangilinan J, Lindquist E, Lamers C, Grigoriev I V., Geiser DM, Covert SF, Temporini E, Vanetten HD. 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet 5:e1000618.

31. Idnurm A, Howlett BJ. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. Fungal Genet Biol 39:31-37.

32. Horns F, Petit E, Yockteng R, Hood ME. 2012. Patterns of repeat-induced point mutation in transposable elements of basidiomycete fungi. Genome Biol Evol 4:240-247.

33. 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.

34. Hane JK, Anderson JP, Williams AH, Sperschneider J, Singh KB. 2014. Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. PLoS Genet 10:e1004281.
35. Zuo W, Odiekmen 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.
36. Möller M, Stukenbrock EH. 2017. Evolution and genome architecture in fungal plant pathogens. Nat Rev Microbiol 15:756–771.
37. Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol 10:417–430.
38. 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.
39. 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.
40. 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.
41. 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.
42. Maire RCJ. 1919. Une ustilaginale nouvelle de la flore nord-Africaine. Bull la Société d'Histoire Nat l'Afrique du Nord 10:46–47.
43. 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.
44. 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.
45. 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.
46. Raudaskoski M, Kothe E. 2010. Basidiomycete mating type genes and pheromone signaling. Eukaryot Cell 9:847-859.
47. 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.
48. 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.
49. 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.
50. 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.
51. 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.
52. Zhang H, Meltzer P, Davis S. 2013. RCircos: an R package for Circos 2D track plots. BMC Bioinformatics 14:244.
53. Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER, p. 65–95. *In* Methods in Molecular Biology. Humana Press Inc.
54. Depotter JRL, Zuo W, Hansen M, Zhang B, Xu M, Doehlemann G. 2021. Effectors with different gears: divergence of *Ustilago maydis* effector genes is associated with their temporal expression pattern during plant infection. J Fungi 7:16.
55. 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.
56. 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.
57. Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ. 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. PLoS Pathog 6:e1001180.
58. Gorinšek B, Gubenšek F, Kordiš D. 2004. Evolutionary genomics of chromoviruses in eukaryotes. Mol Biol Evol 21:781-798.
59. Wagner A. 2002. Selection and gene duplication: a view from the genome. Genome Biol 3:reviews1012.1.

60. Lercher MJ, Hurst LD. 2002. Human SNP variability and mutation rate are higher in regions of high recombination. Trends Genet 18:337.340.
61. 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.
62. 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.
63. McDonald MJ, Wang WC, Huang H Da, Leu JY. 2011. Clusters of nucleotide substitutions and insertion/deletion mutations are associated with repeat sequences. PLoS Biol 9:e1000622.
64. 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.
65. Lesbats P, Engelman AN, Cherepanov P. 2016. Retroviral DNA integration. Chem Rev 116:12730-12757.
66. Badet T, Croll D. 2020. The rise and fall of genes: origins and functions of plant pathogen pangenomes. Curr Opin Plant Biol 56:65-73.
67. Haas BJ, Kamoun S, Zody MC, Jiang RHY, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T, Bozkurt TO, Ah-Fong AM V, Alvarado L, Anderson VL, Armstrong MR, Avrova A, Baxter L, Beynon J, Boevink PC, Bollmann SR, Bos JIB, Bulone V, Cai G, Cakir C, Carrington JC, Chawner M, Conti L, Costanzo S, Ewan R, Fahlgren N, Fischbach M a, Fugelstad J, Gilroy EM, Gnerre S, Green PJ, Grenville-Briggs LJ, Griffith J, Grünwald NJ, Horn K, Horner NR, Hu C-H, Huitema E, Jeong D-H, Jones AME, Jones JDG, Jones RW, Karlsson EK, Kunjeti SG, Lamour K, Liu Z, Ma L, Maclean D, Chibucos MC, McDonald H, McWalters J, Meijer HJG, Morgan W, Morris PF, Munro C a, O'Neill K, Ospina-Giraldo M, Pinzón A, Pritchard L, Ramsahoye B, Ren Q, Restrepo S, Roy S, Sadanandom A, Savidor A, Schornack S, Schwartz DC, Schumann UD, Schwessinger B, Seyer L, Sharpe T, Silvar C, Song J, Studholme DJ, Sykes S, Thines M, van de Vondervoort PJI, Phuntumart V, Wawra S, Weide R, Win J, Young C, Zhou S, Fry W, Meyers BC, van West P, Ristaino J, Govers F, Birch PRJ, Whisson SC, Judelson HS, Nusbaum C. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461:393–398.
68. 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.
69. Frantzeskakis L, Kusch S, Panstruga R. 2019. The need for speed□: compartmentalized genome evolution in filamentous phytopathogens. Mol Plant Pathol 20:3–7.
70. 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.
71. 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.

72. 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.

73. 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.

74. 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.

75. Horváth V, Merenciano M, González J. 2017. Revisiting the relationship between transposable elements and the eukaryotic stress response. Trends Genet 33:832-841.

76. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114-2120.

77. 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.

78. 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.

79. 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.

80. Ellinghaus D, Kurtz S, Willhoeft U. 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9:18.

81. 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.

82. 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.

83. Bao W, Kojima KK, Kohany O. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mob DNA 6:11.

84. 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.

85. 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.

86. 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.

87. 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.

88. 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.
89. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359.
90. 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, Muehlbauer GJ, Clark MD, Caccamo M, Schulman AH, Mayer KFX, Platzer M, Close TJ, Scholz U, Hansson M, Zhang G, Braumann I, Spannagl M, Li C, Waugh R, Stein N. 2017. A chromosome conformation capture ordered sequence of the barley genome. Nature 544:427-433.
91. Liao Y, Smyth GK, Shi W. 2019. The R package Rsubread is easier, faster, cheaper and better for alignment and quantification of RNA sequencing reads. Nucleic Acids Res 47:e47.
92. Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139-140.
93. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2□: accurate alignment of transcriptomes in the presence of insertions , deletions and gene fusions. Genome Biol 14:R36.
94. Sharma R, Mishra B, Runge F, Thines M. 2014. Gene loss rather than gene gain is associated with a host jump from monocots to dicots in the smut fungus *Melanopsichium pennsylvanicum*. Genome Biol Evol 6:2034–2049.
95. Lefebvre F, Joly DL, Labbe C, Teichmann B, Linning R, Belzile F, Bakkeren G, Belanger RR. 2013. The transition from a phytopathogenic smut ancestor to an anamorphic biocontrol agent deciphered by comparative whole-genome analysis. Plant Cell 25:1946–1959.
96. Morita T, Koike H, Koyama Y, Hagiwara H, Ito E, Fukuoka T, Imura T, Machida M, Kitamoto D. 2013. Genome sequence of the basidiomycetous yeast *Pseudozyma antarctica* T-34 , a producer of the glycolipid biosurfactants. Genome Announc 1:e00064-13.
97. Juan J, Armenteros A, Tsirigos KD, Sønderby CK, Petersen TN, Winther O, Brunak S, Heijne G Von, Nielsen H. 2019. SignalP 5.0 improves signal peptide predictions using deep neural networks. Nat Biotechnol 37:420–423.
98. Krogh A, Larsson B, Von Heijne G, Sonnhammer ELL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete

genomes. J Mol Biol 305:567-580.
99. Jones P, Binns D, Chang HY, Fraser M, Li W, McAnulla C, McWilliam H, Maslen J, Mitchell A, Nuka G, Pesseat S, Quinn AF, Sangrador-Vegas A, Scheremetjew M, Yong SY, Lopez R, Hunter S. 2014. InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240.
100. Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate□: a practical and powerful approach to multiple testing. J R Stat Soc B 57:289–300.
101. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. 2012. DbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40:W445-451.
102. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, Busk PK, Xu Y, Yin Y. 2018. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 46:W95-W101.
103. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol 30:772–780.
104. Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30:1312–1313.
105. 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.
106. Drillon G, Carbone A, Fischer G. 2013. Combinatorics of chromosomal rearrangements based on synteny blocks and synteny packs. J Log Comput 23:815-838.
107. 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.
108. 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.
109. 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.
110. Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.
111. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215:403–410.
112. 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.
113. 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.
114. Miele V, Penel S, Duret L. 2011. Ultra-fast sequence clustering from similarity networks with SiLiX. BMC Bioinformatics 12:116.

115. Pignatelli M, Serras F, Moya A, Guigó R, Corominas M. 2009. CROC: finding chromosomal clusters in eukaryotic genomes. Bioinformatics 25:1552–1553.
116. Heberle H, Meirelles VG, da Silva FR, Telles GP, Minghim R. 2015. InteractiVenn: a web-based tool for the analysis of sets through Venn diagrams. BMC Bioinformatics 16:169.
117. 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.

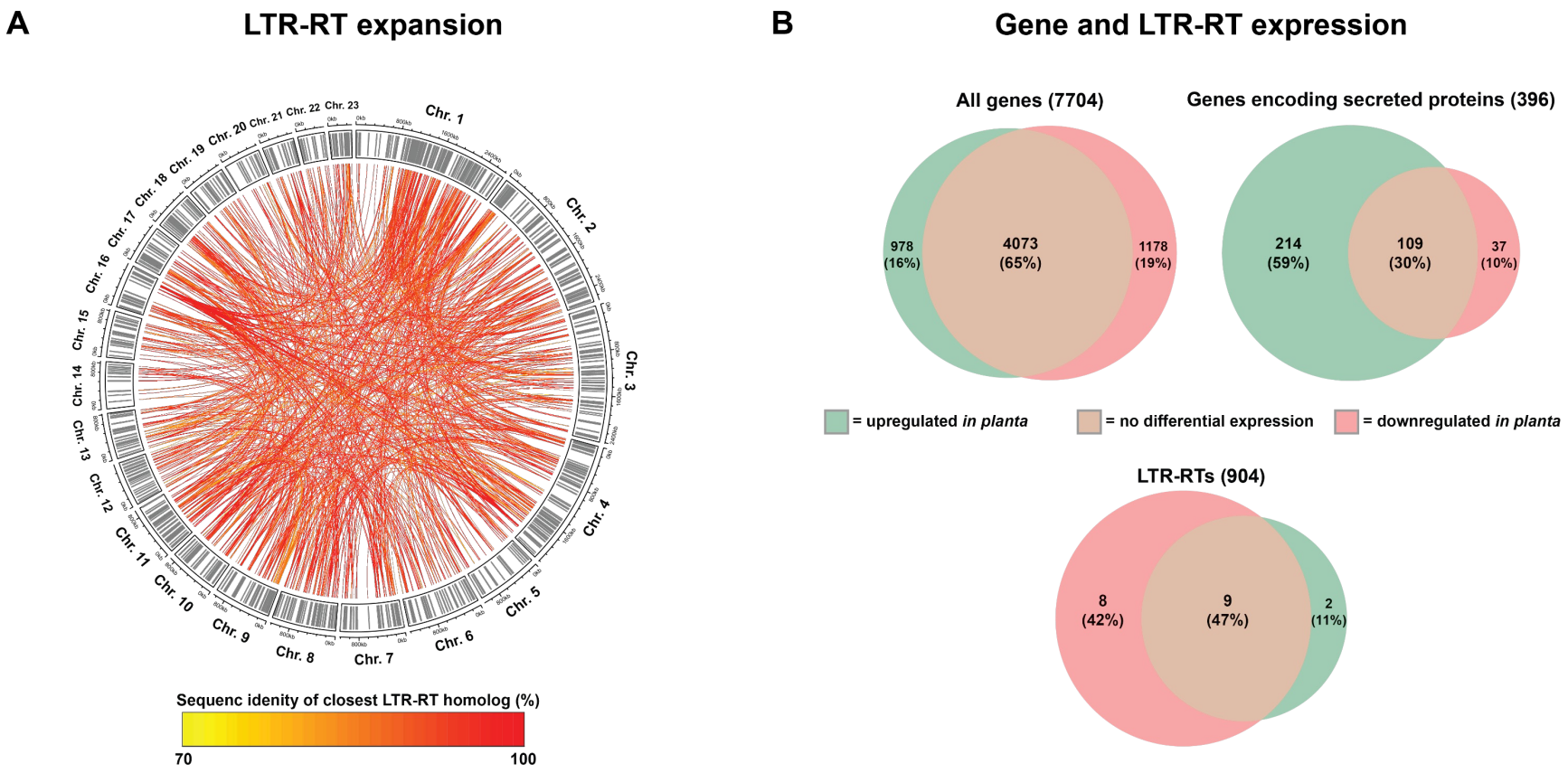

Figure 1: Long terminal repeat retrotransposons (LTR-RTs) recently proliferated in *Ustilago hordei* and are generally transcriptionally inactive. (A) The chromosomes of *U. hordei* strain Uh805 are depicted. The numbering of the chromosomes corresponds to the numbering of the previously published *U. bromivora* assembly (40). The grey lines on the chromosome bars indicate LTR-RTs. The best homolog for every LTR-RT is connected with ribbons in the centre of the circos plot (52). The colour of the ribbon indicates the sequence identity between the homologous region of the two LTR-RTs. (**B**) Differential expression of gene and LTR-RT *U. hordei* loci between growth in liquid culture medium and *in planta*. The numbers between brackets indicate how many genes and LTR-RTs have been annotated in the genome. The significance of differential expression was calculated using *t*-tests relative to a threshold of log2-fold-change of 1 with Benjamini-Hochberg correction. A *p*-value cut-off of 0.05 was used.

The genome architecture of *Ustilago hordei*

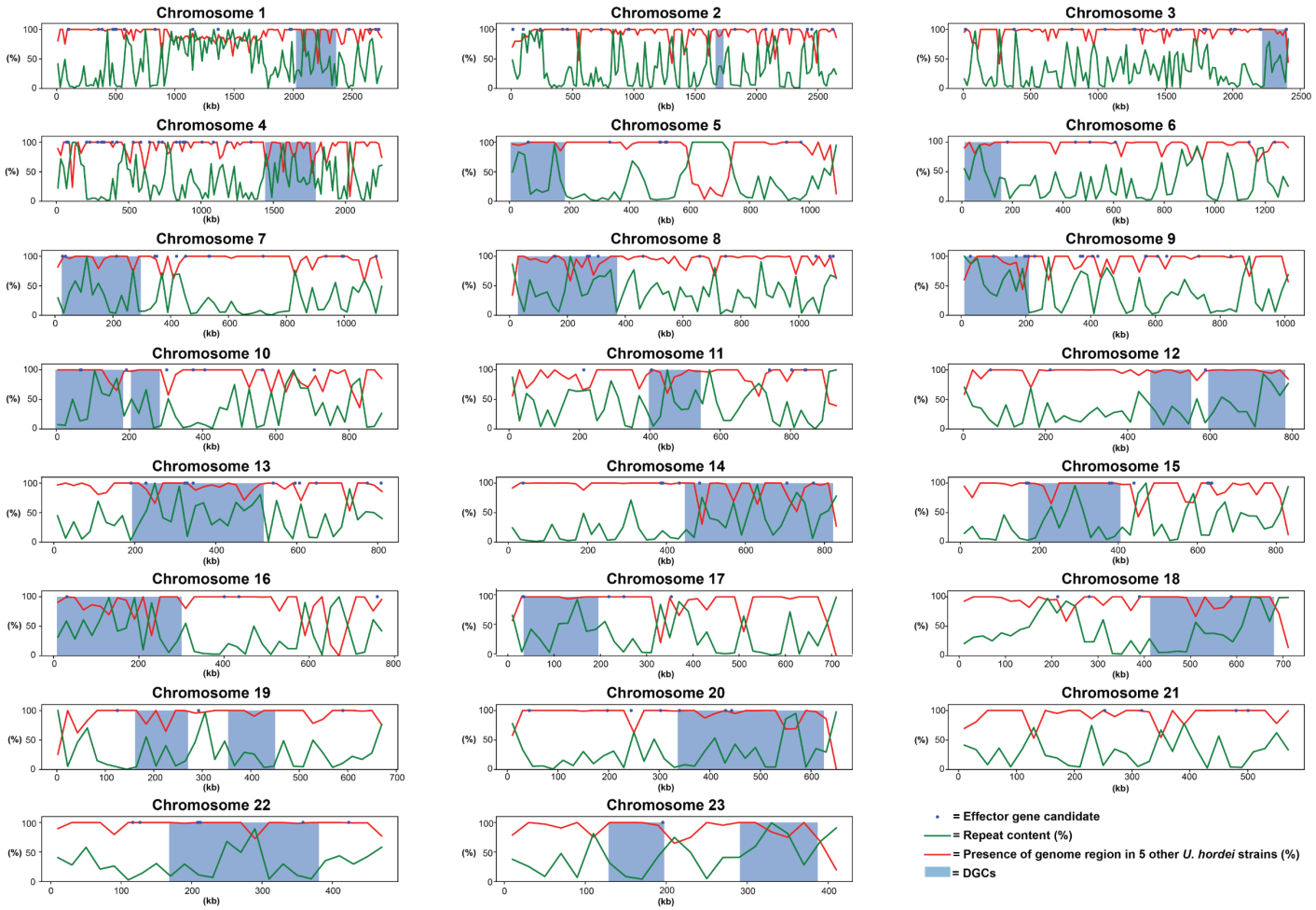

Figure 2: The *Ustilago hordei* genome architecture. The chromosomes of *U. hordei* strain Uh805 are depicted. The numbering of the chromosomes corresponds to the numbering of the previously published *U. brachipodii-distachyi* assembly (40). Repeat content and percentage presence/absence polymorphisms were calculated for 20 kb windows. The presence of Uh805 genome regions in the five other *U. hordei* strains is shown (Table 1). The location of effector gene candidates, i.e. *in planta* upregulated genes encoding secreted proteins, is depicted. Also, the location of the Divergent Gene Clusters (DGCs) is plotted for the 20% most divergent orthologs with *Ustilago brachipodii-distachyi*.

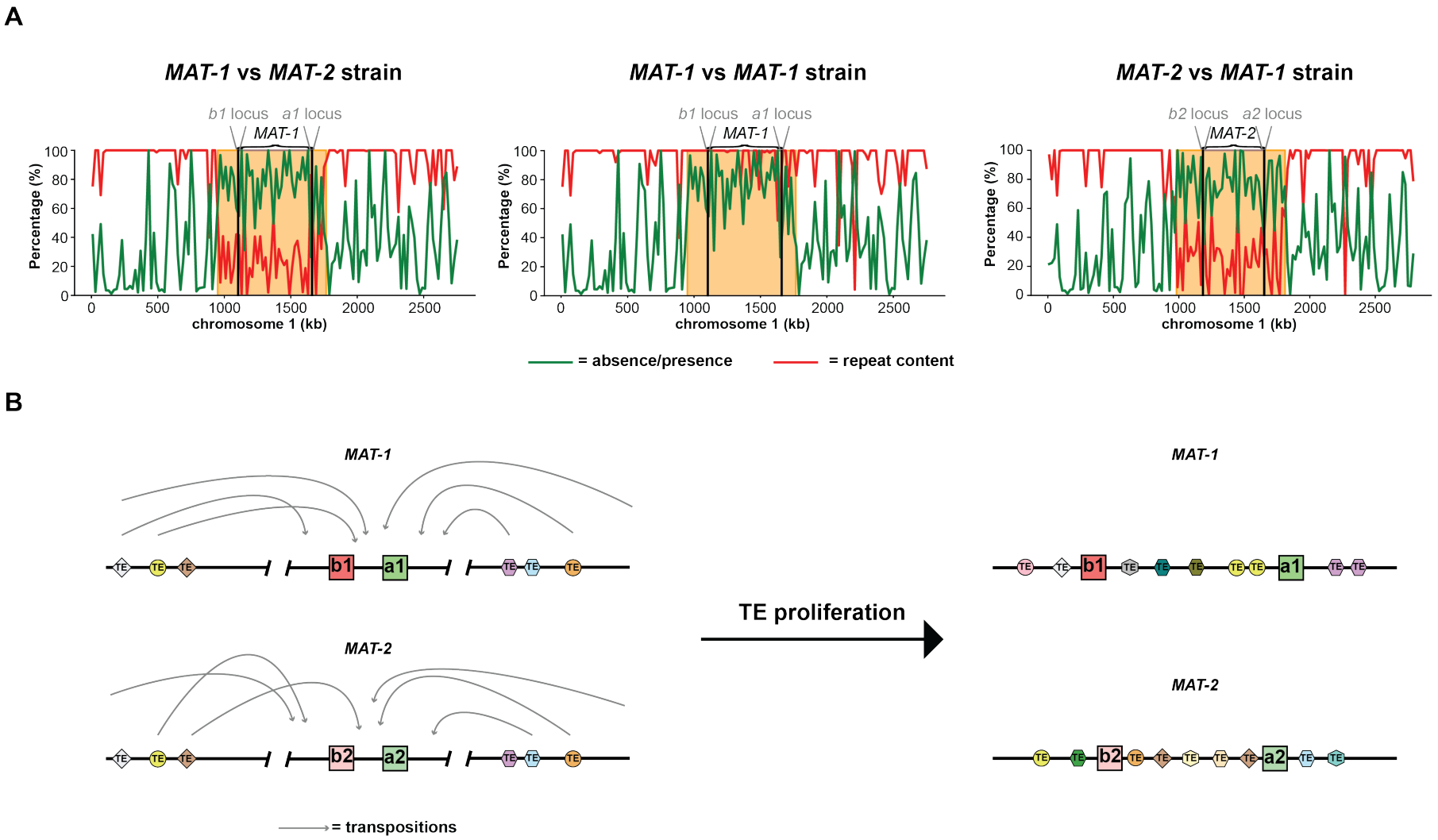

Figure 3: The chromosome 1 architecture of *MAT-1* and *MAT-2 Ustilago hordei*. (A) As reference, the *MAT-1* strain Uh805 and the *MAT-2* strain Uh1278 were used. Repeat content and percentage presence/absence polymorphisms were calculated for 20 kb windows for the reference strains. The presence-absence was determined between the *MAT-1* and *MAT-2* reference strains, in addition to between the *MAT-1* strains Uh805 and Uh811. The *MAT*, *a* and *b* loci are indicated. The orange square includes the mating-type loci and indicates the genome region where the repeat content is high and the sequences are generally mating-type species. (**B**) Model that explains the mating-type specificity of the sequences within and flanking the mating-type loci. The absence of recombination within the mating-type loci and its flanking regions makes that transpositions within these regions are mating-type specific.

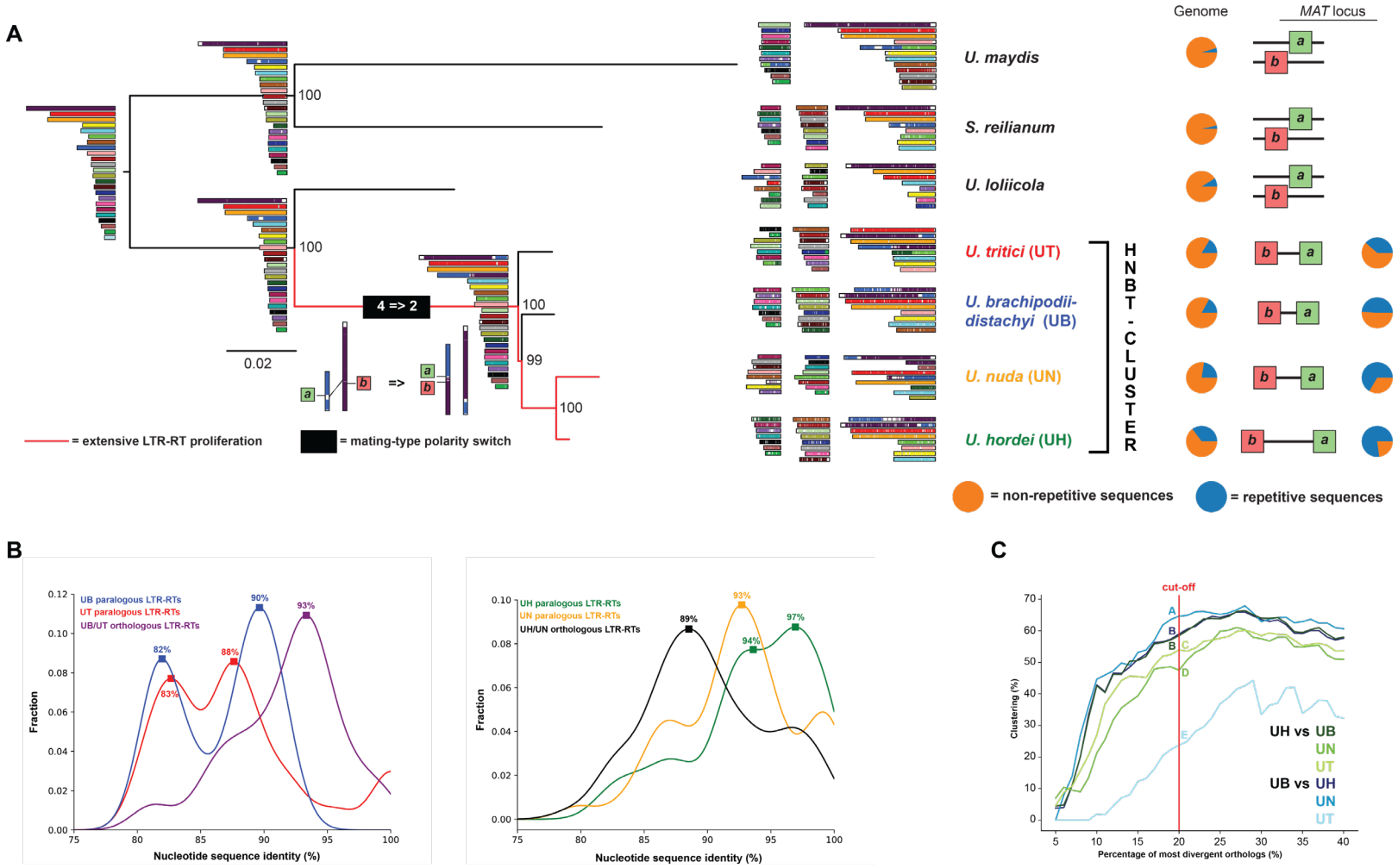

Figure 4: Genome evolution of the HNBT-cluster species. (**A**) Phylogenetic relationship between smut pathogens based on Benchmarking Universal Single-Copy Orthologs (BUSCOs). Phylogenetic relationship between newly and previously sequences smut species with the *Ustilago maydis*/*Sporisorium reilianum* branch was 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-types (**B**) Long terminal repeat retrotransposon (LTR-RT) expansion in the HNBT-cluster. Nucleotide sequence identity distribution of best reciprocal paralogous and orthologous LTR-RT sequences. Squares on the lines display maxima with the corresponding sequence identity value. (**C**) Extent of clustering of most diverged orthologs between different species. *Ustilago hordei* and *Ustilago brachipodii-distachyi* were used as reference genomes to calculated the clustering of the x% most divergent (y-axis) HNBT-cluster orthologs. The degree of clustering is defined as the fraction of the x% most diverged genes that reside with at least 10 together in a window of 20 genes. Significant differences were determined for 20% of the most divergent orthologs with the Fisher exact test for *p*-value < 0.05 and indicated with a capital letter (A-E).

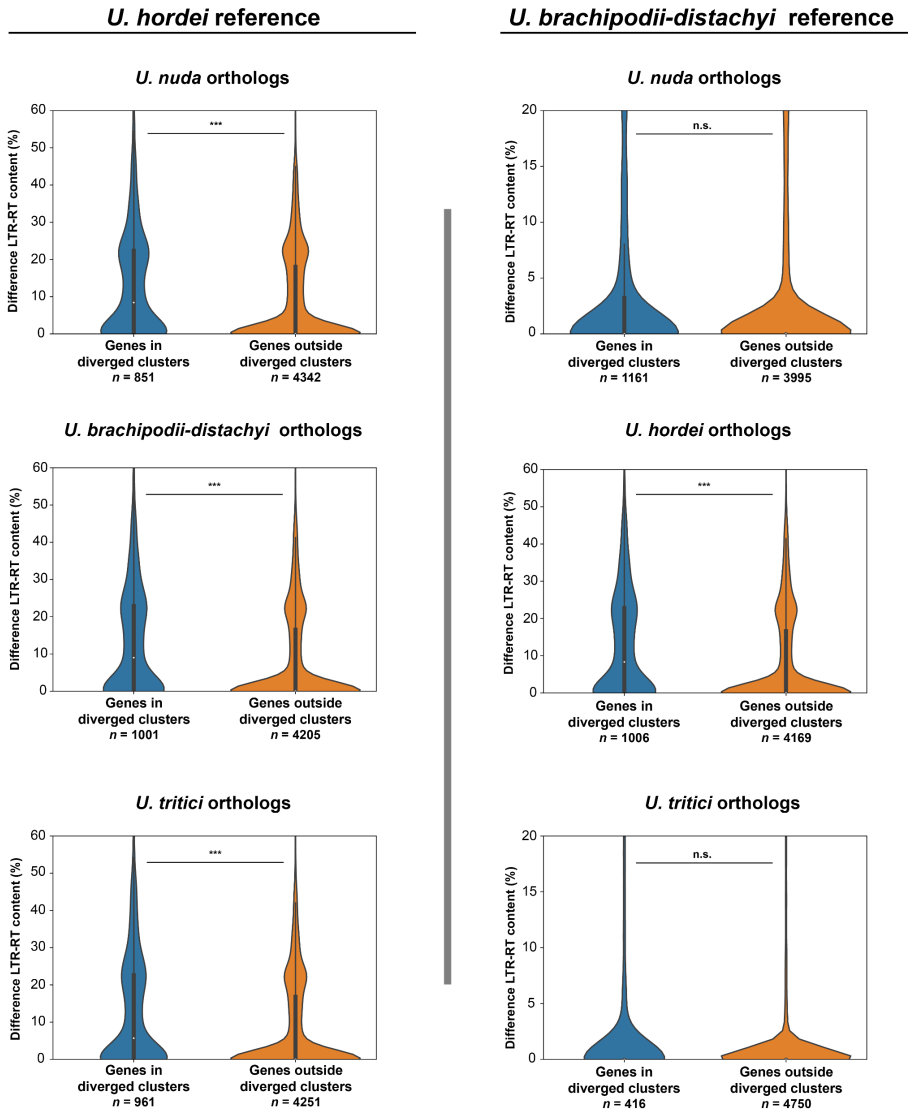

Figure 5: Interspecific long terminal repeat retrotransposon (LTR-RT) alterations in windows around genes within and outside Diverged Gene Clusters (DGCs). *Ustilago hordei* and *Ustilago brachipodii-distachyi* were used as reference genomes and DGCs were determined based on the 20% most diverged genes compared to orthologs from the other HNBT-cluster species. Absolute differences in LTR-RT content between orthologs were determined for a window 10 kb upstream of the 5'-end and 10kb downstream of the 3'-end of the genes. Significant differences were calculated with the Welch's t-test. *** : p-value < 0.001 and n.s. : no significant difference.

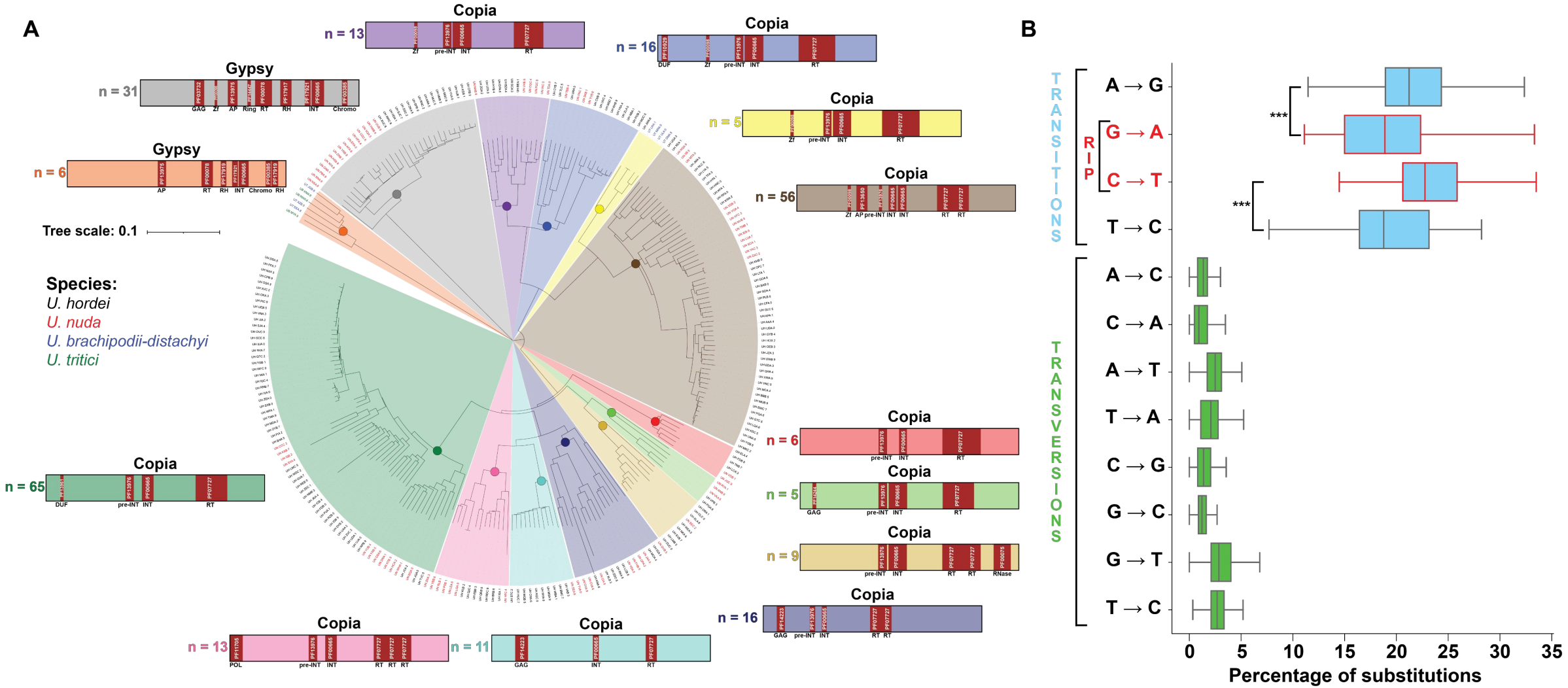

Figure 6: Long terminal repeat retrotransposon (LTR-RT) substitutions. (**A**) In total, 252 LTR-RTs are included in the phylogenetic tree and their species of origin is indicated by their colour name. LTR-RTs families are indicated by the circle segments in different colour. For every family, an ancestor sequence was reconstructed and indicated on the phylogenetic tree by a circle. The superfamily and the functional domains are indicated for the ancestor sequences. (**B**) The substitutions encountered by LTR-RT since their reconstructed ancestor. For transitions, differences between in substitution frequencies were tested with a t-test. *** : p-value < 0.001.

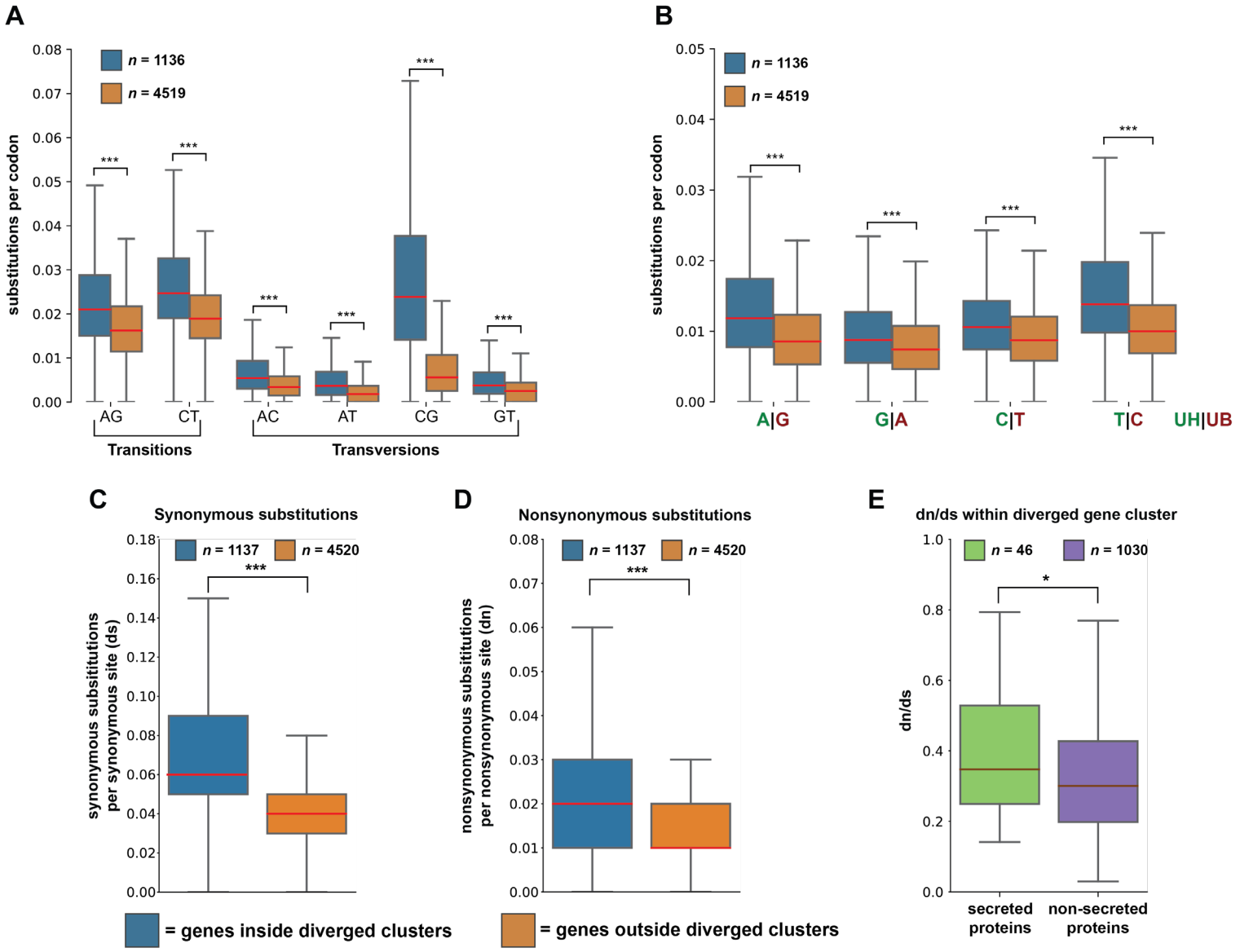

Figure 7: Properties of Diverged Gene Clusters (DGCs). Difference if frequencies of (**A**|**B**) substitution type, (**C**) synonymous and (**D**) nonsynonymous substitutions between genes inside and outside DGCs. (**E**) Difference in fraction of relative nonsynonymous and synonymous levels of genes within DGCs that encode and do not encode a secreted protein. DGCs were defined for *U. hordei* by using the 20% most diverged orthologs with *U. brachipodii-distachyi*. Outliers were not depicted for the boxplots. Significant differences were indicated with "*" for *p*-value < 0.05 and "***" for *p*-value < 0.001. Significant differences were calculated with the Welch's t-test.

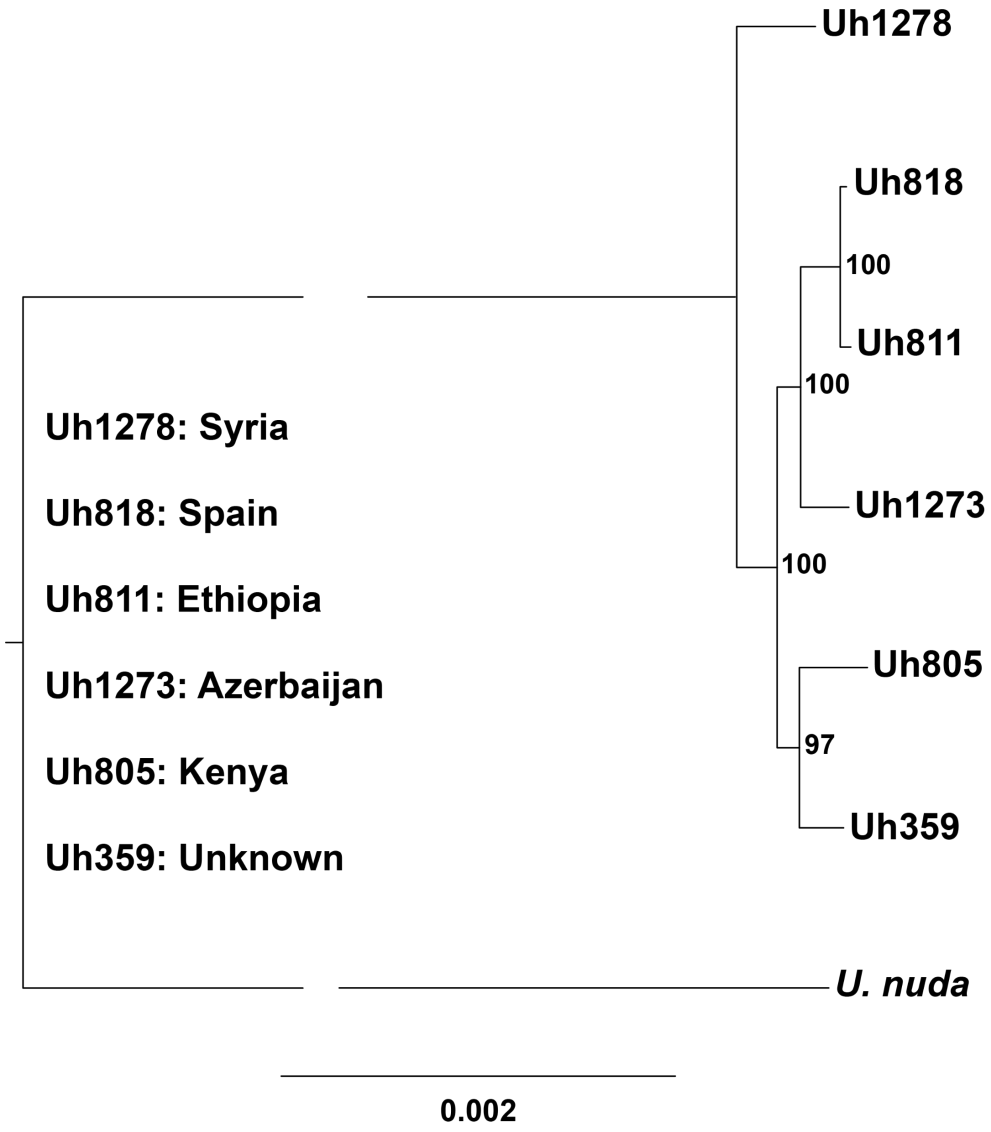

Figure S1 Phylogenetic relationship between *Ustilago hordei* lineages based on Benchmarking Universal Single-Copy Orthologs (BUSCOs). Phylogenetic relationship between the *U. hordei* strains with *U. nuda* as an outgroup. In total, 1,692 BUSCOs were used for tree construction. The robustness of the trees was assessed using 100 bootstrap replicates.

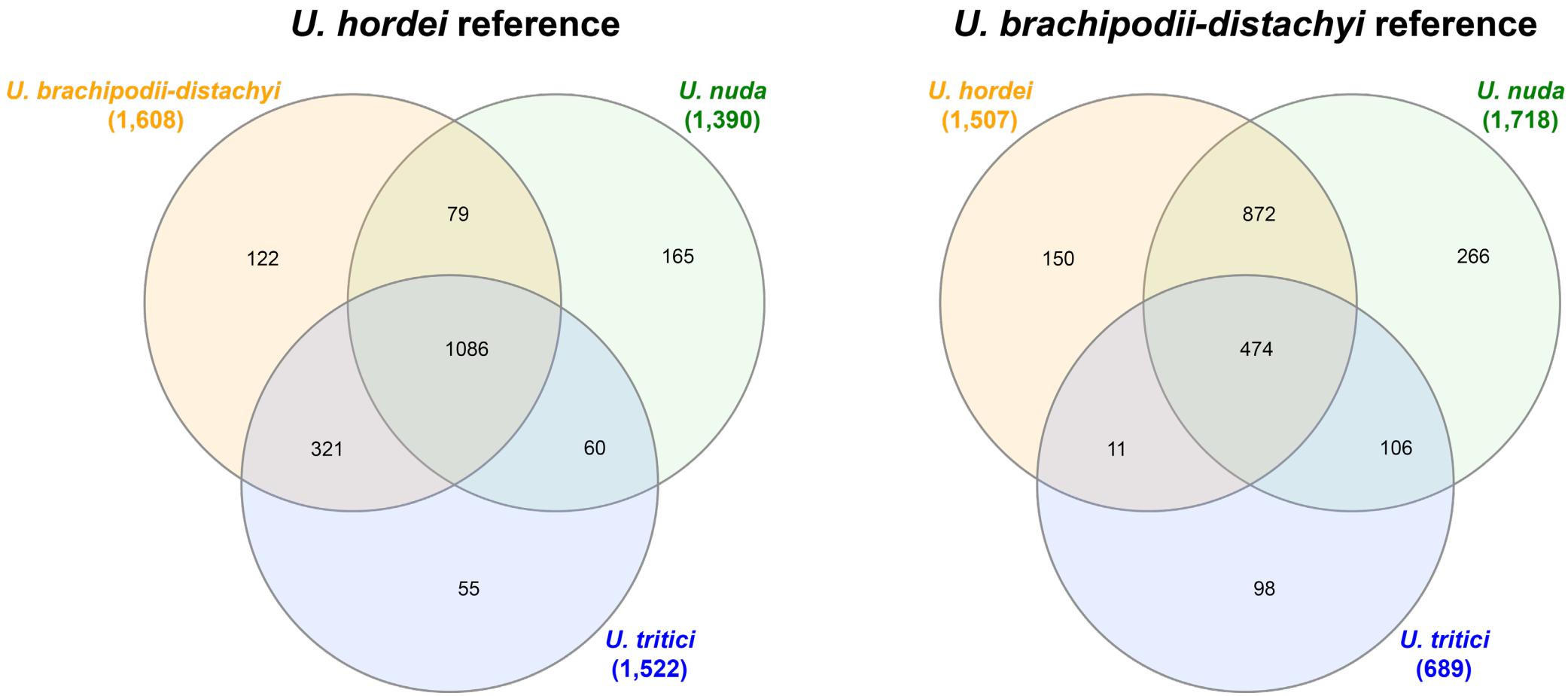

Figure S2: Overlap Diverged Gene Cluster (DGC) genes using orthologs of different species. *Ustilago hordei* and *Ustilago brachipodii-distachyi* were used as reference genomes to determine the clustering of the 20% most diverged genes using orthologs of different species.

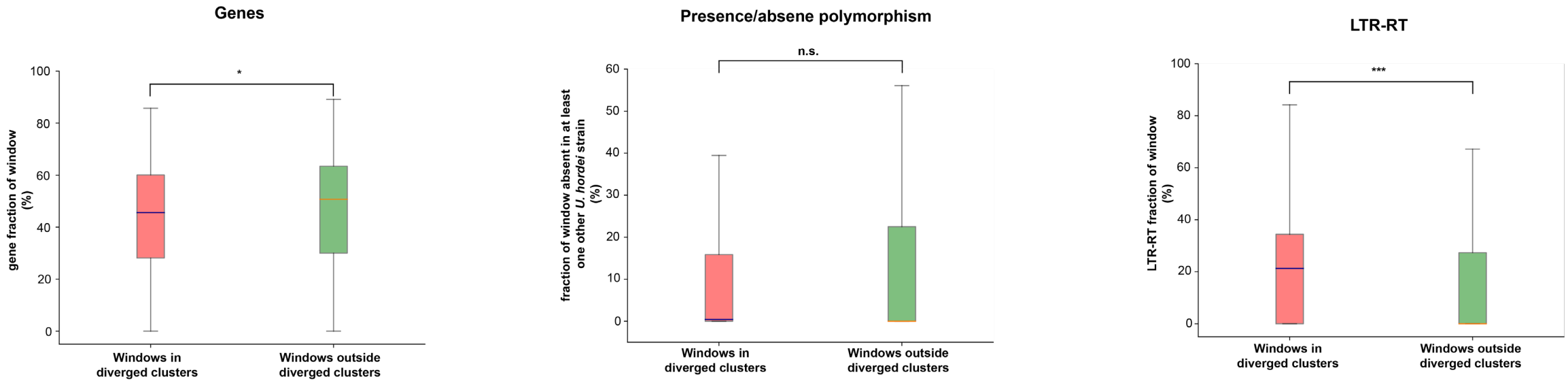

Figure S3: Gene density, presence/absence polymorphisms and long terminal repeat retrotransposons (LTR-RTs) of Diverged Gene Clusters (DGCs). DGCs were determined based on the 20% most diverged *Ustilago hordei* genes with *Ustilago brachipodii-distachyi* orthologs. The various features were calculated over 20kb windows. Outliers were not depicted for the boxplots. Significant differences were calculated with the t-test. * : p-value < 0.05 , *** : p-value < 0.001 and n.s. : no significant difference.