Oggenfuss *et al. Mobile DNA (2024) 15:2*
https://doi.org/10.1186/s13100-024-00312-1

Mobile DNA

RESEARCH

Open Access

A systematic screen for co-option of transposable elements across the fungal kingdom

Ursula Oggenfuss[1,2], Thomas Badet[1] and Daniel Croll[1*]

Abstract

How novel protein functions are acquired is a central question in molecular biology. Key paths to novelty include gene duplications, recombination or horizontal acquisition. Transposable elements (TEs) are increasingly recognized as a major source of novel domain-encoding sequences. However, the impact of TE coding sequences on the evolution of the proteome remains understudied. Here, we analyzed 1237 genomes spanning the phylogenetic breadth of the fungal kingdom. We scanned proteomes for evidence of co-occurrence of TE-derived domains along with other conventional protein functional domains. We detected more than 13,000 predicted proteins containing potentially TE-derived domain, of which 825 were identified in more than five genomes, indicating that many host-TE fusions may have persisted over long evolutionary time scales. We used the phylogenetic context to identify the origin and retention of individual TE-derived domains. The most common TE-derived domains are helicases derived from *Academ*, *Kolobok* or *Helitron*. We found putative TE co-options at a higher rate in genomes of the Saccharomycotina, providing an unexpected source of protein novelty in these generally TE depleted genomes. We investigated in detail a candidate host-TE fusion with a heterochromatic transcriptional silencing function that may play a role in TE and gene regulation in ascomycetes. The affected gene underwent multiple full or partial losses within the phylum. Overall, our work establishes a kingdom-wide view of putative host-TE fusions and facilitates systematic investigations of candidate fusion proteins.

Introduction

Proteomes are diverse and variability extends to the population and individual level [1]. Causes of proteome diversity include alternative splicing, presence-absence polymorphisms, single nucleotide polymorphisms or larger structural variations, such as duplications, reshuffling of protein domains, partial deletions or translocations [2]. Aneuploidy or gene duplication, followed by neofunctionalization due to relaxed purifying selection, can lead to diversification [3]. Gene gains can also be mediated by horizontal gene transfer from other species or by de novo gene birth [4, 5]. Proteome evolution can also entail pseudogenization, with pseudogenes expected to eventually get lost or regain function. Genetic variation can provide a highly dynamic proteome, allowing populations to rapidly adapt to new or changing environments.

Mutations, rearrangements, losses or acquisitions of protein-coding genes may be facilitated by co-localization with transposable elements (TEs). In some species, TEs are clustered into gene-poor islands [6, 7]. TE rich islands are under relaxed purifying selection, often leading to retention of single nucleotide polymorphisms or structural variants, and a higher rate of ectopic

*Correspondence:
Daniel Croll
daniel.croll@unine.ch
[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland
[2] Department of Microbiology and Immunology, University of Minnesota, Medical School, Minneapolis, Minnesota, United States of America

recombination caused by repetitive sequences [8]. Genes located in such TE islands are thought to be under purifying selection as well, allowing the accumulation of mutations at a fast rate [9]. Ectopic recombination of two copies of the same TE family can lead to the deletion of genes between them [10]. Some TEs such as *Helitrons* in maize or *Pack-Mules* in rice have the ability to capture and amplify segments containing coding sequences [11–17]. Finally, TEs can capture genes and horizontally transfer them to other species, including by the means of *Starships* in fungi or virus-like *Mavericks* in nematodes [18–23]. Exon-shuffling via the activity of TEs can lead to novel transcription factor binding sites, providing novel regulatory dynamics and ultimately new functions to proteins [24].

TE insertions into coding regions are typically deleterious and are therefore under strong purifying selection [25, 26]. Yet, TE insertions into duplicated genes, pseudogenes or non-essential genes are less likely to be deleterious and may lead to neofunctionalization or exonisation of the gene [4]. If retained over time, such host-TE fusions may lose functions related to TE proliferation and become essential, a process also identified as TE domestication or co-option [24, 27]. Host-TE fusions that provide essential new functions are expected to be retained, although the evolutionary timeframes of such domestication events remain poorly understood.

The initial function of TE encoding sequences is typically restricted to a few functions related to the mobilization and duplication of the elements [28, 29]. Yet, how TE sequences provide additional functions for existing coding sequences remains understudied. A well-studied example of host-TE fusion is the V(D) J recombination that leads to immunoglobin diversification and provides highly conserved adaptive immunity in jawed vertebrates [30, 31]. The recombination activating genes *RAG1* and *RAG2* retained mobility and can re-shuffle recombination signal sequences, creating the basis for rapid sequences changes in the face of new antigens [31]. Even though the V(D) J recombination is not conserved across all vertebrates, the fusion is thought to have occurred ~ 500 million years ago [31, 32]. *RAG1* is a host-TE fusion gene, containing the transposase of the *Transib*-like DNA transposon and a RING finger ubiquitin ligase at the N-terminal that probably acts in dimerization and as a ligase for ubiquitination [33]. Another example is *KRABINER*, a host-TE fusion in vespertilionid bats consisting of a *Mariner* DNA transposon and *ZNF112* [34]. *KRABINER* controls the regulation of a large network of genes [34]. In the fission yeast *Schizosaccharomyces pombe*, *Abp1*, *Cbh1*, *Cbh2* are centromeric *pogo* derived host-TE fusions that led to retrotransposon silencing [35–37]. A *Bel-Pao* derived *gag* sequence was recently shown to have fused with *PEX14* gene, acquiring an intron and creating a host-TE fusion in fungi [38].

TEs are highly diverse in fungal genomes, even between closely related species, indicating independent TE activity [39–41]. TEs have played an important role in the evolution of host-associated lifestyles or local adaptation to external stress including tolerance of pesticides [42–45]. Many fungal species show distinct genome compartmentalization, featuring TE-rich and gene-poor islands, and a fungal specific defense against repetitive sequences further increases the differentiation [9, 46–48]. Fungi associated with animals and pathogenic lifestyles in general tend to have higher numbers of TE insertions into genes, which could either be recent insertions in non-essential genes or host-TE fusions [49]. Old TE insertions are more likely to affect genes with enzymatic rather than protein-protein interaction functions [49]. The TE content and diversity observed today may not necessarily correlate with the number of host-TE fusions, as TE activity is expected to occur in random or stress-induced bursts of proliferation [50]. Ancient and ongoing TE activity in many lineages of the fungal kingdom and the exceptional genomic resources available for such compact genomes provide a vast potential to retrace the emergence of host-TE fusions over deep evolutionary timeframes.

Here, we used a systematic approach to detect host-TE fusions in the genomes of 1237 fungal isolates. To identify host-TE fusions, we used gene orthology and phylogenomic analyses to detect the emergence and retention of TE-derived domains in fungal proteomes. We found that TE-derived helicases are the dominant TE partner in likely host-TE fusions. The subphylum Saccharomycotina, which includes model yeasts like *Saccharomyces cerevisiae* and *Candida albicans*, shows elevated contents of host-TE fusions despite typically having small and repeat-poor genomes. Host-TE fusions are enriched for binding functions to heterocyclic compounds, organic cyclic compounds ATP, adenyl ribonuclease and adenyl nucleotide. Additionally, we identified widespread candidate host-TE fusions in ascomycetes involved in gene silencing, originating from *Helitron*, *AcademH* or *Kolobok* and *Maverick* domains. Phylogenetic analyses suggest independent origins of identical host-TE fusions, uneven rates of gene retention and secondary losses.

Methods

Retrieval of genomes and gene annotations

We obtained genomes and gene annotations for 1237 fungal isolates from two different sources. A total of 994 genomes belong to the phylum Ascomycota, 195 Basidiomycota, 28 Mucoromycota, 12 Chytridiomycota, 8 Zoopagomycota (see Supplementary Table S1 for

full references and additional data). The budding yeast genomes were retrieved from Shen and colleagues [51]. We retrieved additional genomes and gene annotation from fungal and Oomycetes genomes from NCBI. Sixteen oomycetes were used as outgroup to root the phylogenetic trees in downstream analyses (Supplementary Table S1).

Phylogenomic reconstruction

To build a tree, we followed the approach by Li et al. [52] to reconstruct the fungal tree of life. Briefly, we first identified a set of single-copy orthologous genes in each of the 1237 genomes using BUSCO v 4.1.4 searching the fungi or oomycote orthology database version 10 for fungi and oomycetes, respectively [53]. The pipeline identified a maximum set of 756 BUSCO genes in the genome of the fungus *Colletotrichum plurivorum*. The identified BUSCO genes were then translated into protein sequences respecting the relevant genetic code (code 12 for Saccharomycotina isolates except for *Pachysolen tannophilus* (Pactanno for which code 26 was used, and code 1 for all other genomes) [54]. Of the 756 BUSCO genes identified, a random sample of 100 of the resulting BUSCO protein sequences was then concatenated using the geneStitcher.py script (https://github.com/ballesterus/Utensils) and aligned using mafft v 7.475 with the parameters *--maxiterate 1000 --auto* [55]. The resulting alignment was then trimmed using trimAl v 1.4.rev15 with the *-gappyout* option [56]. We estimated the best-fitting evolutionary models for the concatenated 100 protein sequences using partitionfinder v 2 with the quick option *-q* and default RAxML v 8.2.12 [57, 58]. The resulting partitioned model was then applied for phylogenetic inference using iqtree2 v 2.1.2 after 1000 replicates for ultrafast bootstrap and 2 independent runs with *-B 1000 --runs 2* [59]. We rooted the tree with the non-fungal oomycete *Phytophthora parasitica* and visualized the tree using the R packages ggtree, ggtreeExtra and treeio [60–62].

Annotation of functional domains in the proteomes

To identify putative functional domains across the analyzed proteomes, we downloaded the annotated domains hidden Markov models from the PFAM release 31 [63]. We used the hmmsearch function from the HMMER package v 3.3.2 to scan all proteomes for functional domains with the *--noali* option to speed up the process [64]. We then filtered the matching domains for a minimal bitscore of 50 and a maximal e-value of 1e-17 using the HmmPy.py script (https://github.com/EnzoAndree/HmmPy).

Inference of trophic modes

We categorized genomes using the CATAStrophy v 0.1.0 pipeline [65]. Using the predicted proteins from all genomes, we searched for genes encoding carbohydrate-degrading enzymes (CAZymes) with dbCAN v 8 [66]. As for the PFAM annotation, we performed hmmscans on each proteome using the dbCAN hidden Markov models as query. We then applied the CATAStrophy algorithm to predict the most likely trophic mode based on the set of encoded CAZymes.

Gene orthology analysis

We inferred gene orthology among all genomes based on protein sequence identity. We used Orthofinder v 2.4.1, which implements diamond blast v 0.9.24 for homology searches across the pool of predicted proteins [67, 68]. From the initial set of 13,863,658 individual proteins encoded by all genomes combined, Orthofinder grouped 7,860,083 proteins into 299,713 orthogroups.

Detection of candidate host-TE fusions

We retrieved previously reported PFAM domains associated with fungal TE superfamilies ([49], https://www.mrc-lmb.cam.ac.uk/genomes/boris/12genomes/summary_for_CB) and filtered for genes encoding TE-associated PFAM domains. In a second filtering step, we removed proteins annotated exclusively with TE-associated PFAM domains. We excluded PFAM with similarity to any of the fungal TE PFAM based on SCOOP and HHSearch [69] (Supplementary Table S2). We removed all oomycete genes. We filtered out genes if the identified TE and non-TE PFAM domains had an overlap of more than 5% in the amino acid sequence. Such overlaps may indicate that the two annotations identify the same protein domain. Overlaps were identified using bedtools v 2.30.0 with the *intersect* function [70]. We retained candidate orthogroups including host-TE fusion proteins if genes encoding independent TE and non-TE PFAM domains were represented in at least five genomes and belong to the same orthogroup (Fig. 2A).

Indication of repeat-induced point mutations

Given that host-TE fusions likely emerge after gene duplication, and gene duplication is reduced in many Ascomycete species due to repeat-induced point mutations (RIP), we compared the number of host-TE fusion candidates to the percentage of RIP affected regions of a subset of 48 genomes, previously reported [48].

Gene ontology term enrichment analyses

We analyzed the enrichment of specific gene ontology terms among host-TE fusion genes compared to

the background of all genes. To reduce the computational load, we defined the background as a 1% random subset of the entire set of genes (subset: $n = 358{,}350$). Gene ontology terms were assigned to genes using a GO-PFAM term translation based on Mitchell et al. [71]. We created a GOAllFrame object with the AnnotationDbi package v 1.54.1 and constructed a GeneSetCollection with GSEABase v 1.54.0 [72, 73]. We calculated enrichment *p*-values using the *hyperGTest* in the Category package v 2.58.0 [74]. For each MF (molecular function), BP (biological process) and CC (cellular component) Gene ontology term enrichment, a *p*-value cut-off of 1e-10 and a minimum term size of 20 was applied.

Filtering for copy-number variation in host-TE fusion genes

To detect potential activity of the TEs represented by the identified PFAM domains in individual genomes, we analyzed potential copy-number variation of the host-TE fusion genes and their respective PFAM terms. To reduce conservatively detecting host-TE fusion genes generated by recent TE insertion events, we required host-TE fusion genes to be present in at least 5 genomes belonging to the same order and present in at least 20 genomes. Furthermore, we analyzed candidate host-TE fusion genes for their PFAM domain order along the amino acid sequence and removed orthogroups without a conserved domain order. After filtering, we extracted the predicted function of the non-TE candidate function based on the information provided in the PFAM database [63]. To remove host-TE fusion gene candidates potentially erroneously identified due to the physical proximity of genes in fungal gene clusters, we performed gene cluster analyses using antiSMASH v 3.0 using the list of predicted gene clusters from Kautsar et al. [75, 76]. Finally, we generated a subset of host-TE fusion gene candidates based on manual filtering for known TE domains. We performed this additional filtering starting from host-TE fusion candidates present in ≥5 genomes as described above. We manually removed weakly supported candidates presenting either a candidate TE domain that is not unambiguously recognized as TE-derived or a candidate host domain that may potentially be of TE origin (Supplementary Table S6).

Phylogeny of helicase-related protein families

To infer the evolutionary relationship between proteins that encode a helicase conserved C terminal domain, we first recovered all proteins that harbor such PF00271 domains with a minimal bitscore of 50 and a maximal e-value of 1e-17 after hmmsearch. Using the samtools faidx function, we recovered all protein sequences corresponding to the PF00271 domain. We removed proteins lacking methionine at the start or containing in-frame stop codons using a custom script (https://github.com/milesroberts-123/extract-weird-proteins). The PF00271 encoding protein sequences were aligned using Clustal Omega allowing for 5 iterations and trimmed using trimAl with the gappyout method [77, 78]. Sequences with more than 20% gaps over the trimmed alignment were removed using fasta_drop.py (https://www.biostars.org/p/9512948/). A phylogenetic tree was finally inferred using FastTree version 2.1.11 with the Whelan-And-Goldman 2001 model (wag) and 1000 bootstraps [79]. Clusters based on the tree were computed using the TreeCluster version 1.0.4 and the median pairwise distance method (med_clade) for threshold values ranging from 1 to 2.5 [80]. The tree was visualized with the R package ggtree.

Results

Phylogeny and genomic landscape show variation among genomes in the fungal kingdom

We analyzed genomes of 1237 fungal isolates belonging primarily to phyla of ascomycetes and basidiomycetes (Fig. 1A). Based on a set of 100 single-copy genes we constructed a maximum likelihood phylogenetic tree (Fig. 1B; Supplementary file F1). The tree resolves the fungal phylogeny consistently with recent analyses of similar scope [81, 82]. Ascomycetes are segregated into three larger groups including the Saccharomycotina, Taphrinomycotina and Pezizomycotina. The analyzed genomes are generally of high completeness based on BUSCO analyses with a mean number of complete genes of 94.97%, and with 95.71% higher than 80% (Fig. 1C). The number of detected genes varied from 602 to 22,164, with generally lower gene numbers in the Saccharomycotina (Fig. 1B). Assembled genome sizes were highly variable and ranged between 7.37–773.10 Mb (mean = 34.43 Mb; Fig. 1C). Genome-wide GC content is on average 46.4% with an observed range between 16.3–67.8% (Fig. 1C). Genomes in Saccharomycotina, Chytridiomycota, Mucoromycota and Zoopagomycota typically exhibit GC contents below 50%.

The number of host-TE fusions across the fungal kingdom is highly variable

We next analyzed coding sequences for conserved domains present in the PFAM database. To define candidate host-TE fusion we required that at least one conserved domain matches a domain thought to be exclusively associated with TEs and at least one domain not associated with TEs. The stringent filtering, which required candidates to be detected in at least 20 genomes, allowed us to focus the analyses on conserved host-TE fusions over deep evolutionary times and to exclude pseudogenes. From a total of 39,655 unique

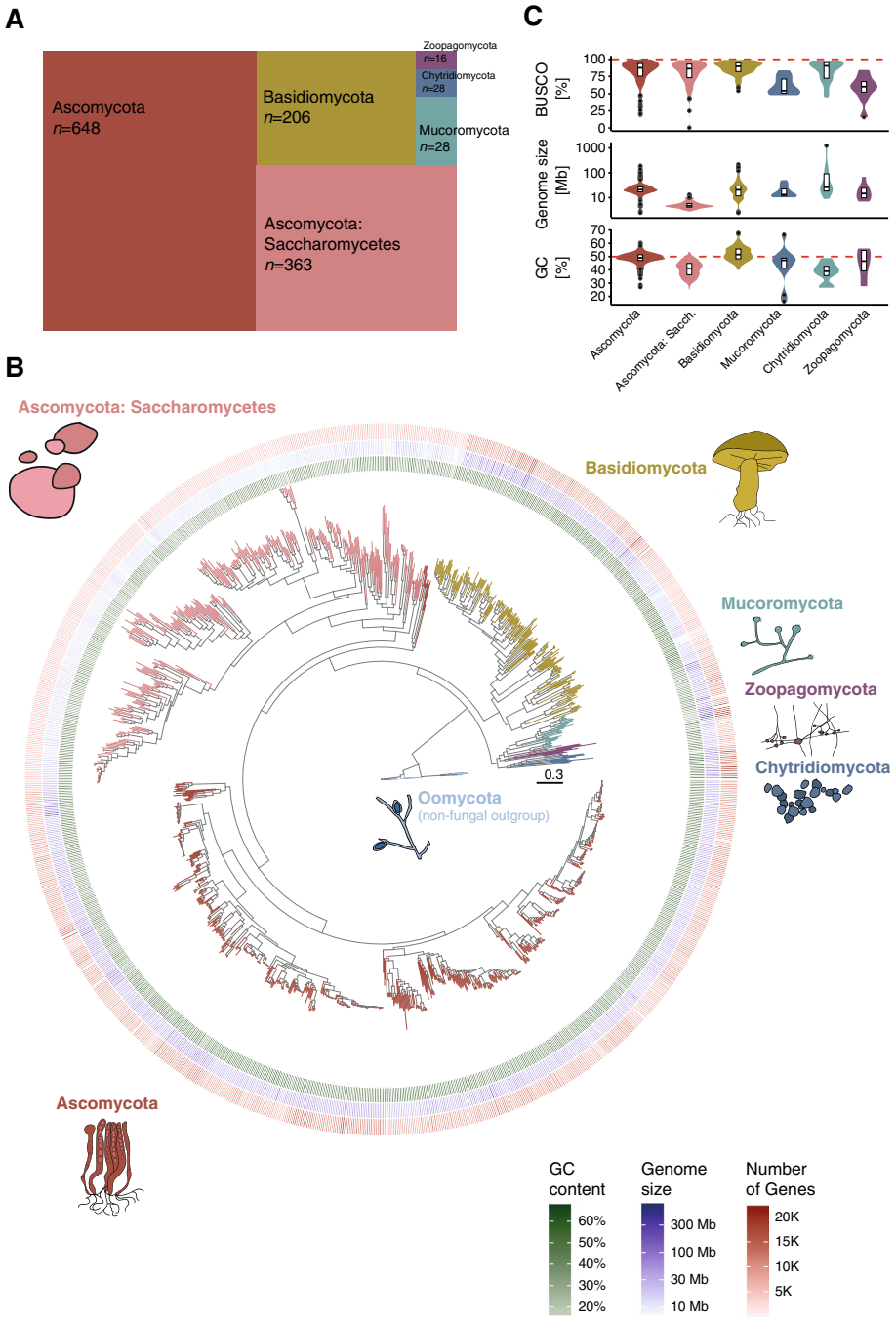

Fig. 1 Phylogenomic analysis and genome properties across the fungal kingdom and oomycetes: **A** Number of species per phylum analyzed. The subphylum of Saccharomycotina is shown separately from the other ascomycetes. **B** Phylogenomic tree of the fungal kingdom based on 100 concatenated orthologous protein alignments. Oomycetes were defined as the outgroup. From inside to outside: genome-wide GC content (green), total genome size (blue) and number of annotated genes (red). The tree is missing the phyla of Blastocladiomycota, Cryptomycota and Microsporidia. **C** Distribution of gene completeness score (BUSCO), genome size and genome-wide GC content

proteins associated with a TE-associated domain across all genomes, we found 13,342 to also contain a non-TE domain (Fig. 2A; Supplementary Table S3). A total of 1205 genomes (98.3%) carry at least one gene matching our criteria for host-TE fusion genes. We found on average 297.5 host-TE fusions (range 0–3311) per genome (Fig. 2B). Overall, 0.6–17.0% of all annotated genes of a genome are host-TE fusions. Genomes belonging to

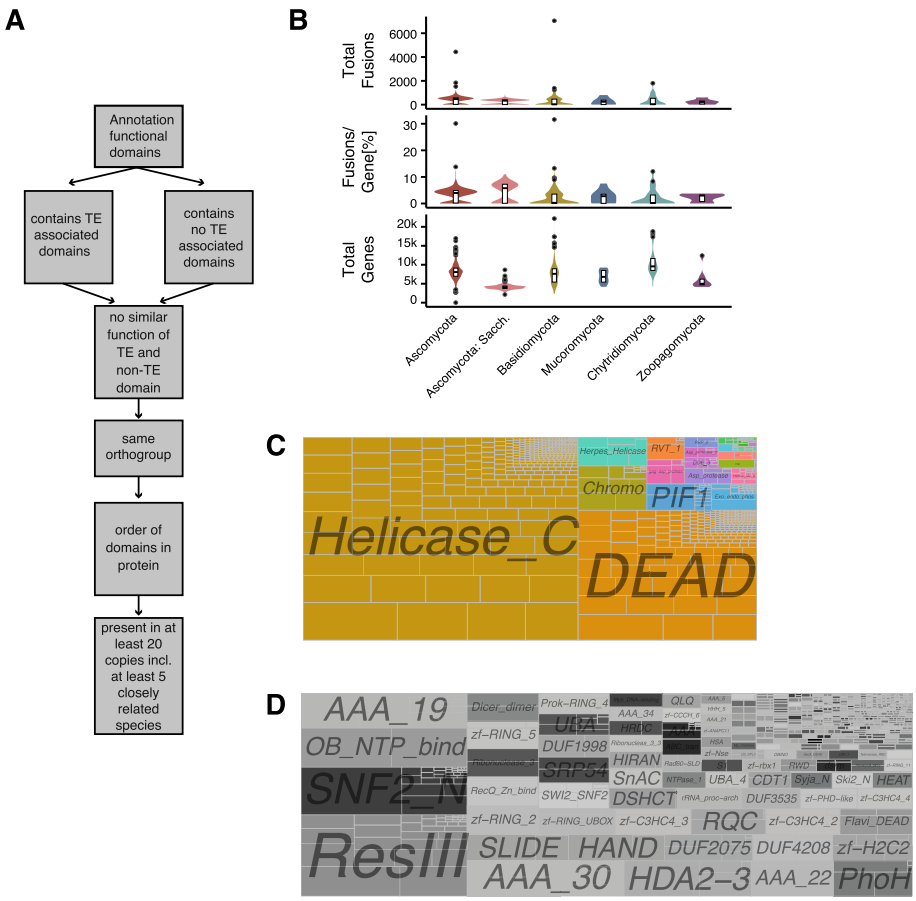

Fig. 2 Host-TE fusion events identified across the fungal kingdom: **A** Overview of host-TE fusion detection steps. **B** Number of fusions detected per species, number of fusions detected per gene and number of annotated genes per genome. **B** Function of TE derived domains in all detected host-TE fusions. Squares indicate the number of individual fusions. **D** Function of non-TE derived domains of host-TE fusion genes

Saccharomycotina, on average, have a higher proportion of host-TE genes per genome (240 compared to 372 across all other genomes; Fig. 2B), even though they generally contain fewer genes compared to other ascomycetes (5837 compared to 12,502; Fig. 2B). Generally, Saccharomycotina have a higher percentage of genes that are containing a TE derived sequence (4.18% compared to 3.04%; Fig. 2B). Outlier species with high proportions of candidate host-TE derived proteins include the plant pathogens *Armillaria ostoyae* (31.6%) and *Fusarium poae* (30.1%), as well as nine additional genomes with a proportion > 10% (Supplementary Fig. S1A). The two highlighted species carry moderate proportions of TEs in the genome compared to other fungi [83, 84]. *Fusarium* species have previously been shown to encode a large number of diverse Helitrons in their genome, which would explain the high number of potential host-TE fusions in these species [83]. We found no correlation between the number of detected host-TE fusions, BUSCO completeness scores, GC content or genome size suggesting the variation in host-TE among lineages is not meaningfully explained by variation in genome assembly quality (Supplementary Fig. S1B). Repeat-induced point mutations (RIP) may impact the ability to retain duplicated sequences and early-stage host-TE fusions in particular. In a subset of genomes that were previously analyzed on the strength of RIP, we detected no indication of host-TE fusions in genomes covered by more than 10% of RIP affected regions (Supplementary Fig. S1C). Genomes with lower coverage of RIP affected regions vary between 0 and 10% of predicted proteins that are part of a host-TE fusion.

Transposable elements provide DNA binding sites to a wide range of functions

We restricted our analysis to 824 (115,497 occurrences) host-TE fusion orthogroups where an ortholog is present in at least five isolates, thereby retaining the evolutionarily conserved host-TE fusions. From the set of 824 individual host-TE fusions, we identified 29 distinct

TE-associated PFAM (Supplementary Table S4). The TE-related PFAM only includes domains of *Helicase_C* (PF00271) *DEAD* (PF00270) helicases, *PIF1* (PF05970) and *Helitron_like_N* (PF14214) from the *AcademH*, *KolobokH* or *Helitron* TE superfamilies (Fig. 2C). Non-Helicase domains with more than 1000 candidates include *Chromo* domains (CHRromatin Organization Modifier; PF00385) from the *Maverick* TE superfamily and specific retrotransposons, *Exo_endo_phos* (Endonuclease / Exonuclease / phosphatase; PF03372) from the LINE TE order and *DDE_1* (DDE superfamily endonuclease; PF03184) from *Tc1–Mariner* TE superfamily. Importantly, *Helicase* domains, *Chromo* domains, *Exo_endo_phos* are not exclusively TE-derived.

The diversity in non-TE PFAM domains consistently found across all orthologs of a host-TE fusion protein is substantially higher with 383 individual non-TE PFAMs. In particular, the domains included *ResIII*, *SNF2_N*, *OB_NTP_bind* and *AAA_19* functions (Fig. 2D). The 383 non-TE PFAM are associated with 66 gene ontology terms, with highest associations in ATP binding, nucleosome-dependent ATP activity, nucleic acid binding, protein binding, methyltransferase activity, GTP binding, nucleus, zinc ionic binding, RNA processing and hydrolase activity, acting on acid anhydrides, in phosphorus containing anhydrides being the prevalent functions (Supplementary Fig. 2A). Among our list of candidates, we find the centromere protein *CENP-B* that is known to have originated from a *pogo*-like transposase domestication event in yeast (also known as *Abp1*, *Cbh1* and *Cbh2* centromere protein N-domain in *S. pombe*; PF03184 and PF18107) [35–37, 85]. More than 40% of all non-TE domains could not be associated with a gene ontology term.

We then focused on a more restricted set of candidates including the most evolutionarily conserved host-TE fusions by requiring an ortholog to be present in at least 20 genomes (and at least five species belonging to the same order). The resulting subset of host-TE fusions contains predominantly TE-associated functions related to helicases (*Helicase_C*, *DEAD* helicase) and 241 genes encoding 125 distinct non-TE PFAM domains, including *SNF2_N*, *UBA*, *zf-H2C2*, *Rad60-SLD* and *UBA_4* (Supplementary Fig. S2B). Domains with functions related to nucleotide binding are enriched in this set of 241 candidates (Fig. 3; Supplementary Table S5). We also identified a fusion between a *DEAD* helicase and a *Dicer_dimerization* domain (PF00270 and PF03368). The Dicer protein is involved in RNA interference and protection against TE activity or viral infection and has been previously identified as containing a helicase domain [86].

We highlight the detection of a previously described *S. pombe* host-TE fusion with homology to the Mit1 domain [87]. The Mit1 domain is a component of an effector complex for heterochromatic transcriptional silencing (SHREC) with a function in heterochromatin silencing (PF00271 and PF00385) [85]. SHREC is a host-TE fusion that includes a *Helicase_C* derived from *AcademH*, *KolobokH* or *Helitron*, an additional TE-derived *Chromo* domain from *Maverick* TEs and a conserved non-TE domain *zf-CCCH_6*. In addition to the conserved non-TE domain *zf-CCCH_6* and the two TE domains *Helicase_C* and *Chromo*, almost all copies of SHREC contain *SNF2_N* and *zf-PHD-like* domains. Approximately half of the fusion protein variants contain *ResII* or *PHD* domains, in addition to 88 more rarely associated domains (Fig. 4A; Supplementary Fig. S3). The Mit1 homology domain is primarily present in ascomycetes, with the highest representation in the Eurotiomycetes ($n=169$), Dothideomycetes ($n=115$) and Leotiomycetes ($n=34$). Lower numbers are found in Lecanoromycetes ($n=4$), Orbiliomycetes ($n=5$), Pezizomycetes ($n=9$), and Xylonomycetes ($n=1$). The Mit1 homology domain is largely absent in the large class of Saccharomycotina ($n=1$) and was only detected in *Schizosaccharomyces cryptophilus*, *S. japonicus* and *S. pombe* of the Taphrinomycotina. Weak representation is also found in basidiomycetes of the classes Agaricomycetes ($n=4$) and Dacrymycetes ($n=1$). In two ascomycetes (*Aspergillus carbonarius* and *Phialophora americana*), SHREC has a paralog, with one duplication that affected the gene with both TE domains and one duplication that affected the *Helicase_C* domain gene. A multiple sequence alignment of the duplicated genomic regions confirms the conservation of the individual domains (Fig. 4B). We further investigated evidence for helicase domains associated with host-TE fusion candidates focusing on the PF00271 underpinning helicase conserved C-terminal domains. The phylogenetic tree of proteins encoding PF00271 domains across the fungal kingdom includes both TE-host fusion candidates and other proteins. We found that helicase host-TE fusions were heterogeneously distributed across the tree with multiple terminal branches carrying host-TE fusion candidates (Supplementary Fig. S4). This is consistent with a pattern of repeated emergence of host-TE fusion candidates.

Manually refined inspection of host-TE fusion candidates

Host-TE fusion candidates likely include many false positives due to uncertainty in assigning protein domains to be of TE origin. To investigate a reduced but more strongly supported set of host-TE fusion candidates, we manually curated proteins shared by ≥5 genomes both for carrying an unambiguous TE-derived domain and at least one unambiguous non-TE domain (Supplementary Table S6). The stringent curation strongly reduced

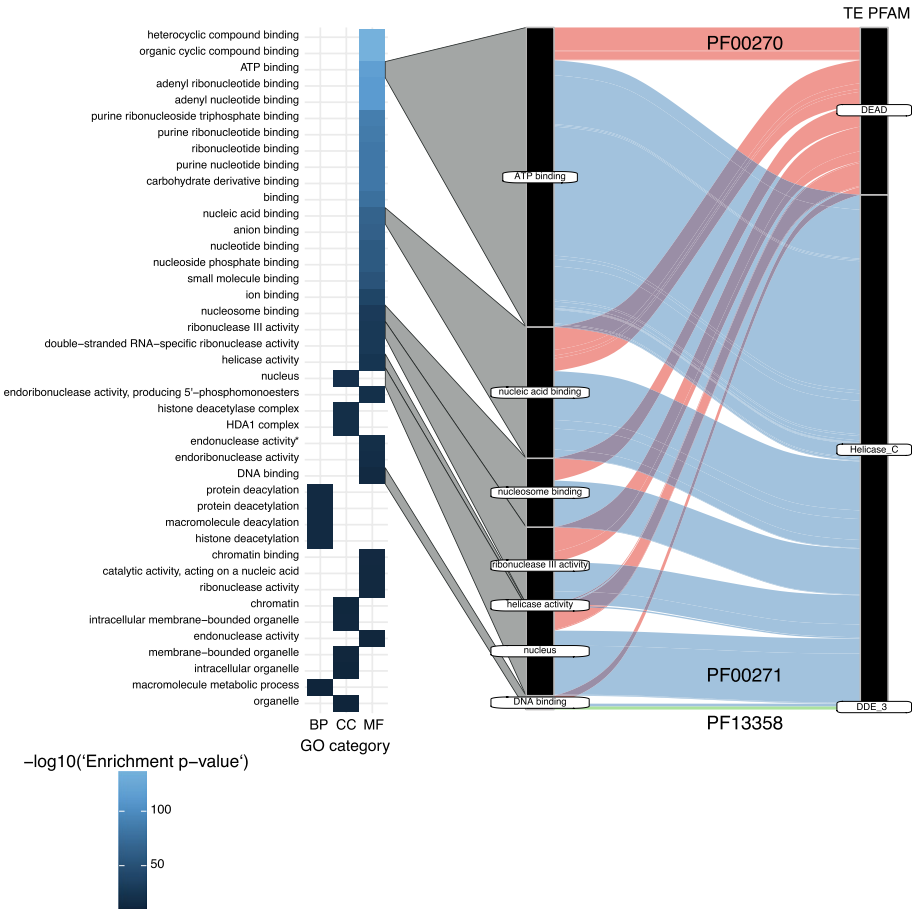

Fig. 3 Gene enrichment analysis: gene enrichment analysis of the non-TE derived domains and the corresponding TE-derived domains. * active with either ribo- or deoxyribonucleic acids and producing 5′-p hosphomonoesters

the candidate set to only 152 species. Outliers for high proportions of host-TE fusion candidates remained *A. ostoyae* ($n=668$) and *F. poae* ($n=108$). The curation removed also most helicase domains as these often are ambiguous in their origins. The most frequent retained domains include retroelements (i.e. RVT_1/ PF00078, reverse transcriptase, $n=417$), followed by gag-polyprotein putative aspartyl protease/PF13975 ($n=296$) and RVP_2/ PF00077 (retroviral aspartyl protease, $n=232$).

Discussion

Transposable elements are important facilitators of genome evolution, by providing regions of relaxed selection, positions of breakpoints from chromosomal rearrangements, by providing the means for horizontal gene transfer, gene mobility and reshuffling in the genome, or by providing coding regions, transcription factor binding sites and other structures to facilitate de novo proteins. The impact of TEs can be reversible, locally limited and short-lived. Yet, TEs may also have an impact over longer evolutionary time frames on the proteome diversification. Our analyses across 1237 fungal genomes revealed an uneven distribution of potential host-TE fusions among major fungal phylogenies, with a higher percentage of genes involved in host-TE fusions in Saccharomycotina. Opposed to vertebrates, plants and nematodes, where terminal inverted repeat transposase domains are predominantly associated with host-TE fusions, we detected helicases as the most abundant TE-derived domains [20, 27, 34]. The host domains of host-TE fusions show a broader diversity in function, but tend to be associated with processes involved in genome integrity and likely defense against foreign sequences including TEs.

TE-driven dynamics in the Saccharomycotina

We observed that the compact genomes of Saccharomycotina contain a higher proportion of host-TE fusions per gene compared to other fungi, while maintaining similar absolute numbers of host-TE fusions per genome. Given that all analyzed Saccharomycotina genomes have extremely low TE counts, we hypothesize that a significant proportion of detectable TEs in these species may

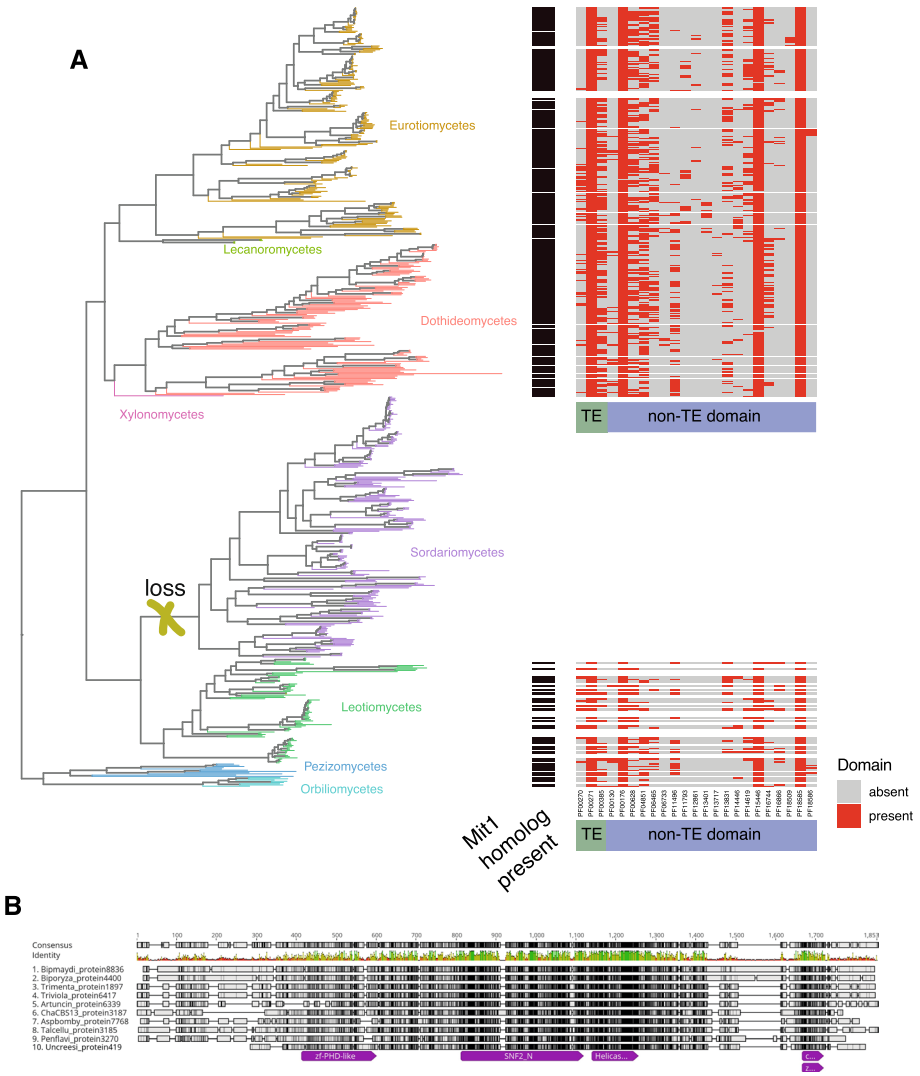

Fig. 4 Host-TE fusion candidate Mit1 domain homolog distribution in the fungal kingdom: **A** Subset of the phylogenetic tree for species with indication of a presence of Mit1 domain homologs from SHREC. The phylogenetic tree only shows Ascomycetes classes, not including the class of Saccharomycotina. Color indicates the class. Presence of the Mit1 domain homolog in the species is indicated by a black rectangle, and the presence of TE-derived domains and host-derived domains are represented with a red rectangle. **B** Multiple Sequence Alignment of a selected number of proteins that are homologs to the Mit1 domain from SHREC

be integrated into host-TE fusions [88–91]. Even tough TEs are rare and potentially less active than in other species, they might still play crucial roles in Saccharomycotina evolution, as seen for *CENP-B*. Notably, the absence of the ascomycete-specific defense mechanisms known as RIP (repeat-induced point mutations) against TEs in Saccharomycotina and Taphrinomycotina increases the potential for gene duplication followed by insertions of TEs and subsequent host-TE fusions events [48]. RIP is a mechanism that induces point mutations in all copies of duplicated regions of a certain length, affecting both transposable elements and genes [92–94]. RIP can introduce early stop codons or other deleterious mutations in coding regions, leading to loss-of-function of duplicated sequences [95]. Active RIP in a lineage can significantly limit the evolution of essential gene functions through gene duplication [96]. Consequently, RIP may underpin low rates of gene duplicates in ascomycetes [97]. In this context, host-TE fusions of essential genes could plausibly emerge after gene duplications, where one copy remains essential, while the other copy is under relaxed purifying selection, potentially leading to the gain of new functions through TE domain fusions. While RIP is elevating mutation rates for genes close to

TEs, RIP may reduce the potential to create new host-TE fusions. With the absence of RIP in Saccharomycotina and Taphrinomycotina, host-TE fusions could arise and be retained at higher rates. Having lower numbers of host-TE fusions in genomes highly affected by RIP is an indication that this might hold true, yet the absence of RIP is not leading to high amounts of host-TE fusions. Future studies may test in Saccharomycotina whether host-TE fusion proteins could be themselves involved in defenses against TEs in providing a mechanism to repress TEs low in the subphylum.

Helicase domains are predominant in fungal host-TE fusions

Most detected host-TE fusions encode helicase domains of likely TE origin. The most common source of helicases appears to be TEs of the DNA TE superfamilies *AcademH*, *KolobokH* or *Helitron* [98]. The specific *DEAD* and *Helicase_C* helicase domains were only recently recognized as of TE origin likely due to the recent discovery of *AcademH* and *KolobokH* TEs [49]. *AcademH* has been found as low-copy TEs in Basidiomycota, Ascomycota as well as Mucoromycota [98]. Helicases in general provide functions for the unwinding of DNA, DNA binding, and they are involved in DNA repair pathways [99]. Helicases from *Helitrons* though are known to be able to capture neighboring regions during transposition events [12, 100]. *Helitrons* might thus generate host-TE fusions through the capture of genes by a TE, rather than their own insertion into coding sequences. Once established, helicase-containing host-TE fusions might remain able to capture surrounding regions, which could be explained by a high presence-absence polymorphism of additional domains in most potential host-TE fusions involving helicases. A gene capture mechanism would also explain the high diversity of functions in host-TE fusions involving helicases, as well as the putative repeated fusion events across clades of PF00271-encoding proteins (Supplementary Fig. S4). Whether the preponderance of host-TE fusions with *AcademH*, *KolobokH* or *Helitron* helicases is related to such a promiscuous mechanism to capture neighboring genes remains unknown. Recurrent gene capture by helicase containing TEs could explain the high helicase diversity in fungi and their dominance among host-TE fusion genes [100].

Fungal host-TE fusions might be involved in silencing of repetitive regions

DNA binding activity is predominant among fungal host-TE fusion genes and is also featured among the most phylogenetically conserved fusions. The overrepresentation of DNA binding activity likely stems from the broad roles helicases play in nuclear functions. The domain Mit1 in the Snf2/Hdac repressive complex (SHREC) host-TE fusion candidate shows a patchy distribution in some classes of ascomycetes, with sparse presence in other clades. Some classes of ascomycetes do not contain the Mit1 homology domain, which could be an indication that this host-TE fusion was randomly lost. In *S. pombe*, SHREC is known to transcriptionally silence genes and TEs [87]. The host-TE fusion consistently contains two TE-derived domains, *Helicase_C* and *Chromo*. *Chromo* domains are known in the superfamilies of *Maverick* (alternatively *Polinton*) and chromoviruses (a group of RLG, formerly known as *Gypsy*) and are located at the C-terminus of the integrase [101–103]. *Chromo* domains interact mostly with methylated histones [102, 104]. The patchy distribution of the Mit1 homology domain and the prevalent partial loss of the *Chromo* domain indicate that the complex might not be present or not functional in all fungi, respectively [105].

Fungal proteomes have been significantly shaped by ancient and ongoing TE insertions, which may increase functional diversity and influence speciation. The exact mechanisms for creating functional proteins remain poorly documented. However, screens of populations will improve our understanding of these mechanisms. Identifying the processes responsible for creating host-TE fusions remains challenging. Regardless of genomic defenses, non-deleterious insertions of TEs into open reading frames of existing genes are likely very rare. We suggest gene capture by TEs (i.e., *Helicases*) as an alternative mechanism to TE insertion into introns followed by alternative splicing to create host-TE fusion. Detecting host-TE fusions in genomes presents several challenges due to the complex nature of the events and the fact that most fusions are ancient and likely no longer recognizable as host-TE fusion events. Accurately detecting host-TE fusions is further complicated by fragmentated genome assemblies, incomplete knowledge of TE-derived domains, and the rarity of events leading to host-TE fusions. Additionally, bioinformatics-based approaches often cannot predict novel functions of host-TE fusion genes. Future research with improved genome assembly quality, refined curation, improved computational tools and functional investigations will expand our understanding of contemporary and historic host-TE fusions in the fungal kingdom.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13100-024-00312-1.

Additional file 1.

Additional file 2.

Additional file 3.

Acknowledgements
We thank Hadi Quesneville, Casey Bergman, Ksenia Krasileva, Anne Nakamoto and Anna Muszewska for help with TE associated PFAM. We also thank Sabina Tralamazza for help with gene cluster analysis.

Authors' contributions
UO and DC conceived the study. UO and TB analyzed data. TB contributed datasets. UO and DC wrote the manuscript with input from TB. All authors approved the final manuscript.

Funding
UO and DC were supported by Swiss National Science Foundation grants 206850 and 201149.

Availability of data and materials
All generated data is reported as Supplementary Tables S1-S6 and File S1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 21 October 2023 Accepted: 4 January 2024
Published: 20 January 2024

References

1. Lofgren LA, Ross BS, Cramer RA, Stajich JE. The pan-genome of *Aspergillus fumigatus* provides a high-resolution view of its population structure revealing high levels of lineage-specific diversity driven by recombination. PLoS Biol. 2022;20(11):e3001890.
2. Graveley BR. Alternative splicing: Increasing diversity in the proteomic world. Trends Genet. 2001;17:100–7.
3. Andersson DI, Jerlström-Hultqvist J, Näsvall J. Evolution of New Functions De Novo and from Preexisting Genes. Cold Spring Harb Perspect Biol. 2015;7:a017996.
4. Van Oss SB, Carvunis AR. De novo gene birth. PLoS Genet. 2019;15:1–23.
5. Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20:1313–26.
6. Torres DE, Thomma BPHJ, Seidl MF. Transposable Elements Contribute to Genome Dynamics and Gene Expression Variation in the Fungal Plant Pathogen Verticillium dahliae. Genome Biol Evol. 2021;13:1–19.
7. Schrader L, Kim JW, Ence D, Zimin A, Klein A, Wyschetzki K, et al. Transposable element islands facilitate adaptation to novel environments in an invasive species. Nat Commun. 2014;5:1–10.
8. Fouché S, Oggenfuss U, Chanclud E, Croll D. A Devil's Bargain with transposable elements in plant pathogens. Trends Genet. 2022;38:222–30.
9. Torres DE, Oggenfuss U, Croll D, Seidl MF. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biol Rev. 2020;34:136–43.
10. Devos KM, Brown JKM, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Res. 2002;12:1075–9.
11. Grabundzija I, Messing SA, Thomas J, Cosby RL, Bilic I, Miskey C, et al. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nat Commun. 2016;7:10716.
12. Morgante M, Brunner S, Pea G, Fengler K, Zuccolo A, Rafalski A. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nat Genet. 2005;37:997–1002.
13. Kolkman JA, Stemmer WP. Directed evolution of proteins by exon shuffling. Nat Biotechnol. 2001;19:423–8.
14. Ejima Y, Yang L. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Hum Mol Genet. 2003;12:1321–8.
15. Goodwin TJD, Ormandy JE, Poulter RTM. L1-like non-LTR retrotransposons in the yeast *Candida albicans*. Curr Genet. 2001;39:83–91.
16. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR. Pack-MULE transposable elements mediate gene evolution in plants. Nature. 2004;431:569–73.
17. Gupta S, Gallavotti A, Stryker GA, Schmidt RJ, Lal SK. A novel class of Helitron- related transposable elements in maize contain portions of multiple pseudogenes. Plant Mol Biol. 2005;57:115–27.
18. Gluck-Thaler E, Ralston T, Konkel Z, Ocampos CG, Ganeshan VD, Dorrance AE, et al. Giant Starship Elements Mobilize Accessory Genes in Fungal Genomes. Mol Biol Evol. 2022;39:msac109.
19. Urquhart AS, Vogan AA, Gardiner DM, Idnurm A. Starships are active eukaryotic transposable elements mobilized by a new family of tyrosine recombinases. Proc Natl. 2023;15:2017.
20. Widen SA, Bes IC, Koreshova A, Pliota P, Krogull D, Burga A. Virus-like transposons cross the species barrier and drive the evolution of genetic incompatibilities. Science (1979). 2023;380:2022.07.12.499685.
21. Zhang HH, Peccoud J, Xu MRX, Zhang XG, Gilbert C. Horizontal transfer and evolution of transposable elements in vertebrates. Nat Commun. 2020;11:1–10.
22. Vogan AA, Ament-Velásquez SL, Bastiaans E, Wallerman O, Saupe SJ, Suh A, et al. The Enterprise, a massive transposon carrying Spok meiotic drive genes. Genome Res. 2021;31:789–98.
23. Tralamazza SM, Gluck-Thaler E, Feurtey A, Croll D. Copy number variation introduced by a massive mobile element underpins global thermal adaptation in a fungal wheat pathogen. bioRxiv. 2023;1–44.
24. Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, et al. Ten things you should know about transposable elements. Genome Biol. 2018;19:199.
25. Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass *Brachypodium distachyon*. Genome Biol Evol. 2017;10:1–38.
26. Baucom RS, Estill JC, Leebens-Mack J, Bennetzen JL. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Res. 2009;19:243–54.
27. Hoen DR, Bureau TE. Discovery of novel genes derived from transposable elements using integrative genomic analysis. Mol Biol Evol. 2015;32:1487–506.
28. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
29. Wells JN, Feschotte C. A Field Guide to Transposable Elements. Annu Rev Genet. 2020;54:7–34.
30. Koonin EV, Makarova KS, Wolf YI, Krupovic M. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nat Rev Genet. 2020;21:119–31.
31. Agrawal A, Eastman QM, Schatz DG. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature. 1998;394:744–51.
32. Kapitonov VV, Jurka J. RAG1 core and V(D) J recombination signal sequences were derived from Transib transposons. PLoS Biol. 2005;3:0998–1011.
33. Yurchenko V, Xue Z, Sadofsky M. The RAG1 N-terminal domain is an E3 ubiquitin ligase. Genes Dev. 2003;17:581–5.
34. Cosby RL, Judd J, Zhang R, Zhong A, Garry N, Pritham EJ, et al. Recurrent evolution of vertebrate transcription factors by transposase capture. Science (1979). 2021;371:eabc6405.
35. Cam HP, Noma KI, Ebina H, Levin HL, Grewal SIS. Host genome surveillance for retrotransposons by transposon-derived proteins. Nature. 2008;451:431–6.
36. Mateo L, González J. Pogo-Like transposases have been repeatedly domesticated into CENP-B-Related Proteins. Genome Biol Evol. 2014;6:2008–16.
37. Smit AFA, Riggs AD. Tiggers and other DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A. 1996;93:1443–8.
38. Wang J, Han GZ. Unearthing LTR Retrotransposon gag Genes Co-opted in the Deep Evolution of Eukaryotes. Mol Biol Evol. 2021;38:3267–78.

39. Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol. 2012;10:417–30.
40. Hess J, Skrede I, Wolfe BE, La BK, Ohm RA, Grigoriev IV, et al. Transposable element dynamics among asymbiotic and ectomycorrhizal amanita fungi. Genome Biol Evol. 2014;6:1564–78.
41. Lorrain C, Feurtey A, Möller M, Haueisen J, Stukenbrock EH. Dynamics of transposable elements in recently diverged fungal pathogens: Lineage-specific transposable element content and efficiency of genome defenses. G3 Genes Genom Genet. 2021;11:0–12.
42. Gladyshev E. Repeat-Induced Point Mutation (RIP) and Other Genome Defense Mechanisms in Fungi. Microbiol Spectr. 2017:5.
43. Omrane S, Audéon C, Ignace A, Duplaix C, Aouini L, Kema G, et al. Plasticity of the MFS1 Promoter Leads to Multidrug Resistance in the Wheat Pathogen Zymoseptoria tritici. mSphere. 2017;2:1–42. Mitchell AP, editor
44. Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, et al. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. Environ Microbiol. 2015;17:2805–23.
45. Hartmann FE, Sánchez-Vallet A, McDonald BA, Croll D. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. ISME J. 2017;11:1189–204.
46. Frantzeskakis L, Kusch S, Panstruga R. The need for speed: compartmentalized genome evolution in filamentous phytopathogens. Mol Plant Pathol. 2019;20:3–7.
47. Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 2015;35:57–65.
48. van Wyk S, Wingfield BD, De Vos L, van der Merwe NA, Steenkamp ET. Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota. Front Microbiol. 2021;11:622368.
49. Muszewska A, Steczkiewicz K, Stepniewska-Dziubinska M, Ginalski K. Transposable Elements Contribute To Fungal Genes and Impact Fungal Lifestyle. Sci Rep. 2019;9:1–10.
50. Belyayev A. Bursts of transposable elements as an evolutionary driving force. J Evol Biol. 2014;27:2573–84.
51. Shen XX, Opulente DA, Kominek J, Zhou X, Steenwyk JL, Buh KV, et al. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell. 2018;175:1533–45.
52. Li Y, Steenwyk JL, Chang Y, Wang Y, James TY, Stajich JE, et al. A genome-scale phylogeny of Fungi; insights into early evolution, radiations, and the relationship between taxonomy and phylogeny. 2020;1–51.
53. Manni M, Berkeley MR, Seppey M, Simão FA, Zdobnov EM. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Mol Biol Evol. 2021;38:4647–54.
54. Mühlhausen S, Findeisen P, Plessmann U, Urlaub H, Kollmar M. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Res. 2016;26:945–55.
55. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
56. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
57. Stamatakis A. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
58. Lanfear R, Calcott B, Ho SYW, Guindon S. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Mol Biol Evol. 2012;29:1695–701.
59. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, Von Haeseler A, et al. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Mol Biol Evol. 2020;37:1530–4.
60. Wang LG, Lam TTY, Xu S, Dai Z, Zhou L, Feng T, et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Mol Biol Evol. 2020;37:599–603.
61. Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods Ecol Evol. 2017;8:28–36.
62. Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, et al. GgtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. Mol Biol Evol. 2021;38:4039–42.
63. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer EL, et al. Pfam: The protein families database in 2021. Nucleic Acids Res. 2021;49:D412–9.
64. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7:e1002195.
65. Hane JK, Paxman J, Jones DAB, Oliver RP, de Wit P. "CATAStrophy," a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There? Front Microbiol. 2020;10:3088.
66. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46:95–101.
67. Emms DM, Kelly S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. bioRxiv. 2018; 1–14.
68. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using diamond. Nat Methods. 2014;12:59–60.
69. Hoede C, Arnoux S, Moisset M, Chaumier T, Inizan O, Jamilloux V, et al. PASTEC: An automatic transposable element classification tool. PLoS One. 2014;9:1–6.
70. Quinlan AR, Hall IM. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26:841–2.
71. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, et al. The InterPro protein families database: The classification resource after 15 years. Nucleic Acids Res. 2015;43:213–21.
72. Pagès H, Carlson M, Falcon S, Li N. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor. 2023. https://doi.org/10.18129/B9.bioc.AnnotationDbi. https://bioconductor.org/packages/AnnotationDbi.
73. Morgan M, Falcon S, Gentleman R. GSEABase: Gene set enrichment data structures and methods. 2023. https://doi.org/10.18129/B9.bioc.GSEABase. https://bioconductor.org/packages/GSEABase.
74. Gentleman R. Category: Category Analysis. 2021.
75. Blin K, Wolf T, Chevrette MG, Lu X, Schwalen CJ, Kautsar SA, et al. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Res. 2017;45:W36–41.
76. Kautsar SA, van der Hooft JJJ, de Ridder D, Medema MH. BiG-SLiCE: A highly scalable tool maps the diversity of 1.2 million biosynthetic gene clusters. Gigascience. 2021;10:1–17.
77. Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539. Internet, Cited 2023 Dec 20, Available from. https://doi.org/10.1038/msb.2011.75.
78. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, et al. A new bioinformatics analysis tools framework at EMBL–EBI. Nucleic Acids Res. 2010;38:W695–9. Internet, Cited 2023 Dec 20, Available from. https://doi.org/10.1093/nar/gkq313.
79. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. PLoS One. 2010;5:e9490. Internet, Cited 2023 Dec 20, Available from. https://doi.org/10.1371/journal.pone.0009490.
80. Balaban M, Moshiri N, Mai U, Jia X, Mirarab S. TreeCluster: Clustering biological sequences using phylogenetic trees. PLoS One. 2019;14:e0221068. Internet, Cited 2023 Dec 20, Available from. https://doi.org/10.1371/journal.pone.0221068.
81. Shen X-X, Steenwyk JL, LaBella AL, Opulente DA, Zhou X, Kominek J, et al. Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. Sci Adv. 2020;6:eabd0079.
82. Li Y, Steenwyk JL, Chang Y, Wang Y, James TY, Stajich JE, et al. A genome-scale phylogeny of the kingdom Fungi. Curr Biol. 2021;31:1653–1665.e5.
83. Chellapan BV, Van Dam P, Rep M, Cornelissen BJC, Fokkens L. Non-canonical Helitrons in Fusarium oxysporum. Mob DNA. 2016;7:1–16. Internet, Cited 2023 Dec 20, Available from:. https://doi.org/10.1186/s13100-016-0083-7.
84. Heinzelmann R, Rigling D, Sipos G, Münsterkötter M, Croll D. Chromosomal assembly and analyses of genome-wide recombination rates in the forest pathogenic fungus Armillaria ostoyae. Heredity. 2020;124:699–713. Internet, Cited 2023 Dec 20, Available from: https://www.nature.com/articles/s41437–020-0306-z
85. Nakagawa H, Lee JK, Hurwitz J, Allshire RC, Nakayama JI, Grewal SI, et al. Fission yeast CENP-B homologs nucleate centromeric heterochromatin

by promoting heterochromatin-specific histone tail modifications. Genes Dev. 2002;16:1766–78.
86. Hammond SM. Dicing and slicing: The core machinery of the RNA interference pathway. FEBS Lett. 2005;579:5822–9.
87. Sugiyama T, Cam HP, Sugiyama R, Noma KI, Zofall M, Kobayashi R, et al. SHREC, an Effector Complex for Heterochromatic Transcriptional Silencing. Cell. 2007;128:491–504.
88. Bleykasten-Grosshans C, Neuvéglise C. Transposable elements in yeasts. C R Biol. 2011;334:679–86.
89. Holton NJ, Goodwin TJD, Butler MI, Poulter RTM. An active retrotransposon in *Candida albicans*. Nucleic Acids Res. 2001;29:4014–24.
90. Zhu CX, Yan L, Wang XJ, Miao Q, Li XX, Yang F, et al. Transposition of the zorro2 retrotransposon is activated by miconazole in *candida albicans*. Biol Pharm Bull. 2014;37:37–43.
91. Potocki L, Kuna E, Filip K, Kasprzyk B, Lewinska A, Wnuk M. Activation of transposable elements and genetic instability during long-term culture of the human fungal pathogen *Candida albicans*. Biogerontology. 2019;20:457–74.
92. Selker EU, Garrett PW. DNA sequence duplications trigger gene inactivation in Neurospora crassa. Proc Natl Acad Sci U S A. 1988;85:6870–4.
93. Galagan JE, Selker EU. RIP: the evolutionary cost of genome defense. Trends Genet. 2004;20:417–23.
94. Selker EU. 15 Repeat-induced gene silencing in fungi. Adv Genet. 2002;46:439–50.
95. Dhillon B, Cavaletto JR, Wood KV, Goodwin SB. Accidental Amplification and Inactivation of a Methyltransferase Gene Eliminates Cytosine Methylation in Mycosphaerella graminicola. Genetics. 2010;186:67–77.
96. Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP. Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. Genet Transform Syst Fungi. 2015;2:55–68.
97. Skamnioti P, Furlong RF, Gurr SJ. The fate of gene duplicates in the genomes of fungal pathogens. Commun Integr Biol. 2008;1:196–8.
98. Kojima KK. AcademH, a lineage of Academ DNA transposons encoding helicase found in animals and fungi. Mob DNA. 2020;11:1–11.
99. Croteau DL, Popuri V, Opresko PL, Bohr VA. Human RecQ Helicases in DNA Repair, Recombination, and Replication. Annu Rev Biochem. 2014;83:519–52.
100. Chellapan BV, Van Dam P, Rep M, Cornelissen BJC, Fokkens L. Non-canonical Helitrons in Fusarium oxysporum. Mob DNA. 2016;7:1–6.
101. Pritham EJ, Putliwala T, Feschotte C. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 2007;390:3–17.
102. Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Res. 2008;18:359–69.
103. Quesneville H. Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. Mob DNA. 2020;11:1–13.
104. Brehm A, Tufteland KR, Aasland R, Becker PB. The many colours of chromodomains. BioEssays. 2004;26:133–40.
105. Lei B, Capella M, Montgomery SA, Borg M, Osakabe A, Goiser M, et al. A Synthetic Approach to Reconstruct the Evolutionary and Functional Innovations of the Plant Histone Variant H2A.W. Curr Biol. 2021;31:182–191.e5.

Publisher's Note