

# Cut-and-paste transposons in fungi with diverse lifestyles

Anna Muszewska<sup>1,\*#</sup>, Kamil Steczkiewicz<sup>2,#</sup>, Marta Stepniewska-Dziubinska<sup>1</sup>, Krzysztof Ginalski<sup>2</sup>

<sup>1</sup> Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Pawinskiego 5A, 02-106 Warsaw, Poland

<sup>2</sup> Laboratory of Bioinformatics and Systems Biology, CeNT, University of Warsaw, Zwirki i Wigury 93, 02-089 Warsaw, Poland

# The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors

**\*Author for Correspondence:** Anna Muszewska, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Poland, +48 592 57 59, [musze@ibb.waw.pl](mailto:musze@ibb.waw.pl)

**Running title:** DDE transposons in fungi

**Keywords:** DNA transposon, fungi, genome architecture, fungal ecology

## Abstract

Transposons (TEs) shape genomes *via* recombination and transposition, lead to chromosomal rearrangements, create new gene neighbourhoods and alter gene expression. They play key roles in adaptation either to symbiosis in *Amanita* genus or to pathogenicity in *Pyrenophora tritici-repentis*. Despite growing evidence of their importance, the abundance and distribution of mobile elements replicating in a “cut and paste” fashion is barely described so far. In order to improve our knowledge on this old and ubiquitous class of transposable elements, 1,730 fungal genomes were scanned using both *de novo* and homology-based approaches. DNA TEs have been identified across the whole dataset and display uneven distribution from both DNA TE classification and fungal taxonomy perspectives. DNA TE content correlates with genome size, which confirms that many transposon families proliferate simultaneously. In contrast, it is independent from intron density, average gene distance and GC content. TE count is associated with species’ lifestyle and tends to be elevated in plant symbionts and decreased in animal parasites. Lastly, we found that fungi with both RIP and RNAi systems have more total DNA TE sequences but less elements retaining a functional transposase, what reflects stringent control over transposition.

## Introduction

Transposable elements (TEs) have been long neglected, considered genomic dark matter. Currently, TEs are commonly recognised as ubiquitous and vital components of almost all prokaryotic and eukaryotic genomes. Strikingly, TEs have recently been reported even in the giant virus *Pandoravirus salinus* genome (Sun, et al. 2015). The broad taxonomic distribution of major TE lineages is likely a sign of their ancient origin and evolution mainly through vertical transmission (Daboussi, et al. 2003).

Transposons extensively shape eukaryotic genomes by chromosomal rearrangements, pseudogenisation, domestication and gene shuffling (Feschotte and Pritham 2007; Pritham 2009). There are well documented cases of spectacular impact of TEs on specific processes such as karyotype instability in gibbons (Carbone, et al. 2014), genome rearrangements in ciliates (Yerlici and Landweber 2014) and development of immune systems in both prokaryotes (CRISPR) and vertebrates (V(D)J recombination) (Koonin and Krupovic 2015). Transposon proliferation and accumulation is one of important sources of raw material for regulatory sequences for host genes (Rebollo, et al. 2012). Under normal conditions, TEs usually remain inactive and silenced, however, some of them become activated at certain points of ontology, e.g. LINE1 in neurogenesis (Erwin, et al. 2014) and ERV in embryogenesis (Grow, et al. 2015), or in stressful conditions (Makarevitch, et al. 2015; Rey, et al. 2016). There have also been cases of documented correlation between TE abundance and host lifestyle, e.g. in symbiotic *Amanita* mushrooms (Hess, et al. 2014) and in pathogenic *Pyrenophora tritici-repentis*. The latter has a genome remarkably abundant in young elements neighbouring effector genes (Manning, et al. 2013).

### DNA TE

Eukaryotic transposons are classified into two main classes, which are further split into orders, superfamilies, families and subfamilies (Kapitonov and Jurka 2008). Class I elements, retrotransposons, use an RNA intermediate during transposition, synthesised based on a DNA template. Class II DNA transposons constitute a large and diverse group of mobile elements that utilise a DNA intermediate during mobilization. Based on the transposition mechanism, they can be classified into three major types:

i) Maverick/Polintons with unknown mechanism of transposition, ii) helitrons with a putative mechanism similar to rolling circle replication and iii) "cut and paste" DNA transposons with a DDE nuclease for excising the element directly as a double-stranded DNA. The latter TEs possess a simple architecture comprising of a transposase and terminal inverted repeats (TIR), and therefore are referred to as TIR transposable elements. Here, the transposase itself is a nuclease, in all eukaryotes commonly retaining RNase H-like fold (Yuan and Wessler 2011), decorated with plethora of DNA binding and other accessory elements. The transposase endonucleolytic activity is essential for excising the mobile element from the donor site, what generates double strand breaks (DSBs) eventually repaired by the host (Liu and Wessler 2017). The insertion of TE into an acceptor site often generates short duplications of the flanking region, target-site duplication (TSD), which are characteristic for a given TE type.

From the perspective of protein structure evolution, DNA transposons harbour protein domains classified to structural folds that had originated before cellular organisms emerged (Abrusán, et al. 2013). Consistently, the presence of most of the "cut and paste" superfamilies in many eukaryotic lineages and their similarity to the prokaryotic insertion sequences suggest that DNA TEs may be older than the last common eukaryotic ancestor (Pritham 2009).

## **Classification**

RepBase (Kapitonov and Jurka 2008) is the golden standard collection of transposon consensus sequences used extensively by the genomic community as a reference in annotation of genomes of non-model organisms. In 2011, Yuan and Wessler revised the RepBase classification of DNA "cut and paste" transposons and grouped the existing families into superfamilies based on conserved motifs in their nuclease domain, TIRs and TSDs (Yuan and Wessler 2011). This classification reasonably reflects evolutionary relationships and will be used in this work hereafter. The most conserved TE sequence regions, e.g. those of RNase H-like endonuclease domains, are broadly used for detection and classification of transposons (Yuan and Wessler 2011). However, fast substitution rate in concert with recursive losses of the whole TE superfamilies and families along the tree of life render deep phylogenetic inference risky and error prone. TE regions that display sequence conservation sufficient for phylogenetic

inference are very short, even within the commonly conserved RNase H-like nuclease domain, and in consequence, provide limited signal suitable for tree topology searches. According to the acclaimed models of evolution, TEs are subject to either neutral or negative selective pressure due to genome defense mechanisms, what results in even faster mutation rate than an average for a given taxon. In consequence, relationships between TE superfamilies still remain unresolved.

## DNA TE in Fungi

Fungi with sequenced genomes represent a variety of lifestyles, genome sizes and taxonomic lineages, what makes them the eukaryotic kingdom of choice for comparative genomics. Retrotransposons, particularly LTR retrotransposons (Gypsy/Ty3), are the best studied, the easiest to annotate and the most abundant transposons in fungi. DNA transposons are perhaps equally ubiquitous but understudied compared to the former. Initially, only Tc1/Mariner, hAT, MULE and MITEs were reported from fungi (Daboussi, et al. 2003). Although EnSpm, hAT, PiggyBac, PIF-Harbinger, MULE, Merlin and Tc1/Mariner superfamilies are currently reported from many fungal lineages (Pritham 2009), there are some DNA TEs to be detected in fungi. Some DNA transposon families have been identified in single fungal taxa, e.g. P element was reported from *Allomyces* (Yuan and Wessler 2011), Sola transposon in *Rhizophagus irregularis* (Bao, et al. 2009) and Dada in *Laccaria laccata* (Kojima and Jurka 2013), exclusively.

TE fate in fungi is determined by multiple factors, among them genome defense mechanisms such as repeat-induced point mutation (RIP, discovered in *Neurospora crassa* (Selker, et al. 1987; Singer and Selker 1995)), sex-induced silencing (SIS, in *Cryptococcus neoformans* (Wang, et al. 2010)), methylation induced pre-meiotically (MIP, in *Ascobolus immersus* (Barry, et al. 1993)), meiotic silencing process (MSUD, discovered in *N. crassa* ((Shiu, et al. 2001)) and quelling (discovered in *N. crassa* (Rountree and Selker 2010)). MSUD, MIP and SIS occur during meiosis, whereas quelling takes place in the vegetative phase. MIP and RIP processes require a specialised C5-cytosine methylase Masc1 (*Ascobolus*) or RID (*N. crassa*), which is a member of an ancient protein family, but this specific fungal

subfamily of methylases has been documented solely in Pezizomycotina (Gladyshev and Kleckner 2016). In MSUD, SIS and quelling, the machinery providing genome defense is composed of conserved proteins from RNAi pathway, such as RNA-dependent RNA polymerase (RdRP, QDE-1 in *N. crassa*), piwi-Argonaute protein (QDE-2 in *N. crassa*), RecQ helicase (QDE-3 in *N. crassa*), Dicer-like proteins (DCL in *N. crassa*) and other components of the RISC complex (Chang, et al. 2012). RNAi-mediated mechanisms seem to be ancient and widespread means of defense against transposition and viral invasion, while RIP and MIP possibly evolved as a secondary weapon to control repetitive genome content.

Here, we present a comprehensive analysis of DNA transposons in publicly available fungal genomes. We have developed a semi-automated approach for detecting typical TIR-containing DNA TEs and their annotation to superfamilies. We raise questions regarding TE abundance and its correlation with host lifestyle, genome defense mechanisms and genome complexity.

## Materials and methods

### Genomic sequences

Genomic sequences were downloaded from NCBI genome database on the 18th of August 2016. Initially, 1,746 assemblies were obtained but only full assemblies were considered, resulting in a dataset of 1,730 genomes belonging to 847 fungal species. 1,726 out of 1,730 assemblies had taxonomic information needed for statistical analyses. The complete list of genomes together with their references is listed in [Supplementary Table S1a](#).

### Transposase-based classification

All DNA transposon sequences stored in RepBase (Kapitonov and Jurka 2008) were downloaded in September 2016 and scanned against Pfam database v. 30 with pfam\_scan.pl script as a wrapper for HMMer (Mistry, et al. 2013). The obtained protein domain architectures were used to determine the characteristics of each DNA transposon superfamily. After manual curation of spurious cases, we worked out a one-to-one assignment for nine DNA transposon superfamilies (Tc1-Mariner, hAT, CMC-Enspm,

Merlin, MULE, PiggyBac, PIF-Harbinger, Transib and P) to Pfam domains. Maverick/Polintons, LTR retrotransposons, retroviruses and Ginger encode an endonuclease domain from integrase core domain family (rve, PF0065), which is one of the largest families in Pfam database. RNase H profiles for Sola1-3 and Zator elements were defined in RNase H-like superfamily classification by Majorek and others in 2014 (Majorek, et al. 2014) (Sola1 as E.10, Sola2 as E.12, Sola3 as E.13 and Zator as A.28) and the alignments provided in their work were used to build the respective HMM profiles. Academ (Kapitonov and Jurka 2010), Novosib (Kapitonov and Jurka 2006) and Kolobok (Kapitonov and Jurka 2007) nucleases were defined by Yuan and Wessler (Yuan and Wessler 2011). RNase H profiles for Academ and Kolobok were built using the alignments provided in the supplementary materials. Additionally, for Academ, we added more distant RepBase representatives described after Yuan and Wessler publication. In the case of Dada transposons, a sequence profile was built using RepBase transposase sequences as a reference for the sequence alignment building, with RNase H borders as described by Kojima and Jurka (Kojima and Jurka 2013). RNase H profiles for KDZ (Kyakuja, Dileera, and Zisupton) and Plavaka were built using representative sequences for these superfamilies from Iyer and colleagues study (supplementary materials) (Iyer, et al. 2014), the RNase H definition was based on their definition. Profile-profile comparisons were performed using META-Basic (Ginalski, et al. 2004), multiple sequence alignments were built with MAFFT 7 (Yamada, et al. 2016), and HMM profiles were built with HMMER3 (Mistry, et al. 2013) ([Supplementary File S1](#)). A schematic representation of the key steps of the analysis is summarised on [Figure 1](#).

## Detection of DNA transposons

DNA transposons were detected within 1,728 out of 1,730 genome assemblies. In order to ensure a reference-free initial detection, we used inverted repeat finder, irf (Warburton, et al. 2004), to identify TE candidates. Irf program scans for pairs of repeats within a given distance (20 kb in our searches). The resulting set was expected to group many possible false positives, among them simple repeats, repeat-rich protein coding regions, etc. To increase the method specificity, we filtered out (using pfam\_scan.pl with

an e-value confidence threshold of 0.001) sequences lacking RNase H or specific accessory domains (OTU, helicase, C48 protease). Since irf produces multiple overlapping hits, we removed them by sequence clustering using CD-HIT (Fu, et al. 2012) with sequence similarity threshold set to 100 and query coverage set to 99% of the shorter sequence. In parallel, we predicted transposons *de novo* with RepeatModeler and filtered the obtained TE candidates with the same procedure described above. The selected hits from irf and RepeatModeler were merged together with RepBase and the resulting database was used as a custom library for subsequent RepeatMasker searches. RepeatMasker hits were also scanned for the presence of transposase domains. All transposons including non-autonomous copies are classified by RepeatMasker based on the provided reference. Autonomous copies have a refined classification based on transposase similarity to known protein domains. Eventually, two groups of TEs were obtained: i) RepeatMasker hits with transposase domain, and ii) RepeatMasker hits with DNA TE annotation and score higher than 200, not necessarily containing a detectable transposase domain. The first group consists of elements, which are more likely to be active; the latter harbours all elements with a strong score regardless of their state of deterioration. Such data separation enables us to speculate about the overall abundance of TEs and in parallel focus on more reliable younger TEs.

## **DDE superfamilies**

Based on the manually curated one-to-one mappings between TE and Pfam domains, the obtained TEs were classified into distinct DNA transposon superfamilies. The resulting collections of transposon sequences were used to determine DNA transposon abundance in each of the analysed genomes. Detailed counts of all DNA transposons within each of the superfamilies are listed in [Supplementary Table S1b](#).

## **Additional datasets and statistical analysis**

Genome statistics (size, density, intron per gene) were computed based on the assembly sequences and gff annotation files derived from the NCBI database, where available (see [Supplementary Table S1c](#)). GC content of all DNA TEs and those with a potential transposase, together with total length of DNA transposons, are also included in [Supplementary Table S1c](#).



RID homologues were collected using jackhmmer web-search (Mistry, et al. 2013) with *Neurospora* RIP-defective sequence (XP\_011392925) as a query against the fungal subset of UNIPROT database, and subsequently clustered with CLANS (Frickey and Lupas 2004) in order to separate the RID methylase from proteins retaining a C-5 cytosine-specific DNA methylase domain (PF00145) in different functional contexts. RNAi proteins (Argonaute, Dicer-like and RNA-dependent RNA polymerase) were searched using reference sequences derived from funRNA as queries (Choi, et al. 2014). The obtained RIP- and RNAi-related protein reference sequences were aligned, converted to HMM profiles, merged into a single database for hmmscan and subsequently used as a target database for scanning all analysed 634 fungal proteomes. The obtained fungal sequences were clustered in CLANS (Frickey and Lupas 2004) together with the respective reference sequences. Sequences belonging to each protein family were aligned using MAFFT iterative alignment method (Yamada, et al. 2016). Each alignment was manually curated; all proteins with deletions in the conserved regions of the enzymatic domain were excluded (accession numbers of RNAi and RIP sequences are provided in [Supplementary File S3](#), counts of all RNAi and RIP proteins per assembly is present in [Supplementary Table S1d](#)). Only 633 assemblies with predicted proteomes were considered in genome defense analyses.

The summary of fungal lifestyles was derived based on the available literature. Categories including host type (plant, animal, fungus), main habitat (soil/dung, water) and lifestyle (pathogenic, symbiotic and saprotrophic) were assigned to every species. Noteworthy, a single fungus could represent multiple categories, if applicable, e.g. species functioning both as a plant symbiont and animal pathogen (see [Supplementary Table S1e](#)). The taxonomical annotation was derived from the NCBI taxonomy database, with manual fine tuning, when needed (see [Supplementary Table S1b](#)).

Taxonomic categories were chosen differently for separate taxons: at class level for over-represented Agaricomycotina and Pezizomycotina and at subphylum level for smaller taxa in order to obtain groups of a more comparable size. Taxonomic categories with at least 5 observations were encoded with separate binary vectors. In order to retain maximum biological information for taxa with less than 5 members, lifestyle features were kept, while taxonomy was represented with a null vector. Exploratory

analysis and basic statistics for the dataset were carried out using pandas and seaborn Python packages. Statistical tests were performed in Python with the scipy package. Influence of different lifestyle and taxonomic factors on transposon abundance was determined using Mann-Whitney U test.

The whole dataset comprised 1,726 assemblies (four assemblies were rejected due to incomplete information) belonging to 847 species, of which 633 had proteome data available. While majority of analyses in the manuscript were based on the set of 1,726 assemblies, regardless of the phylogenetic proximity, the analyses of genome defense mechanisms are based on a set of 633 assemblies with predicted proteomes. In order to verify the impact of phylogenetic covariance, additional analyses were performed in which all assemblies for one genus were considered together. For numeric variables an average value was taken for analyses, while in the case of binary and categorical variables (i.e. lifestyle features, defense mechanisms) the most frequent (dominant) category was used in the analyses. In order to deal with covariance of distinct features we have also tested the dataset building linear models considering all variables at a time using scikit-learn package. The model was trained with Stochastic Gradient Descent (SGD) which allows for efficient training of linear regression models. This approach enabled identification of covering factors which are listed in [Supplementary File S2](#).

The whole dataset is available as [Supplementary Tables S1b-e](#) and the code for statistical procedures is available as a Python code in a Jupyter Notebook (Kluyver, et al. 2016) [Supplementary File S2](#).

## **Phylogenetic analyses**

Sequences were aligned using MAFFT iterative alignment method (linsi, 100 iterations). Sequences lacking key catalytic residues were discarded from the alignment. Uncertain regions of the alignment were removed with TrimAl (Capella-Gutierrez, et al. 2009) using automated1 mode. Best suiting substitution model was selected with ProtTest 3.4 (Darriba, et al. 2011). Phylogeny inference was performed with the best fitting settings in PhyML 3.1 (Guindon, et al. 2010).

# Results

## **Pfam domains describe TE superfamilies**

By annotating RepBase reference sequences with the Pfam domains, we have built a high quality dictionary for discerning TE superfamilies based solely on their encoded protein domains composition. Such approach is especially feasible for large scale analyses by allowing automatic classification of extensive datasets. However, Sola, Zator, Kolobok, Novosib, Academ, Dada and KDZ RNase H-like domains lack detectable sequence similarity to any Pfam domain. For these superfamilies, described in the last decade, we built separate, dedicated protein sequence profiles based on sequences (KDZ, Dada) and, when available, multiple sequence alignments provided in literature (Sola1-3, Zator, Kolobok, Academ). Novosib RNase H-like domain was proposed by Yuan and Wessler (Yuan and Wessler 2011) with caution due to limited representation and, in consequence, uncertainties in profile building. In our study, the presence of RNase H for Novosib elements was not confirmed and hence, we did not validate the identified Novosib with Pfam domains. Interestingly, Ginger TEs (Bao, et al. 2010) encode transposase related to retrotransposon and retroviral integrases (rve, PF00665) and therefore, we selected rve domain as a determinant for this TE superfamily. [Table 1](#) presents the assignment of protein domains to DNA TE superfamilies in Fungi.

## **TE abundance**

7,411,508 DNA TE fragments, 216,933 of which contain a potentially active DDE transposase domain (coordinates of identified DNA TEs with a transposase are available by the corresponding author upon request), have been identified in 1,728 of the analysed 1,730 genomes (see [Supplementary Table S1b](#)). [Table 1](#) contains the following information: which transposon superfamilies were previously described in Fungi, how many RepBase references were available on 23th of March 2017 and also our findings. In 2011, Yuan and Wessler (Yuan and Wessler 2011) reported presence of 8 superfamilies of DNA TEs in Fungi, whereas RepBase had references for 14 superfamilies in March 2017 (two of those

were described after 2011). The differences between RepBase and Yuan & Wessler are limited to single occurrences of understudied elements. Great differences in TE abundance can be noticed both from DNA TE superfamily perspective and from fungal taxonomy side ([Figure 2](#)).

Sola, Zator, Dada and P elements, each with less than a thousand representatives encoding DDE transposase, constitute the least abundant DNA TE superfamilies in the dataset, whereas Tc1/Mariner with more than 4,414 thousand copies, 93 thousand of which contain a DDE transposase, are the most ubiquitous in the analysed dataset. However, these TEs are extraordinarily abundant not only in Fungi. EnSpm and PiggyBac have limited taxonomic distribution and seem to remain in a few fungal taxa only. Merlin and PiggyBac have comparable taxonomic distributions, the former being present in most basal lineages, while the latter prevailing in terrestrial fungi (Mucoromycotina and Ascomycota). Transposons with patchy distribution are likely products of horizontal transposon transfer (HTT) or multiple loss history (Wallau, et al. 2012). P, Sola1-3 and Zator were identified only in a handful of isolates and likely have been acquired *via* HTT. Only one DNA TE superfamily, Transib, is missing from Fungi. One might define fungal core DNA TE dataset as a composition of Tc1/Mariner, Ginger, hAT, PIF/Harbinger, MuLE and Kolobok elements often accompanied by CMC/EnSPM elements.

### **Genome size, genome defense and non-coding genome**

There is a moderate correlation ( $r=0.6$ ,  $p=3.5 \times 10^{-178}$ ) between total TE abundance and genome size ([Figure 3](#)). The correlation between genome size and TE content is present for both functional and remnant copies ( $r=0.64$ ). Big genomes tend to be rich in multiple types of mobile elements at once. However, the correlation is significantly higher, when phyla are analysed separately, reaching  $r=0.98$  for Mucoromycotina in its extreme ( $p=4.1 \times 10^{-51}$ ). This discrepancy can be explained by the huge diversity of Fungi resulting from ancient lineage separation – the divergence of main fungal lineages predates land colonisation (Berbee and Taylor 2010) what could have led to different patterns of genome architecture evolution between fungal phyla. The abundance of Tc1/Mariner, the most successful superfamily of DNA TEs in Fungi, correlates well with genome size ( $r=0.75$ ,  $p=4.3 \times 10^{-58}$  for Basidiomycota and  $r=0.94$  ,

$p=2.0 \times 10^{-34}$  for Mucoromycota). Both Basidiomycota and Mucoromycotina span species with big genomes, display a high fraction of repeats and considerable diversity of DNA TEs. In contrast, rare TE superfamilies display weaker, if any, correlation with genome sizes, which can be a derivative of limited taxonomic sampling, especially when basal fungal lineages are considered. The distribution of elements per genome for each superfamily is summarised in [Supplementary Figure S1](#).

A question arises, whether TE occurrence is correlated in any way with an internal genome structure. Intron abundance counted per gene, as well as averaged gene distance, display only weak correlation with TE number ( $r=0.28$ ,  $p=1.7 \times 10^{-13}$  and  $r=0.33$ ,  $p=3.3 \times 10^{-20}$ , respectively; [Supplementary Figures S2 and S3](#), respectively) what might suggest that more elaborate factors impact DNA TE proliferation. GC content had been previously considered as one of the key determinants of TE insertion preference, since AT rich regions are favoured by some TE types and are often more accessible for transposase (Sultana, et al. 2017). Our results indeed demonstrate an elevated capacity of AT-rich genomes to host DNA TEs ( $r=0.12$ ,  $p=1.1 \times 10^{-13}$ , [Supplementary Figure S4](#)). The discussed genome characteristics explain only partially the observed richness of DNA TEs and suggest that DNA TE abundance is not only a result of massive accumulation of sequences in an inert fashion, but rather a complex process with different influencing factors.

The population of mobile elements in a genome is a derivative of two opposing forces: a proliferative capacity of the transposable elements themselves and defensive capability of the host to eliminate these potentially aberrative factors. The mechanisms of mobile element elimination in fungi are not well studied. Transposons are continuously eliminated by ectopic recombination and deletions (Feschotte and Pritham 2007). Some of the genome defense mechanisms depend on meiosis while others do not. Sex in fungi is a complex phenomenon since many fungal pathogens are clonal, some are obligatory biotrophs; others are selfing, some are outcrossing (Heitman, et al. 2013). In order to juxtapose genome defense mechanisms with the observed DNA TE distribution, we assessed the conservation of the core elements of RNAi associated genes (involved in MSUD, SIS and *quelling*) as well as Masc1/RID. Canonical RNAi components are missing in the analysed Saccharomycotina (some Saccharomycotina use

non-cannonical Dicer proteins to generate small interfering RNAs to silence TEs (Drinnenberg, et al. 2009)), *Cryptococcus gattii* (Tremellomycetes), all *Pneumocystis* (Taphrinomycotina) and most of Microsporidia, or are incomplete in some species across all main fungal lineages (Billmyre, et al. 2013). Surprisingly, three representatives of Microsporidia (*Vavraia culicis* subsp. *floridensis*, *Vittaforma corneae* and *Nosema ceranae*) retain all analysed RNAi components, despite having significantly reduced genomes. What is somehow characteristic for RNAi machinery, its components may occur in multiple copies (paralogues), especially in symbiotic fungi: Glomeromycotina and Agaricomycotina. For instance, *R. irregularis* DAOM 197198w has 30 Argonaute, two Dicer-like proteins and 10 RdRP paralogues, while *R. irregularis* DAOM 181602 has 16 Argonaute paralogues. For comparison, these fungi possess more Argonaute and RdRP paralogues than *Arabidopsis thaliana* (Argo: 14, Dicer: 8, RdRP: 6) and *A. lyrata* (Argo: 10, Dicer: 6, RdRP: 7) have according to FunRNAi database (Choi, et al. 2014). Accessions for all identified RNAi components and Rid/Masc homologues are available in [Supplementary File S3](#).

Fungi that retain both RIP and RNAi systems have more total DNA TE sequences but less elements retaining a transposase ([Figure 4](#)). The abundance of fragmented copies is not impacted by the presence of defense mechanisms what shows that MITES and other remnant fragments accumulate in a stochastic fashion due to drift. Nonetheless, our results may appear to be only approximative, because Masc1/RID presence is neither a direct proof nor a sufficient condition of RIP activity (Hane, et al. 2014), but is a requirement for RIP only. Fungi lacking both RIP and RNAi systems have the lowest number of elements. This category, however, covers the majority of organisms with reduced genomes (e.g. parasites), which are often almost devoid of repetitive content and lack some of the core defense mechanisms as well. These genomes are under particular selection pressure for genome compactness, what prevents fixation of nearly neutral parasitic sequences. Masc1/RID homologues were identified in Pezizomycotina, two Taphrinomycotina and five Agaricomycotina representatives; a highly similar methylase was also present in *Batrachochytrium dendrobatidis* ([see Supplementary Figure 5](#)). However, the latter is more similar to Bacillus sequences than to fungal Masc1/RID with high sequence identity (70-

80%), which points at a likely HGT from bacteria to *B. dendrobatidis* and an unlikely Masc1/RID function.

## Old fungal lineages harbour diverse DNA transposons

The oldest fungal lineages: Cryptomycota, Microsporidia, Chytridiomycota and Blastocladiomycota mostly host very few DNA transposons. TE reduction and elimination is expected for obligate parasites with compact genomes, like Microsporidia (Parisot, et al. 2014). However, not all of them have reduced genomes and, in consequence, reduced the DNA TE repertoire. For instance, *Anncaliia algerae* is known to possess a high-repeat genome with more than 240 transposon families (Parisot, et al. 2014). Apart from *A. algerae*, our dataset included a handful of non-reduced Microsporidia genomes from *Hamiltosporidium tvaerminnensis*, *Nosema bombycis* and *Pseudoloma neurophilia*, which were also colonised by diverse mobile elements. Parisot and colleagues report Merlin, Tc/Mariner and piggyBac to be present in *A. algerae* genome, a list significantly extended in this work by hAT, MuLE, Ginger and Academ. On the other hand, Microsporidia with strongly reduced genomes from Nematocida and Encephalitozoon genera harboured less than ten reliably predicted transposon candidates per genome.

Cryptomycota, represented in the database by the only sequenced species, *Rozella allomycis*, have single representatives of 7 out of 18 TE superfamilies considered in this study. However, taking into account the huge variability within high rank taxons, genome sequencing quality and the stringent criteria applied in this project, much more TEs might be encoded by Cryptomycota members. *Rozella* has transposons from Tc/Mariner, Merlin, hAT, MuLE, Academ, Ginger and PIF/Harbinger superfamilies.

Chytridiomycota, represented by five distant taxa (*Spizellomyces punctatus*, *Batrachochytrium dendrobatidis*, *Homoloaphlyctis polyrhiza*, *Synchytrium endobioticum* and *Gonapodya prolifera*), and Blastocladiomycota with only one representative, *Allomyces macrogynus*, contain very few DNA TEs, if any. Their DNA TE composition is similar to Microsporidial. However, there are several peculiarities observed within this quite limited dataset: an expansion of MuLE in *H. polyrhiza*, an expansion of PIF-Harbinger in *G. prolifera* and single Kolobok elements in *Allomyces* and *Spizellomyces*. The observed

low TE content is surprising in the context of big genomes of *A. macrogynus* and *G. prolifera*, which contain more than 3 thousand remnant elements, but only 52 and 177 transposons with a transposase coding region, respectively. The absence of active elements and presence of only 335 remnant copies in the *Synchytrium endobioticum* genome is a possible effect of the spectacular genome reduction to the size of 2Mb.

The complex landscape of DNA TE families both in Microsporidia and Rozella, regardless of their compact genomes, suggest a great diversity of DNA TE repertoire in the last common fungal ancestor. This is consistent with the presence of multiple DNA TE superfamilies also in the majority of other eukaryotic lineages (Pritham 2009).

### **Terrestrial fungi vary in their DNA TE composition**

#### Mucoromycota

Mucoromycotina, Mortierellomycotina and Glomeromycotina, grouped together in Mucoromycota (Spatafora, et al. 2016), have genomes, which contain different sets of DNA TE superfamilies. Glomeromycotina have big genomes, each with more than 80,000 copies of DNA TEs representing majority of superfamilies analysed (except for Dada, Merlin and Sola2). *Rhizophagus irregularis* strains are unique sequenced Glomeromycotina representatives here and the only organisms to harbour additionally Sola3 and Zator elements. In contrast, Mortierellomycotina DNA TE composition is very limited, as compared to Glomeromycotina, with about four thousand remnant copies and 59 containing transposase for *Mortierella alpina* up to 165 transposons with DDE transposase in *M. elongata* representing only Academ, hAT, Kolobok, MuLE, PIF-Harbinger, PiggyBac and Tc1/Mariner (out of 18 superfamilies). Mucoromycotina, the most abundant Mucoromycota in our dataset, constitute a more heterogeneous taxon in terms of DNA TE spectrum. They are relatively TE-rich; they lack only IS3EU, P, Sola1, Sola2, Sola3 and Zator elements. *Rhizopus delmar* genome underwent whole genome duplication and has been described as abundant in transposons (Ma, et al. 2009). However, there are Mucoromycotina (Umbelopsidales) with compact genomes and DNA TE composition resembling that of



Mortierellomycotina, what clearly divides Mucoromycota into three clusters regarding DNA TE abundance (**Figure 3C**): the first one is formed by Mortierellomycotina and Umbelopsidales bearing few elements, the second cluster groups repeat abundant *Rhizophagus irregularis* together with some Mucorales (e.g. *Mucor racemosus* B9645 and *Rhizopus microsporus* B9738) and the last cluster spans the remaining Mucorales with intermediate TE abundance. Umbelopsidales are a basal group within Mucoromycotina (Spatafora, et al. 2016) with streamlined genomes compared to Mucorales.

### Zoopagomycota

Zoopagomycota, formed by Entomophthoromycotina, Kickxellomycotina and Zoopagomycotina (Spatafora, et al. 2016), are represented in NCBI database by two Basidiobolus, two Conidiobolus and a single Kickxellomycotina genome (*Capniomyces stellatus*), all of which differ in their DNA TE composition. The former are very TE-rich with huge expansions of CMC-EnSpm and PIF-Harbinger elements, whereas the remaining *C. stellatus* and *Conidiobolus coronatus* possess Tc1/Mariner and few other transposons only. Basidiobolus and *Conidiobolus incongruus* genomes are TE-rich whereas *C. coronatus* and *C. stellatus* have up to 30 elements with a transposase.

### Dikarya

Dikarya constitute the majority of the analysed dataset with taxonomic sampling best of all major branches. Among Basidiomycota, two contrasting genome architectures can be distinguished: compact, low repeat Ustilaginomycotina (with different species of *Malassezia* with less than 10 active elements), Microbotryomycetes with a handful of transposons and Pucciniomycetes with big genomes with many repeat proliferations (with the extreme case of *Uromyces viciae-fabae* with 793 Ginger, 1,422 hAT, 1339 MuLE and 3,146 PIF-Harbinger elements). Agaricomycotina groups Tremellomycotina with less than a hundred elements per genome and Agaricomycetes with up to a thousand active elements. In Agaricomycetes, Zisupton, CMC/EnSpm and PIF-Harbinger elements are significantly more abundant. Also, Ascomycota genome architectures vary significantly. Most members of Saccharomycetes have less than 20 copies with a transposase domain from 3-4 superfamilies. Taphrinomycotina have very few TEs

from only 5 superfamilies (*Taphrina wiesneri* has only 9 DNA TEs with a transposase and *T. defromans* has 127 such elements). Pezizomycotina groups taxa with thousands of DNA TEs such as *Erisiphae pisi*, *Tuber melanosporum* and *Pseudogymnoascus destructans* M1379 and Orbiliomycetes with a handful of elements. Surprisingly, Dothideomycetes, which group plant-associated fungi with big genomes, have few DNA TEs.

## Transposon abundance and fungal lifestyle

This rich dataset enabled us to test hypotheses related to lifestyle and TE content relationships (Figure 5). Each fungus was assigned to general ecological categories (except for undescribed taxa). Statistical analyses showed that plant-related fungi are significantly more prone to DNA TE accumulation, with plant symbionts being the most extreme. PIF/Harbinger and hAT distributions show preference for fungi living with plants (Supplementary File S2). Plant pathogens from Pucciniomycotina harbour huge multiple transposon expansions, among them *Uromyces vicia-fabae*, *Melampsora laricis-populina* and a variety of Puccinia species.

Most of animal-associated fungi have compact genomes with few DNA TEs, however, there is a number of fungi which escape this rule. Ascomycota from Sordariomycetes associated with insects (*Hirsutella*, *Metarhizium*, *Ophiocordyceps* and *Pochonia*), vertebrate-associated Onygenales (*Ajellomyces capsulatus* and *Paracoccidioides*), *Pseudogymnoascus destructans* and opportunistic human pathogen *Curvularia lunata* show high TE abundance. Fungi colonising an animal host have less PIF/Harbinger copies per genome than other fungi. The host (plant vs animal) seems to be a major factor influencing TE abundance while the detailed type of relationship with the host plays a secondary role. In general, pathogens tend to have less DNA TEs than non-pathogenic taxa. Saprotrophy and soil/dung habitat are positively correlated with DNA TE content. However, relationships between pathogenic and saprotrophic lifestyles and TE abundance are weak and do not hold when multiple lifestyle and taxonomic factors are analyzed together (Supplementary File S2). This lack of strong support might be a consequence of used categories' intrinsic dependence e.g. pathogenic fungi are often animal-related and the animal feature has

a stronger statistical signal. Many of the saprotrophic fungi analysed are related to a plant host or live in a soil/dung habitat, which seem to be the major factors shaping TE abundance. Academ elements are more common in saprotrophs than in other fungi. Obligate parasites with contracted genomes are expected to have few elements, what is particularly clear for Microsporidia.

## Discussion

The aim of this study was to assess the abundance, state of conservation and distribution of main superfamilies of transposons encoding RNase H-like transposase in the context of fungal taxonomy, genome architecture and fungal life strategies. Applying a semi-automated approach enabled us to recover most of the already known mobile elements and to detect many previously unreported families and copies for the whole set of 1,730 genomes, what significantly expands the contemporary RepBase collection of repeats. The RepeatMasker RepBase edition has about ten thousand DNA TE references, while RepBase website hosts 536 fungal DNA TE loci. We identified almost 70 thousands of distinct families (after clustering at 80% sequence identity threshold). Pfam domains can be successfully used to automatically classify the identified, potentially active members of most of DNA transposon superfamilies, what is especially feasible for high throughput genome annotation.

One might consider limiting the dataset to assembled genomes as leading to underestimation of TE abundance, what is particularly justified for repeat-rich genomes of animals. Genome size and architecture differs between fungi and animals, and within fungi themselves. A typical fungus has 1-15% of repetitive content (Cuomo and Birren 2010). Many fungi, e.g. *Candida albicans* have compact genomes with few introns. Also the typical genome size is relatively small with an average close to 30Mb with just as much as few genomes being bigger than 60Mb. In consequence, the majority of assemblies in fungi are close to complete, even if fragmented in centromeric and telomeric parts. A study concerning unassembled reads might marginally enrich the set of identified DNA TE and decrease the impact of the assembly quality on repeat-rich genomic regions. However, working with unassembled reads increases dataset size making it computationally inefficient and infeasible on a thousand genomes scale. We

observe variation between assemblies of closely related taxa, which might be either a difference present in the population or assembly quality artefacts. Also, there were studies with several genomes of one species sequenced with non-NGS methods showing big differences in repetitive content between strains yet resulting not solely from sequencing coverage differences but also from diversity at population level (Neafsey, et al. 2010). There is still space for improvement in TE annotation quality resulting both from limited curated TE references and sequence data quality. Transposon annotation using raw reads might help to overcome this limitation. There is also a need for expert curated sequence profiles for non-animal taxa like currently available in Dfam (Hubley, et al. 2016). Results obtained in the course of this project significantly expand the record of fungal transposons which might be included in reference databases.

DNA TE superfamilies have an ancient origin, what is supported by their presence in both old fungal and other eukaryotic lineages (Yuan and Wessler 2011). Already Feschotte and Pirtham in 2007 stated that major types of DNA transposons predate the divergence of Eukaryotes and remain present in distant evolutionary groups (Feschotte and Pritham 2007). Therefore, TEs are very likely to be transmitted vertically in the majority of cases and in the absence of strong selection should remain at least in low copy numbers in many lineages. There are of course documented cases of HTT observed for various TE families (Dotto, et al. 2015) but in the light of works cited above we consider them as a phenomenon rather than a rule. In consequence of vertical transmission, DNA TE taxonomic distribution should be almost continuous when closely related taxa are compared. Our finding expands previous assessments of taxonomic distribution for many of DNA TE families within fungi, what was expected with the advancement of sequencing and broader taxon sampling.

It has been previously shown that fungal genome size positively correlates with a number of hosted transposon families (Elliott and Gregory 2015), at least for genomes smaller than 500 Mbp, above which more complex rules apply and the genome inflates without significant enrichment in TE abundance. Our results are in concordance with the aforementioned phenomenon; the taxa with the biggest genome sizes, such as Rhizophagus, Pucciniomycetes, Mucorales, tend to have more diverse elements and more transposon copies. However, genome size may not be considered as the single factor

influencing TE abundance. Parameters describing genome compactness, such as gene density and number of introns per gene, correlate very weakly with DNA TE abundance, what suggests that complexity assessed that way does not necessarily grow with transposon number or perhaps these features, i.e. gene density and number of introns per gene are simply not meaningful measures of genome complexity. The observed DNA TE distribution and their unexpected abundance in the analysed genomes could serve as one of the arguments for considering some of the fungal genomes (e.g. *Tuber melanosporum*, *Rhizophagus irregularis*, *Rhizopus delmar*) as following the fate of complex eukaryotic genomes that underwent secondary restructurations after passive accumulation of mobile elements and genome inflation as proposed by Lynch and Conery (Lynch and Conery 2003). Our identification of independent expansions of diverse TE types in distant fungal lineages is in agreement with the observation that TE rich regions contribute to genomic complexity in plant associated fungi (Moller and Stukenbrock 2017).

There have been several attempts to link repetitive content with organism's lifestyle both in fungi (*Mycosphaerella graminicola* (Goodwin, et al. 2011), *Nectria haematococca* (Coleman, et al. 2009), *Leptosphaeria maculans* (Rouxel, et al. 2011)) and in fungi-like organisms (*Phytophthora infestans* (Haas, et al. 2009)), especially in the light of adaptations to pathogenicity. TE insertion patterns differ between strains within the same species e.g. *Magnaporthe grisea* (Shirke, et al. 2016), or between closely related species e.g. *Ustilago maydis* and *Sporisorium scitamineum* (Dutheil, et al. 2016), differing in host specificity. Currently, it is broadly recognised that TEs are an important source of regulatory sequences for host genes and shape the genomic landscape for coding sequences (Rebollo, et al. 2012). Our results show that there is an overrepresentation of DNA TEs in plant-related fungi, which have bigger genomes anyway. The expansion of transposable elements had been previously shown to be especially noticeable in plant-symbiotic fungi (Hess, et al. 2014) and plant pathogens (Raffaele and Kamoun 2012), all confirmed in our big-scale analyses. It seems that recurrent adaptation to symbiosis involves not only a reduced number of plant cell wall-degrading enzymes (Martin, et al. 2016), molecular crosstalk with the host and small secreted proteins (van der Heijden, et al. 2015), but also, at least in some cases, relaxed genome control against duplications, mobile element proliferation and overall genome size growth. We

might speculate that symbiosis is so evolutionary challenging to the fungus that: i) the cost of maintaining strict genome defense is too high, and ii) TEs are mobilized and proliferate like in other stressful conditions. In consequence of increased TE mobility there is a new raw material for selection, provided by an expanded genome, which is highly appreciated. Some TEs are known to provide small non-coding RNAs for RNAi, e.g. LTR retransposons in *Magnaporthe oryzae* (LTR-siRNAs) (Nunes, et al. 2011), and therefore seem to be a perfect source of components for developing elaborate regulatory networks. There are examples of gene cluster regulation by neighbouring TEs e.g. the penicillin cluster in *Aspergillus nidulans* has lower expression in the absence of Pbla element (Shaaban, et al. 2010). Also, the presence of defense mechanisms shows a positive correlation with TE abundance, what could be naively explained by the fact that organisms with both systems (RNAi and RIP) have more raw material for selection from inactive TEs with potential promoter sequences, TF binding sites and other reusable modules with a minimised risk of deleterious TE activity resulting from TE excision and insertions. This genome inflation due to accumulation of TE fragments might be a result of genetic drift and limited effective population size (Moller and Stukenbrock 2017). Additionally, an increasing genome complexity was found to correlate with an extended range of hosts and was reported from plant-related fungi (van der Heijden, et al. 2015).

The abundance of DNA TE varies in different fungal lineages. Our results show that already Cryptomycota had most of the core fungal transposon dataset, however, the more complete repertoire of DNA TE emerged in land fungi (Mucoromycotina and Glomeromycotina). *Rhizophagus irregularis*, an exclusive representative of Glomeromycotina in the dataset, has peculiar expansions both in the diversity and in abundance of DNA TE in parallel with multiplication of genes involved in RNAi pathways. A greater number of sequenced Glomeromycotina genomes will reveal whether this is a single case or perhaps a feature linked to the evolutionary group and arbuscular mycorrhiza formation. Meantime, the whole fungal lineages recursively streamlined their genomes and multiple TE types became lost, e.g. in Wallemiomycetes, Mixiomycetes, Microbotryomycetes, Lecanoromycetes, Taphrinomycotina, Saccharomycotina, Schizosaccharomycetes and possibly Microsporidia, which altogether with HTTs

explains the observed patchy distributions of several DNA TE superfamilies. HTT events were documented for Mariner, hAT and P elements (Dotto, et al. 2015), and we speculate here that Sola1-3 and Zator elements might have been transferred as well. We found Sola3 and Zator elements only in *Rhizophagus irregularis*. The rare occurrence of these repeats justifies hypotheses on their emergence in Fungi through HTT rather than vertical descendance followed by multiple losses (Silva, et al. 2004; Wallau, et al. 2012).

Our findings point to several previously unreported correlations between transposon abundance and fungal life strategies, which might provide an inspiration for further studies at the intersection of environmental and molecular biology. Undeservedly, transposable elements are still an understudied part of eukaryotic genomes but their vital roles in shaping the life's complexity are beginning to be understood.

## **ACKNOWLEDGEMENTS**

We thank Julia Pawlowska and Marcin Grynberg for their insightful comments and suggestions.

## **FUNDING**

This work was supported by National Science Centre (2012/07/D/NZ2/04286 to AM and 2014/15/B/NZ1/03357 to KG), Foundation for Polish Science (TEAM to KG), and Ministry of Science and Higher Education scholarship for outstanding young researchers to AM and KS.

## **CONFLICT OF INTEREST**

No conflict of interest declared.

## **AUTHOR CONTRIBUTIONS**

A.M. designed the study, A.M. and K.S. prepared the dataset, implemented software and performed genome analyses, M.S-D. performed statistical analyses, A.M., K.S. and K.G. interpreted the data and wrote the manuscript.





## REFERENCES

- Abrusán G, Zhang Y, Szilágyi A 2013. Structure prediction and analysis of DNA transposon and LINE retrotransposon proteins. *J. Biol. Chem.* 288: 16127-16138. doi: 10.1074/jbc.M113.451500
- Bao W, Jurka MG, Kapitonov VV, Jurka J 2009. New superfamilies of eukaryotic DNA transposons and their internal divisions. *Mol. Biol. Evol.* 26: 983-993. doi: 10.1093/molbev/msp013
- Bao W, Kapitonov VV, Jurka J 2010. Ginger DNA transposons in eukaryotes and their evolutionary relationships with long terminal repeat retrotransposons. *Mob. DNA* 1: 3. doi: 10.1186/1759-8753-1-3
- Barry C, Faugeron G, Rossignol JL 1993. Methylation induced premeiotically in *Ascobolus*: coextension with DNA repeat lengths and effect on transcript elongation. *Proc. Natl. Acad. Sci. U. S. A.* 90: 4557-4561.
- Berbee ML, Taylor JW 2010. Dating the molecular clock in fungi – how close are we? *Fungal Biol. Rev.* 24: 1-16. doi: 10.1016/j.fbr.2010.03.001
- Billmyre RB, Calo S, Feretzaki M, Wang X, Heitman J 2013. RNAi function, diversity, and loss in the fungal kingdom. *Chromosome Res.* 21: 561-572. doi: 10.1007/s10577-013-9388-2
- Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 25: 1972-1973. doi: 10.1093/bioinformatics/btp348
- Carbone L, et al. 2014. Gibbon genome and the fast karyotype evolution of small apes. *Nature* 513: 195-201. doi: 10.1038/nature13679
- Chang S-S, Zhang Z, Liu Y 2012. RNA Interference Pathways in Fungi: Mechanisms and Functions. *Annu. Rev. Microbiol.* 66: 305-323. doi: 10.1146/annurev-micro-092611-150138
- Choi J, et al. 2014. funRNA: a fungi-centered genomics platform for genes encoding key components of RNAi. *BMC Genomics* 15 Suppl 9: S14. doi: 10.1186/1471-2164-15-s9-s14
- Coleman JJ, et al. 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. *PLoS Genet.* 5: e1000618. doi: 10.1371/journal.pgen.1000618
- Cuomo CA, Birren BW 2010. The fungal genome initiative and lessons learned from genome sequencing. *Methods Enzymol* 470: 833-855. doi: 10.1016/S0076-6879(10)70034-3
- S0076-6879(10)70034-3 [pii]
- Daboussi M-J, Marie-Josée D, Pierre C 2003. Transposable Elements in Filamentous Fungi. *Annu. Rev. Microbiol.* 57: 275-299. doi: 10.1146/annurev.micro.57.030502.091029
- Darriba D, Taboada GL, Doallo R, Posada D 2011. ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* 27: 1164-1165. doi: 10.1093/bioinformatics/btr088
- Dotto BR, et al. 2015. HTT-DB: horizontally transferred transposable elements database. *Bioinformatics* 31: 2915-2917. doi: 10.1093/bioinformatics/btv281
- btv281 [pii]
- Drinnenberg IA, et al. 2009. RNAi in budding yeast. *Science* 326: 544-550. doi: 10.1126/science.1176945
- Dutheil JY, et al. 2016. A Tale of Genome Compartmentalization: The Evolution of Virulence Clusters in Smut Fungi. *Genome Biol. Evol.* 8: 681-704. doi: 10.1093/gbe/evw026
- Elliott TA, Gregory TR 2015. Do larger genomes contain more diverse transposable elements? *BMC Evol. Biol.* 15: 69. doi: 10.1186/s12862-015-0339-8
- Erwin JA, Marchetto MC, Gage FH 2014. Mobile DNA elements in the generation of diversity and complexity in the brain. *Nat. Rev. Neurosci.* 15: 497-506. doi: 10.1038/nrn3730
- Feschotte C, Pritham EJ 2007. DNA transposons and the evolution of eukaryotic genomes. *Annu. Rev. Genet.* 41: 331-368. doi: 10.1146/annurev.genet.40.110405.090448
- Frickey T, Lupas A 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics* 20: 3702-3704. doi: 10.1093/bioinformatics/bth444

Fu L, Niu B, Zhu Z, Wu S, Li W 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28: 3150-3152. doi: 10.1093/bioinformatics/bts565

Ginalski K, von Grotthuss M, Grishin NV, Rychlewski L 2004. Detecting distant homology with Meta-BASIC. *Nucleic Acids Res.* 32: W576-581. doi: 10.1093/nar/gkh370

Gladyshev E, Kleckner N 2016. Recombination-independent recognition of DNA homology for repeat-induced point mutation. *Curr. Genet.* doi: 10.1007/s00294-016-0649-4

Goodwin SB, et al. 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet.* 7: e1002070. doi: 10.1371/journal.pgen.1002070

Grow EJ, et al. 2015. Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells. *Nature* 522: 221-225. doi: 10.1038/nature14308

Guindon S, et al. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Syst. Biol.* 59: 307-321. doi: 10.1093/sysbio/syq010

Haas BJ, et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461: 393-398. doi: 10.1038/nature08358

Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP. 2014. Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. In: *Fungal Biology*. p. 55-68.

Heitman J, Sun S, James TY 2013. Evolution of fungal sexual reproduction. *Mycologia* 105: 1-27. doi: 10.3852/12-253

12-253 [pii]

Hess J, et al. 2014. Transposable element dynamics among asymbiotic and ectomycorrhizal *Amanita* fungi. *Genome Biol. Evol.* 6: 1564-1578. doi: 10.1093/gbe/evu121

Hubley R, et al. 2016. The Dfam database of repetitive DNA families. *Nucleic Acids Res* 44: D81-89. doi: 10.1093/nar/gkv1272

gkv1272 [pii]

Iyer LM, et al. 2014. Lineage-specific expansions of TET/JBP genes and a new class of DNA transposons shape fungal genomic and epigenetic landscapes. *Proc. Natl. Acad. Sci. U. S. A.* 111: 1676-1683. doi: 10.1073/pnas.1321818111

Kapitonov V, Jurka J 2010. Academ—a novel superfamily of eukaryotic DNA transposons. *Rebase Reports* 10: 643.

Kapitonov V, Jurka J 2007. Kolobok, a novel superfamily of eukaryotic DNA transposons. *Rebase Rep* 7.

Kapitonov VV, Jurka J 2006. Novosib-1\_CR, a family of autonomous Novosib transposons from the green algae genome. *Rebase Reports* 6(5): 262-262

Kapitonov VV, Jurka J 2008. A universal classification of eukaryotic transposable elements implemented in Rebase. *Nat. Rev. Genet.* 9: 411-412; author reply 414. doi: 10.1038/nrg2165-c1

Kluyver T, et al. 2016. Jupyter Notebooks—a publishing format for reproducible computational workflows. *Positioning and Power in Academic Publishing: Players, Agents and Agendas*: 87.

Kojima KK, Jurka J 2013. A superfamily of DNA transposons targeting multicopy small RNA genes. *PLoS one* 8: e68260. doi: 10.1371/journal.pone.0068260

Koonin EV, Krupovic M 2015. Evolution of adaptive immunity from transposable elements combined with innate immune systems. *Nat. Rev. Genet.* 16: 184-192. doi: 10.1038/nrg3859

Liu K, Wessler SR 2017. Transposition of Mutator-like transposable elements (MULEs) resembles hAT and Transib elements and V(D)J recombination. *Nucleic Acids Res.* doi: 10.1093/nar/gkx357

Lynch M, Conery JS 2003. The origins of genome complexity. *Science* 302: 1401-1404. doi: 10.1126/science.1089370

Ma L-J, et al. 2009. Genomic Analysis of the Basal Lineage Fungus *Rhizopus oryzae* Reveals a Whole-Genome Duplication. *PLoS Genet.* 5: e1000549. doi: 10.1371/journal.pgen.1000549

Majorek KA, et al. 2014. The RNase H-like superfamily: new members, comparative structural analysis and evolutionary classification. *Nucleic Acids Res.* 42: 4160-4179. doi: 10.1093/nar/gkt1414

Makarevitch I, et al. 2015. Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet.* 11: e1004915. doi: 10.1371/journal.pgen.1004915

Manning VA, et al. 2013. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3* 3: 41-63. doi: 10.1534/g3.112.004044

Martin F, Kohler A, Murat C, Veneault-Fourrey C, Hibbett DS 2016. Unearthing the roots of ectomycorrhizal symbioses. *Nat. Rev. Microbiol.* 14: 760-773. doi: 10.1038/nrmicro.2016.149

Mistry J, Finn RD, Eddy SR, Bateman A, Punta M 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* 41: e121. doi: 10.1093/nar/gkt263

Moller M, Stukenbrock EH 2017. Evolution and genome architecture in fungal plant pathogens. *Nat Rev Microbiol.* doi: 10.1038/nrmicro.2017.76  
nrmicro.2017.76 [pii]

Neafsey DE, et al. 2010. Population genomic sequencing of *Coccidioides* fungi reveals recent hybridization and transposon control. *Genome Res* 20: 938-946. doi: 10.1101/gr.103911.109  
gr.103911.109 [pii]

Nunes CC, et al. 2011. Diverse and tissue-enriched small RNAs in the plant pathogenic fungus, *Magnaporthe oryzae*. *BMC Genomics* 12: 288. doi: 10.1186/1471-2164-12-288

Parisot N, et al. 2014. Microsporidian genomes harbor a diverse array of transposable elements that demonstrate an ancestry of horizontal exchange with metazoans. *Genome Biol. Evol.* 6: 2289-2300. doi: 10.1093/gbe/evu178

Pritham EJ 2009. Transposable elements and factors influencing their success in eukaryotes. *J. Hered.* 100: 648-655. doi: 10.1093/jhered/esp065

Raffaele S, Kamoun S 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. *Nat. Rev. Microbiol.* 10: 417-430. doi: 10.1038/nrmicro2790

Rebollo R, Romanish MT, Mager DL 2012. Transposable elements: an abundant and natural source of regulatory sequences for host genes. *Annu. Rev. Genet.* 46: 21-42. doi: 10.1146/annurev-genet-110711-155621

Rey O, Danchin E, Mirouze M, Loot C, Blanchet S 2016. Adaptation to Global Change: A Transposable Element–Epigenetics Perspective. *Trends Ecol. Evol.* 31: 514-526. doi: 10.1016/j.tree.2016.03.013

Rountree MR, Selker EU 2010. DNA methylation and the formation of heterochromatin in *Neurospora crassa*. *Heredity* 105: 38-44. doi: 10.1038/hdy.2010.44

Rouxel T, et al. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nat. Commun.* 2: 202. doi: 10.1038/ncomms1189

Selker EU, Jensen BC, Richardson GA 1987. A portable signal causing faithful DNA methylation de novo in *Neurospora crassa*. *Science* 238: 48-53.

Shaaban M, et al. 2010. Involvement of transposon-like elements in penicillin gene cluster regulation. *Fungal Genet. Biol.* 47: 423-432. doi: 10.1016/j.fgb.2010.02.006

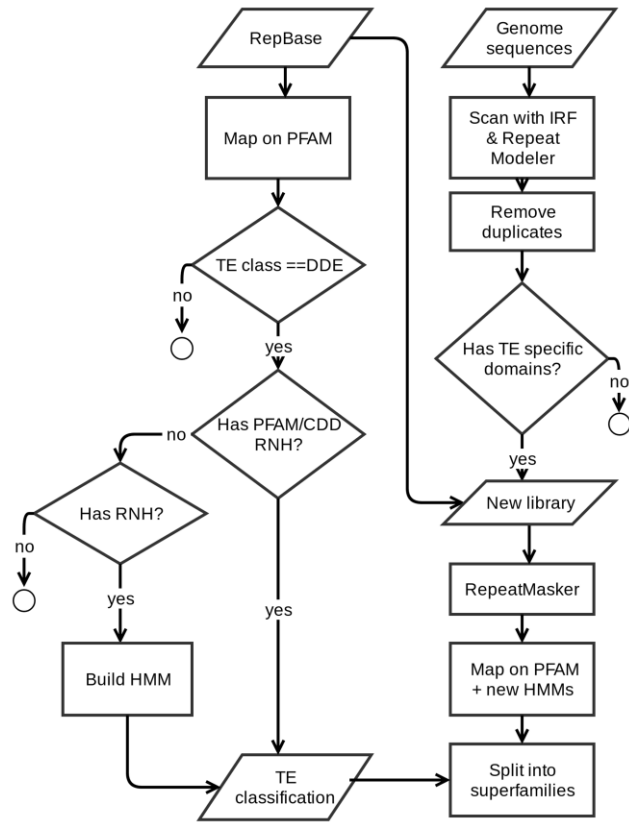
Shirke MD, Mahesh HB, Gowda M 2016. Genome-Wide Comparison of *Magnaporthe* Species Reveals a Host-Specific Pattern of Secretory Proteins and Transposable Elements. *PloS one* 11: e0162458. doi: 10.1371/journal.pone.0162458

Shiu PKT, Raju NB, Zickler D, Metzberg RL 2001. Meiotic Silencing by Unpaired DNA. *Cell* 107: 905-916. doi: 10.1016/s0092-8674(01)00609-2

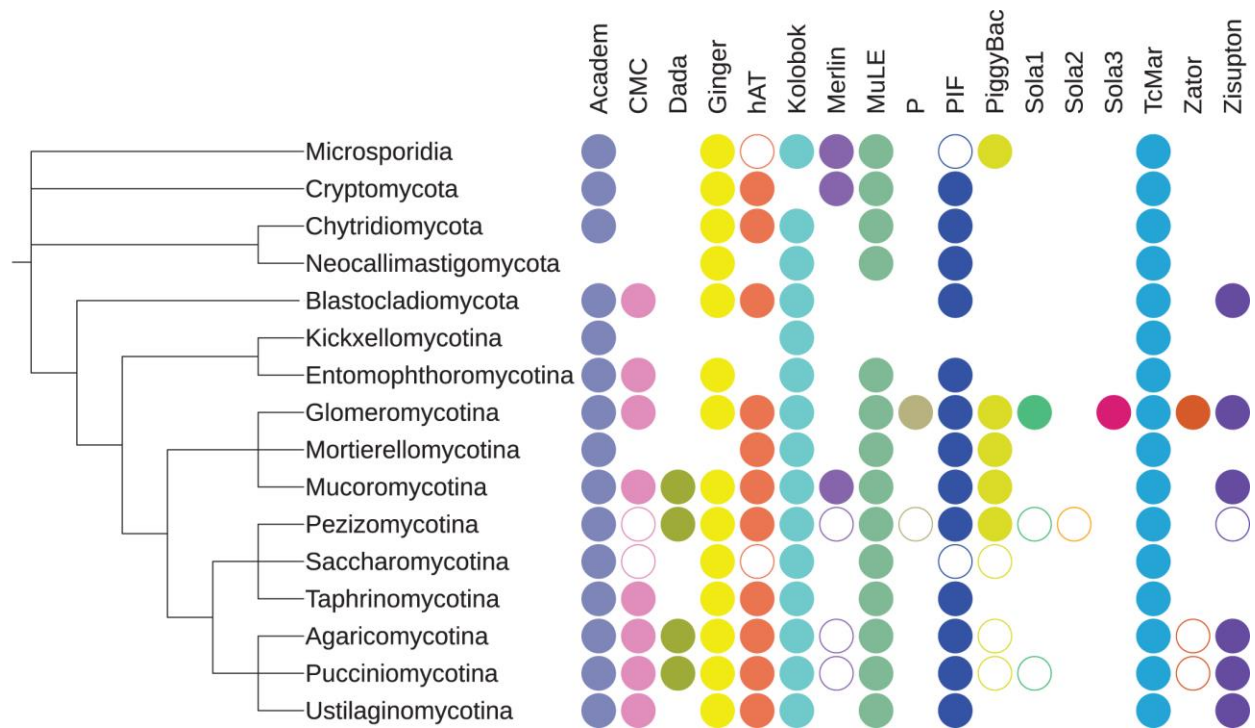
Silva JC, Loreto EL, Clark JB 2004. Factors that affect the horizontal transfer of transposable elements. *Current issues in molecular biology* 6: 57-71.

- Singer MJ, Selker EU. 1995. Genetic and Epigenetic Inactivation of Repetitive Sequences in *Neurospora crassa*: RIP, DNA Methylation, and Quelling. In. *Current Topics in Microbiology and Immunology*. p. 165-177.
- Spatafora JW, et al. 2016. A phylum-level phylogenetic classification of zygomycete fungi based on genome-scale data. *Mycologia* 108: 1028-1046. doi: 10.3852/16-042
- Sultana T, Zamborlini A, Cristofari G, Lesage P 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. *Nat. Rev. Genet.* 18: 292-308. doi: 10.1038/nrg.2017.7
- Sun C, Feschotte C, Wu Z, Mueller RL 2015. DNA transposons have colonized the genome of the giant virus Pandoravirus salinus. *BMC Biol.* 13: 38. doi: 10.1186/s12915-015-0145-1
- van der Heijden MGA, Martin FM, Selosse M-A, Sanders IR 2015. Mycorrhizal ecology and evolution: the past, the present, and the future. *New Phytol.* 205: 1406-1423. doi: 10.1111/nph.13288
- Wallau GL, Ortiz MF, Loreto ELS 2012. Horizontal transposon transfer in eukarya: detection, bias, and perspectives. *Genome Biol. Evol.* 4: 689-699. doi: 10.1093/gbe/evs055
- Wang X, et al. 2010. Sex-induced silencing defends the genome of *Cryptococcus neoformans* via RNAi. *Genes Dev.* 24: 2566-2582. doi: 10.1101/gad.1970910
- Warburton PE, Giordano J, Cheung F, Gelfand Y, Benson G 2004. Inverted repeat structure of the human genome: the X-chromosome contains a preponderance of large, highly homologous inverted repeats that contain testes genes. *Genome Res.* 14: 1861-1869. doi: 10.1101/gr.2542904
- Yamada KD, Tomii K, Katoh K 2016. Application of the MAFFT sequence alignment program to large data—reexamination of the usefulness of chained guide trees. *Bioinformatics* 32: 3246-3251. doi: 10.1093/bioinformatics/btw412
- Yerlici VT, Landweber LF 2014. Programmed Genome Rearrangements in the Ciliate *Oxytricha*. *Microbiol Spectr* 2. doi: 10.1128/microbiolspec.MDNA3-0025-2014
- Yuan Y-W, Wessler SR 2011. The catalytic domain of all eukaryotic cut-and-paste transposase superfamilies. *Proc. Natl. Acad. Sci. U. S. A.* 108: 7884-7889. doi: 10.1073/pnas.1104208108

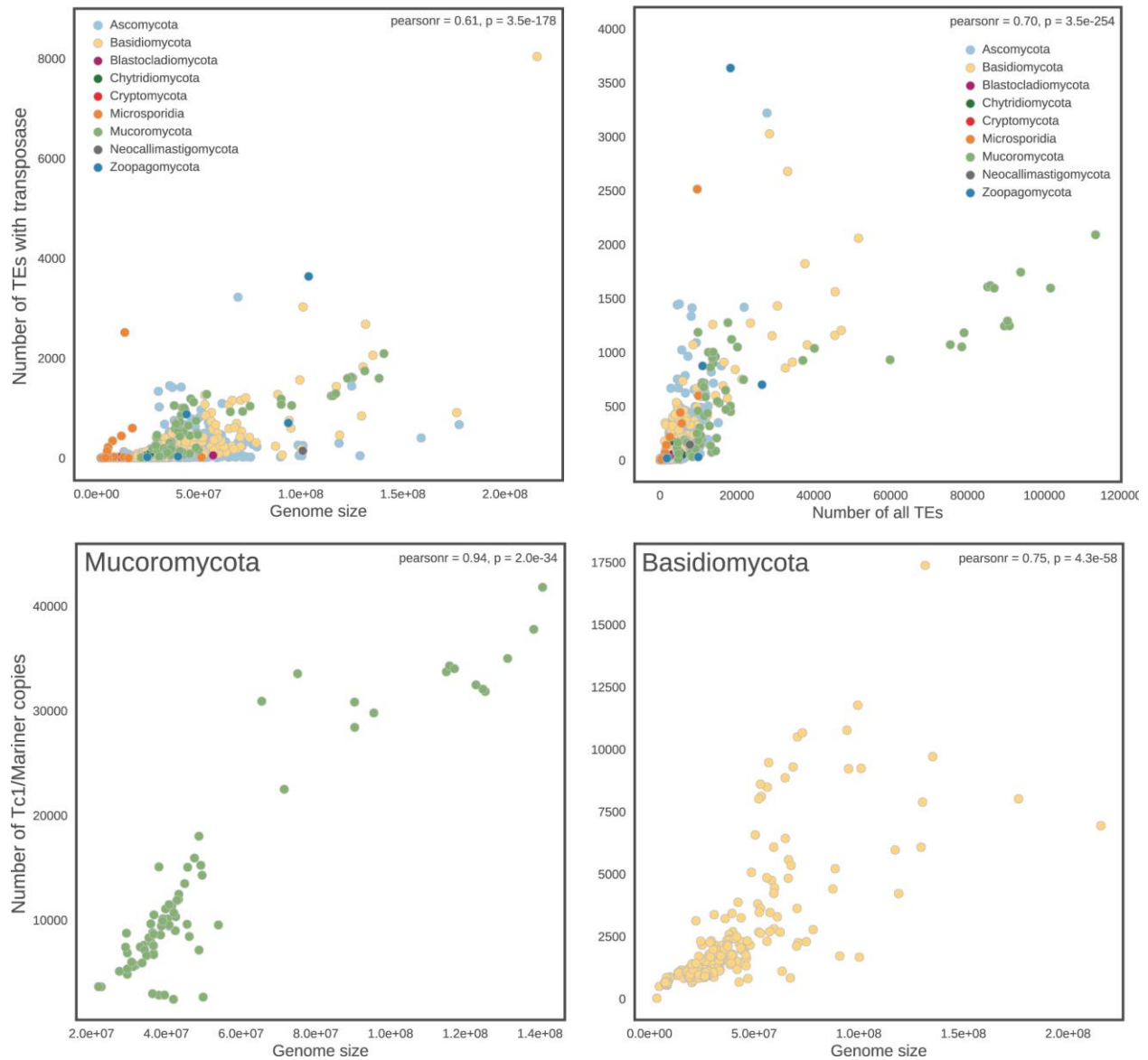
**FIGURES**



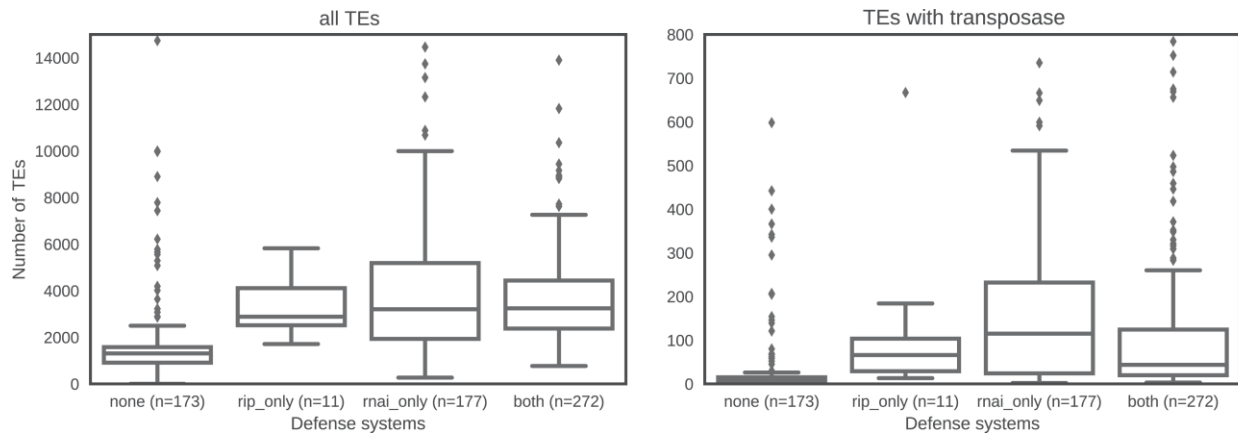
**Fig. 1.** Semi-automatic pipeline for DNA transposon detection and classification. The assignment of protein domains to TE superfamilies was performed manually.



**Fig. 2.** Taxonomic distribution of DNA TE superfamilies in major fungal lineages. Empty and filled circles depict occurrences in less than 10% and more than 10% of given taxon's representatives, respectively. Cryptomycota, Blastocladiomycota, Kickxellomycotina and Neocallimastigomycota are represented only by one isolate.

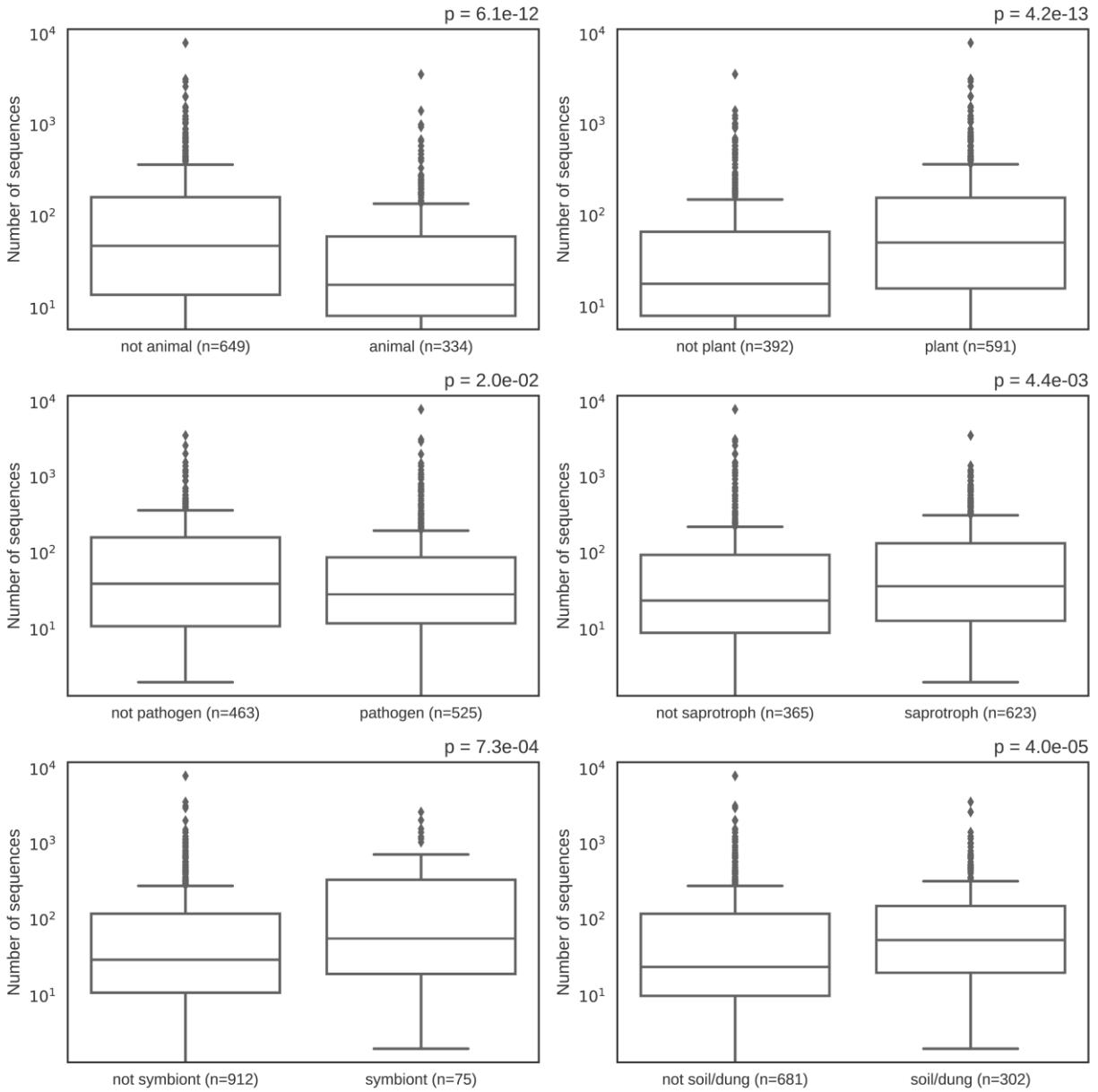


**Fig. 3.** A) Distribution of TE containing transposase versus genome size (n=1726). B) Correlation between the abundance of all and transposase-containing DNA TEs for each genome. *Uromyces viciae-fabae* outlier was not shown (x=91391, y=8039). Tc1/Mariner abundance in C) Mucoromycota (n=74) and D) Basidiomycota (n=314). The figure was prepared using Python in Jupyter (Kluyver, et al. 2016).



**Fig. 4.** Abundance of DNA TEs in fungi depending on RNAi and/or RIP presence for A) all TEs and B) only TEs retaining transposase. All differences between RNAi and both systems possessing fungi is statistically significant for transposons with a transposase (p-val. 2.2e-05).





**Fig. 5.** Distribution of transposase-containing DNA TEs in fungi with a given lifestyle. Significance of differences is confirmed by Mann-Whitney U test. A log scale is used for DNA TE count.

## TABLE LEGENDS

**Table 1.** Summary of DNA TE superfamilies in RepBase, Yuan and Wessler (Yuan and Wessler 2011) and transposons identified in this study. Assignment of protein families to DNA transposon superfamilies resulting from RepBase reference mapping on Pfam database of protein domains is given where applicable, HMM profiles are available as [Supplementary File S1](#) for remaining families. Other domains, e.g. DNA binding, associated with a particular superfamily, are marked with an asterisk.

**Table 1. Summary of DNA TE superfamilies in RepBase, Yuan and Wessler (Yuan and Wessler 2011) and transposons identified in this study.**

Superfamily	RepBase (Fungi)	Count in RepBase	Yuan & Wessler	DNA TE (with domain)	Observed distribution	Domains
Academ	only <i>Puccinia graminis</i>	7	n	9709	low copy, in most taxa, highest abundance in Pucciniomycotina	RNase H-like (Yuan and Wessler 2011)
CMC	Basidiomycota & Mucorales	28	y	11961	broader distribution, expanded in Agaricomycetes Pucciniomycotina and Mucoromycotina	Transposase_21:PF02992, Transposase_23:PF03017*, Transposase_24:PF03004*
Dada	only <i>Laccaria bicolor</i>	2	-	1023	broader distribution (4 phyla)	RNase H-like (Kojima and Jurka 2013)
Ginger	only <i>Malassezia globosa</i>	1	n	6648	ubiquitous, expansions in Dikarya	rve:PF00665
hAT	only Dikarya	37	y	33376	ubiquitous	Dimer_Tnp_hAT:PF05699*, DUF659:PF04937, DUF4371:PF14291, DUF4413:PF14372.1*
KDZ (Zisupton)	only <i>Puccinia graminis</i>	4	-	14607	Basidiomycota, Rhizophagus, Mucoromycotina and Allomyces	RNase H-like (Iyer, et al. 2014)
Kolobok	only <i>Rhizophagus irregularis</i>	5	n	3214	low copy, ubiquitous, highest abundance in <i>Rhizophagus irregularis</i>	RNase H-like (Yuan and Wessler 2011)
Merlin	only <i>Rhizopus oryzae</i>	5	y	4255	single occurrences in Dikarya, expansions in Microsporidia	DDE_Tnp_IS1595:PF12762
MULE	Dikarya & Rhizopus	36	y	17658	ubiquitous	Transposase_mut:PF00872, MULE:PF10551
Novosib	n	0	n	0	only copies without transposase	-
P	only Pucciniales & Allomyces	17	y	11	single occurrences	Tnp_P_element_C:PF12596*, Tnp_P_element:PF12017
PIF/Harb	diverse Fungi	76	y	13443	ubiquitous	Plant_tran:PF04827, DDE_Tnp_4:PF13359
PiggyBac	Mucor & Pezizomycotina	4	y	5965	Mucoromycota, Microsporidia and Pezizomycotina	DDE_Tnp_1_7:PF13843
Sola1	n	0	n	140	Rhizophagus, single occurrences in Dikarya	RNase H-like (Majorek, et al. 2014)
Sola2	n	0	n	1	one case in <i>Aspergillus flavus</i>	RNase H-like (Majorek, et al. 2014)
Sola3	n	0	n	637	only Rhizophagus	RNase H-like (Majorek, et al. 2014)
Tc1/Mariner	diverse Fungi	148	y	93120	ubiquitous	DDE_1:PF03184, DDE_3:PF13358, Transposase_1:PF01359
Transib	n	0	n	0	absent	RAG1:PF12940
Zator	only <i>Puccinia striiformis</i>	2	n	1165	Rhizophagus, single occurrences in Basidiomycota	RNase H-like (Majorek, et al. 2014)

