Fungal Genetics and Biology 172 (2024) 103897

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

Regular Articles

Anaerobic fungi contain abundant, diverse, and transcriptionally active Long Terminal Repeat retrotransposons

Tejas A. Navaratna [a,b,1], Nabil Alansari [a,1], Amy R. Eisenberg [a,b,1], Michelle A. O'Malley [a,b,c,*]

[a] Department of Chemical Engineering, UC Santa Barbara, United States
[b] California NanoSystems Institute, United States
[c] Department of Bioengineering, UC Santa Barbara, United States

A B S T R A C T

Long Terminal Repeat (LTR) retrotransposons are a class of repetitive elements that are widespread in the genomes of plants and many fungi. LTR retrotransposons have been associated with rapidly evolving gene clusters in plants and virulence factor transfer in fungal-plant parasite-host interactions. We report here the abundance and transcriptional activity of LTR retrotransposons across several species of the early-branching *Neocallimastigomycota*, otherwise known as the anaerobic gut fungi (AGF). The ubiquity of LTR retrotransposons in these genomes suggests key evolutionary roles in these rumen-dwelling biomass degraders, whose genomes also contain many enzymes that are horizontally transferred from other rumen-dwelling prokaryotes. Up to 10% of anaerobic fungal genomes consist of LTR retrotransposons, and the mapping of sequences from LTR retrotransposons to transcriptomes shows that the majority of clusters are transcribed, with some exhibiting expression greater than 10^4 reads per kilobase million mapped reads (rpkm). Many LTR retrotransposons are strongly differentially expressed upon heat stress during fungal cultivation, with several exhibiting a nearly three-$\log_{10}$ fold increase in expression, whereas growth substrate variation modulated transcription to a lesser extent. We show that some LTR retrotransposons contain carbohydrate-active enzymes (CAZymes), and the expansion of CAZymes within genomes and among anaerobic fungal species may be linked to retrotransposon activity. We further discuss how these widespread sequences may be a source of promoters and other parts towards the bioengineering of anaerobic fungi.

1. Introduction

Large proportions of many genomes across all kingdoms of life are repetitive in nature (Britten and Kohne, 1968; SanMiguel et al., 1998). These repetitive elements have been classified based on length, structure, origin (e.g. viral), and sometimes function (Wessler, 2006). Transposable elements (TEs) are autonomous and mobile in nature and have the ability to be cut or copied, and then pasted into a different genomic location. TE activity can create population-level diversity, and is thought to contribute to evolution through creation of new variants and selection (Finnegan, 1989; Gozashti et al., 2022). This has been studied extensively in the domestication and selective breeding of crops such as rice (Gao et al., 2004) and wheat (Wicker et al., 2018), plants with a long history of genetics research (Kumar and Bennetzen, 2003). TEs moreover have been used as inspiration to develop genetic tools, for example the retrotransposon *Ty* in *S. cerevisiae* for genomic integration of plasmid DNA (Boeke et al., 1988), and *Tc1/mariner* for genome-scale functional gene screening in various organisms (van Opijnen et al., 2009).

LTR retrotransposons are TEs that consist of two highly similar repeats (LTRs) flanking a relatively long retroviral-like region typically containing two open reading frames that encode the protein machinery for replication (Fig. 1). LTR retrotransposons have been previously characterized in many ascomycete and basidiomycete fungi (Muszewska et al., 2019) and it was found that these elements are widespread and abundant in these later evolutionarily branching fungi, with an average of 1129 LTR retrotransposons per genome. Two LTR retrotransposon superfamilies, *copia* and *gypsy*, characterized by their distinct open reading frame (ORF) organization (Eickbush and Malik, 2002), were identified in these genomes (Fig. 1), with the number of LTR retrotransposons per genome varying significantly, even in closely related species. A particularly interesting finding in this report was that fungi associated with plants (e.g. pathogens such as *Magnaporthe grisea* and non-pathogenic *Phanerochaete* fungi) have a greater number of LTR retrotransposon expansions compared with non-plant-associating fungi. LTR retrotransposons have also been characterized in the amphibian-infecting chytrids *B. dendrobatidis* (Muszewska et al., 2011) *and B. salamandivorans* (Wacker et al., 2023), which are sister species to the *Neocallimastigomycota*.

In this report, we characterize the LTR retrotransposon landscape

* Corresponding author.
E-mail address: momalley@ucsb.edu (M.A. O'Malley).
[1] These authors contributed equally.

https://doi.org/10.1016/j.fgb.2024.103897
Received 5 February 2024; Received in revised form 14 April 2024; Accepted 10 May 2024
Available online 13 May 2024

within the genomes and transcriptomes of anaerobic, rumen-inhabiting *Neocallimastigomycota*. The *Neocallimastigomycota,* or anaerobic gut fungi (AGF), are an early-diverging fungal branch and are thought to have arisen during the emergence of grasses and grass-consuming mammals during the early Mesozoic–Cenozoic radiation (Wang et al., 2019). Their ability to secrete powerful carbohydrate active enzymes (CAZymes) (Dementiev et al., 2023; Lowe et al., 1987; Mountfort and Asher, 1989; Solomon et al., 2016; Teunissen et al., 1991) and break down complex lignocellulosic biomass into sugars and fatty acids contributes to host nutrition (Hartinger and Zebeli, 2021). Along with gut prokaryotes, AGF are thought be important for the evolutionary success of ruminant herbivores (Wang et al., 2019). While they are difficult to genetically transform (Hooker et al., 2023), recent genomic sequencing efforts (Brown et al., 2021; Grigoriev et al., 2012; Youssef et al., 2013; Haitjema et al., 2017) indicate that many of their biomass-degrading genes were horizontally acquired from rumen gut bacteria through an unknown mechanism (Haitjema et al., 2017; Murphy et al., 2019). Here, we report that the genomes of AGF consist of up to 10 % LTR retrotransposons (Table 1), with significant variation between species, and an even larger proportion of these genomes consisting of LTR-bounded sequences not containing retrotransposon element homology, henceforth termed 'unclassified LTRs'. These unclassified LTRs generally lack longer open-reading frames and are generally not found at high genome copy numbers, leading us to hypothesize that these are erstwhile LTR retrotransposon that fragmented or otherwise lost retrotransposon structure through diverse mechanisms, with resulting loss of their ability to proliferate throughout the genome. Interestingly, many unclassified LTR retrotransposons continue to be transcribed at high levels. We furthermore characterize the transcriptional response of these LTR retrotransposons to heat shock, as well as during cultivation with different growth substrates, ranging in complexity from the monosaccharide glucose to lignocellulose-containing grasses. Overall, the many LTR retrotransposons found in AGF are responsive to stress conditions and are a potential cause of evolutionary diversification, and furthermore represent a source of tools for the engineering of these organisms.

2. Results

2.1. LTR retrotransposons are widespread in the genomes of anaerobic gut fungi

The abundance of high quality anaerobic fungal genomes through the efforts of the Joint Genome Institute's Mycocosm program (Grigoriev et al., 2012) has advanced the understanding of their functional biology, including CAZyme production (Hagen et al., 2021; Solomon et al., 2016) and natural product clusters (Swift et al., 2021a). LTR retrotransposons play important roles in specialization and evolution of distinct phenotypes in a broad range of organisms, and we reasoned that AGF, which are highly specialized to their herbivore digestive system niche, may also contain diverse and abundant LTR retrotransposons that may have contributed to their evolutionary trajectory and continue to play important roles in their biology.

LTRharvest (Ellinghaus et al., 2008) was used to identify LTR retrotransposons in the genomes of seven isolates of AGF and found that all contained a remarkably high proportion of LTR retrotransposons. The *Neocallimastix* genus, in particular, had a much greater number of LTR retrotransposons (Table 1, Fig. 2A), as well as a higher genomic proportion of LTR retrotransposons and LTR-bounded sequences (Table 1, Figure S1). It is important to use genome assemblies of high quality, as accurate detection of repetitive elements requires sufficient long-read sequencing accuracy and coverage (Ou et al., 2018). The genome sequence of *Pecoramyces* sp. *C1A* (formerly called *Orpinomyces* sp., Youssef et al., 2013) was excluded from further analysis because its large number of scaffolds (32574) likely resulted in misidentification and underestimation of the number of identified LTR retrotransposons (Figure S1).

2.2. Clustering and classification suggest widespread loss-of-function of LTR retrotransposons in anaerobic gut fungi

CD-HIT (Fu et al., 2012) was used on LTRharvest-identified sequences across the anaerobic fungal genomes for alignment-based clustering using a minimum coverage of 70 % and sequence identity of 90 %. We found that between 40–56 % of LTR retrotransposons in anaerobic fungi clustered (Table 1), but the vast majority of LTR-bounded unclassified sequences did not cluster (Table S1) with clustering histograms for *A. robustus, N. californiae,* and *P. finnis* shown in Figure S3. HMMsearch as implemented in TESorter (Zhang et al., 2022) was further used to classify LTR retrotransposons in anaerobic fungi against the GyDB transposon database (Llorens et al., 2011). Classifiable LTR retrotransposons generally contained well-defined and long ORFs with TEsorter-assigned function, i.e. reverse transcriptase, RNAse H, or integrase (Fig. 3). Unclassifiable LTR-bounded sequences did not contain any HMMER-assigned GyDB database retrotransposon domains at a lenient cutoff of 20 % coverage and e-value of < 0.001 (Zhang et al., 2022) and manual sequence inspection revealed many ORFs (Fig. 3), which is suggestive of retrotransposon inactivation through translocations, inversions, or insertions. We furthermore enumerated open reading frames (ORFs) within LTR-bounded sequences and compared classified LTR retrotransposons and LTR-bounded unclassified sequences in *N. californiae* (Figure S4). The distribution shapes are remarkably different, regardless of the ORF definition used (stop codon to stop codon, or start codon to stop codon), with classified LTR retrotransposons exhibiting a normal-like distribution, but unclassified sequences exhibiting an exponential distribution with a longer tail, indicating loss-of-function mutations have taken place over

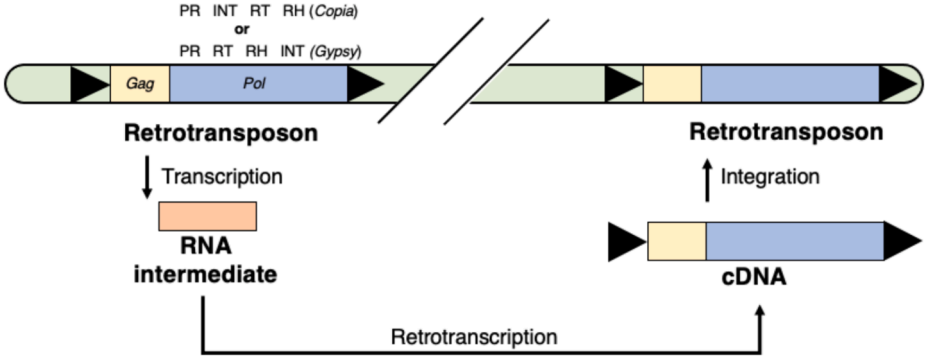

Fig. 1. Full length Long Terminal Repeat (LTR) Retrotransposons contain two near-identical repeats flanking a central region containing viral Gag and Pol sequences that are responsible for the copy-paste mechanism of retrotransposition. The organization of Pol varies by LTR retrotransposon type; Copia and Gypsy-type LTR retrotransposons are shown here (PR, protease; INT, integrase; RT, reverse transcriptase; RH, RNase H).

Table 1
Species or isolates of AGF that have sequenced genomes and transcriptomes, number of identified LTR retrotransposons, and classification information.

Isolate	Genome reference	#	% genome	% Gypsy	% Copia	% other	% transcribed	% clustered
Anaeromyces robustus	(Haitjema et al., 2017)	181	1.4	68	29	3.3	72	40
Caecomyces churrovis	(Henske et al., 2017)	558	2.2	68	29	3.0	72	53
Neocallimastix californiae	(Haitjema et al., 2017)	2618	8.8	58	38	3.3	64	50
Neocallimastix frontalis var. *giraffae*		2737	7.7	59	39	2.2	81	41
Neocallimastix lanati	(Wilken et al., 2021)	3050	10.2	64	33	3.4	97	56
Piromyces finnis	(Haitjema et al., 2017)	589	7.1	80	14	5.9	89	54
Piromyces sp. *UH3-1*		382	3.1	54	41	5.2	96	40

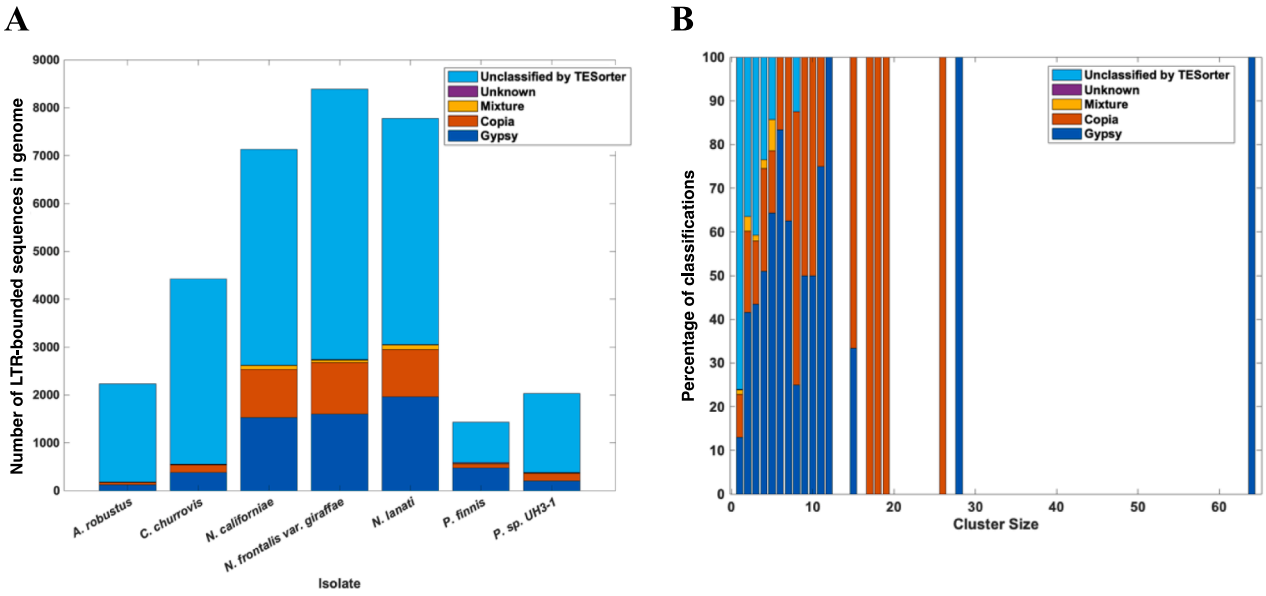

Fig. 2. A: LTR-bounded sequences are highly abundant in anaerobic fungal genomes, with classifiable LTR retrotransposons being Gypsy-type dominant (dark blue). The majority of LTR-containing sequences in all genomes are not classifiable however (light blue), indicating inactivation and loss of Gag and Pol sequences. B: A mapping of cluster size vs. classification type in the isolate N. californiae. The majority of LTR-containing sequences are orphans (cluster size 1), and these are dominated by non-classifiable sequences. Sequences belonging to larger clusters are predominantly classifiable and contain homology to retrotransposon elements. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

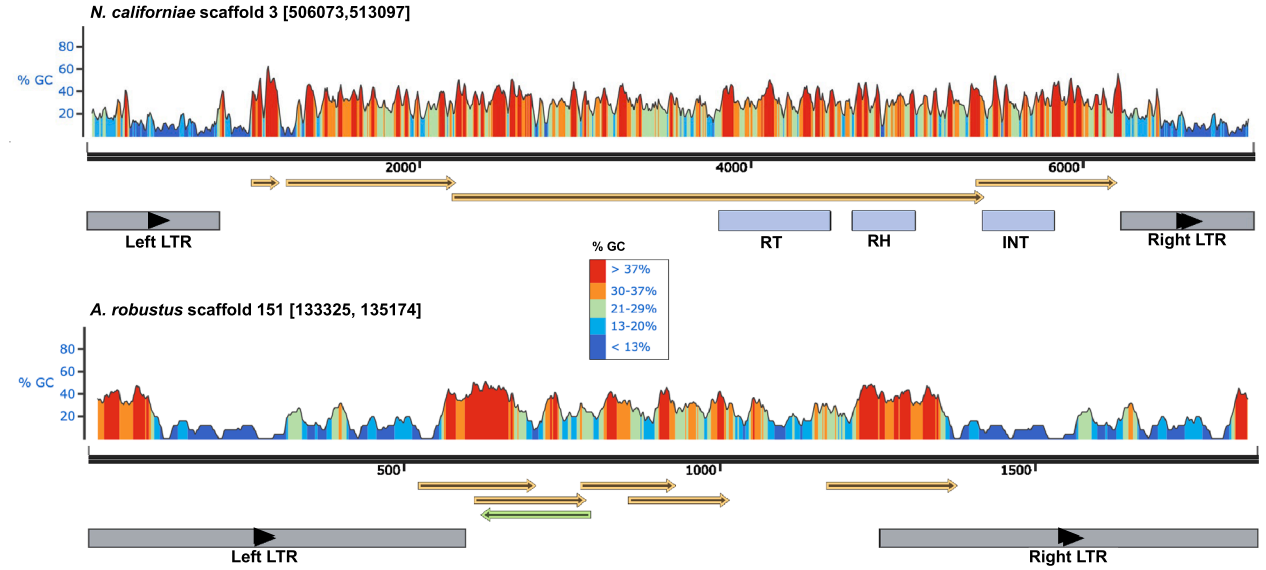

Fig. 3. Examples of a highly clustered, classifiable LTR retrotransposon (cluster size 64, Gypsy) from N. californiae and a moderately clustered, unclassifiable LTR-bounded sequence (cluster size 14) from A. robustus. Note differences in lengths. Open reading frames are indicated with arrows and TEsorter-annotated regions are shown in boxes. Functional annotations correspond to those in Fig. 1A – RT, reverse transcriptase; RH, RNAse H; INT, integrase. Schematic generated with Snapgene Viewer.

evolutionary time through recombination and translocation resulting in extensive fragmentation in unclassified LTR-bounded sequences.

Classified LTR retrotransposons generally demonstrated a single-peak length distribution (Fig. 4A) and a narrower distribution than LTR-containing sequences that were not classified (Fig. 4B), which is consistent with random recombination-based inactivation of LTR retrotransposons throughout the genome, which may in turn have arisen from the same original retrotransposon transposition events. Furthermore, larger clusters are nearly all classifiable (Fig. 2B). All isolates had a higher proportion of LTR retrotransposons classified as Gypsy relative to Copia (Table 1, Figure S2), though the majority of LTR-containing sequences in each isolate were unclassifiable by TESorter using the available databases. Interestingly, the proportion of unclassified LTR retrotransposons varied strongly among the isolates studied here (Table 1, Fig. 2A), reflecting different timings of LTR retrotransposon acquisition, activity, and inactivation.

In addition to performing clustering within a species, we also looked to see if LTR retrotransposons are shared between species. Performing pairwise BLASTn alignments with each genome as the query and database showed generally low levels of conservation of LTRs (Table S2). For example, 14.7 % of LTR retrotransposons in *C. churrovis* are shared in *N. californiae*, but only 2.7 % of LTR retrotransposons in *N. californiae* are shared in *C. churrovis*, which is consistent with the larger number and diversity of LTR retrotransposons in *N. californiae*. The highest similarity was found between *N. californiae* and *N. lanati*, which are known to be very closely related and in the same species complex. Surprisingly, some pairs (*Piromyces* sp. *UH3-1* and *Piromyces finnis*) had no shared LTR retrotransposons.

2.3. A subset of LTR retrotransposons contain carbohydrate-active enzymes

Carbohydrate-active enzymes (CAZymes) are fundamental to the role of anaerobic fungi in host metabolism and are widespread throughout anaerobic fungal genomes (Solomon et al., 2016; Youssef et al., 2013). They are thought to have been introduced through horizontal gene transfer events from other gut dwelling microbes such as methanogenic archaea and anaerobic bacteria (Haitjema et al., 2017; Murphy et al., 2019). Long terminal retrotransposons may have played a role in the propagation of virulence factors in *Batrachochytrium* species (Wacker et al., 2023), and we were intrigued by the possibility that CAZymes may have similarly been propagated by LTR retrotransposons.

There were LTR retrotransposons that contained CAZyme domains in most isolates analyzed. *N. californiae* contained the greatest diversity of within-LTR retrotransposon CAZymes (Figure S5), with 5 glycoside hydrolase (GH) domains, 4 glycosyltransferase (GT) domains, and 6 polysaccharide lyase (PL) domains, as well as 13 dockerin domains typically associated with CAZyme machinery (Haitjema et al., 2017). Several LTR retrotransposons that contain CAZymes are listed in detail in Table S4.

2.4. LTR retrotransposons are transcriptionally active in anaerobic fungi

We analyzed the transcriptomes of *N. californiae, A. robustus*, and *P. finnis* (Solomon et al., 2016) through BLAST searches against LTRharvest-identified LTR-bounded sequences. The percentage of LTR retrotransposons with matching transcripts varied from 64 % in *N. californiae* to 97 % in *N. lanati* (Table 1), although the percentage of all LTR-bounded sequences with matching transcripts was somewhat lower (Table S1). Somewhat surprisingly, we found no discernable

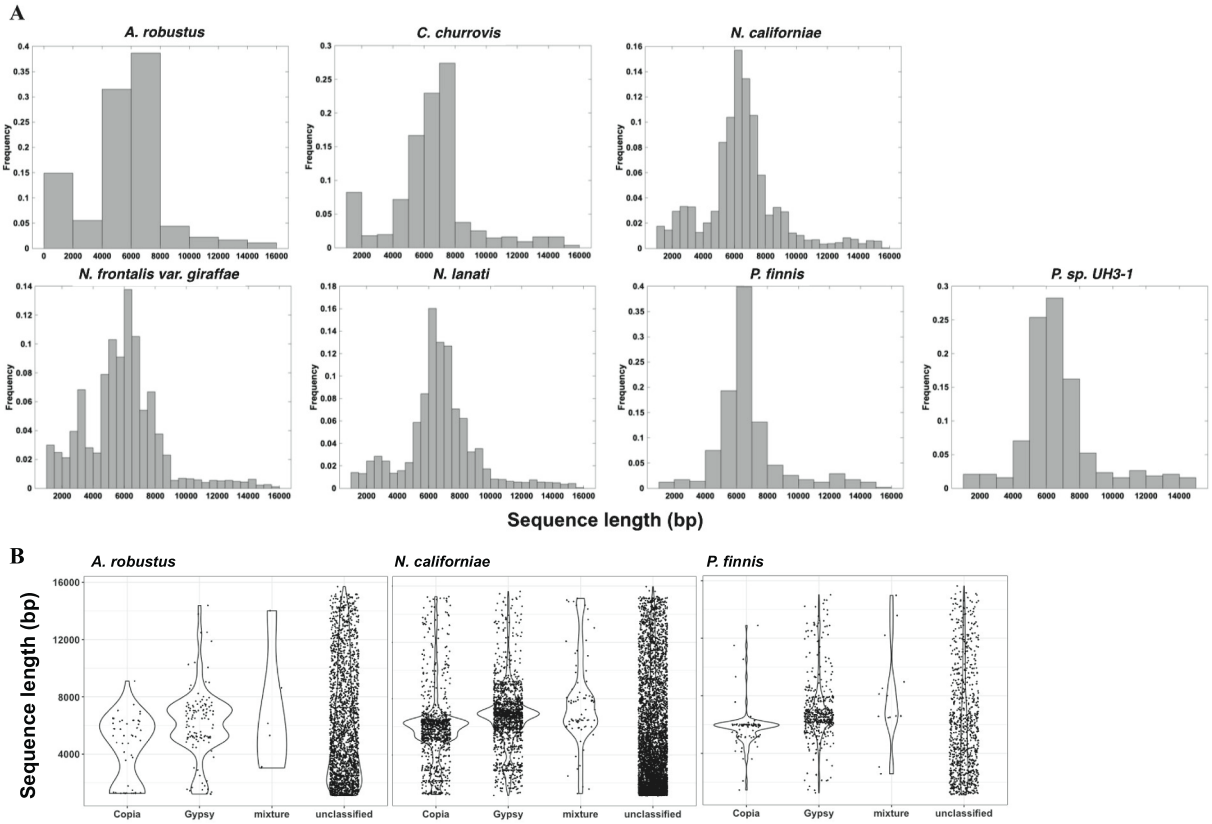

Fig. 4. A: Length distributions of all classifiable LTR retrotransposon sequences across 7 anaerobic fungal genomes. B: Length distributions of LTR-bounded sequences separated by classification for 3 anaerobic fungal genomes. Unclassifiable retrotransposon domain-lacking sequences have the broadest distribution and are skewed towards shorter lengths than Copia or Gypsy retrotransposon-classified sequences.

relationship between LTR retrotransposon length and transcriptional activity (Fig. 5A), and unclassified LTR retrotransposons had a broader range of transcriptional activity than Gypsy or Copia-type LTR retrotransposons (Fig. 5B). This suggests than LTR retrotransposons, regardless of intactness, continue to be transcribed at varying levels. This information could be used to source promoter sequences and transcription start sites to better understand transcriptional regulation in these isolates, as well as to develop genetic tools for these organisms, similar to efforts in the human gene therapy field, where retroviral vectors have been optimized, combining promoters from LTR retrotransposons and other sources with various payloads to balance dosing and adverse effects (Hoffmann et al., 2017; Weber and Cannon, 2007).

2.5. Gene promoters are associated with LTR retrotransposons

Long terminal repeats contain promoter elements that can impact expression of adjacent genes, even when a transposon is no longer autonomous and loses functional machinery on evolutionary timescales (Havecker et al., 2004). To examine the effects of LTR retrotransposition on non-LTR gene expression, putative promoter regions in *N. californiae* upstream of annotated genes were analyzed for the presence or absence of LTR-bounded but unclassified sequences, which contain no homology to *Gypsy* or *Copia* domains.

Interestingly, unclassified LTR sequences that contained promoters had a broader distribution of matched-transcript abundance values (rpkm), which is consistent with the role of these sequences as promoter-acting. Unclassified sequences that do not contain promoters had a much

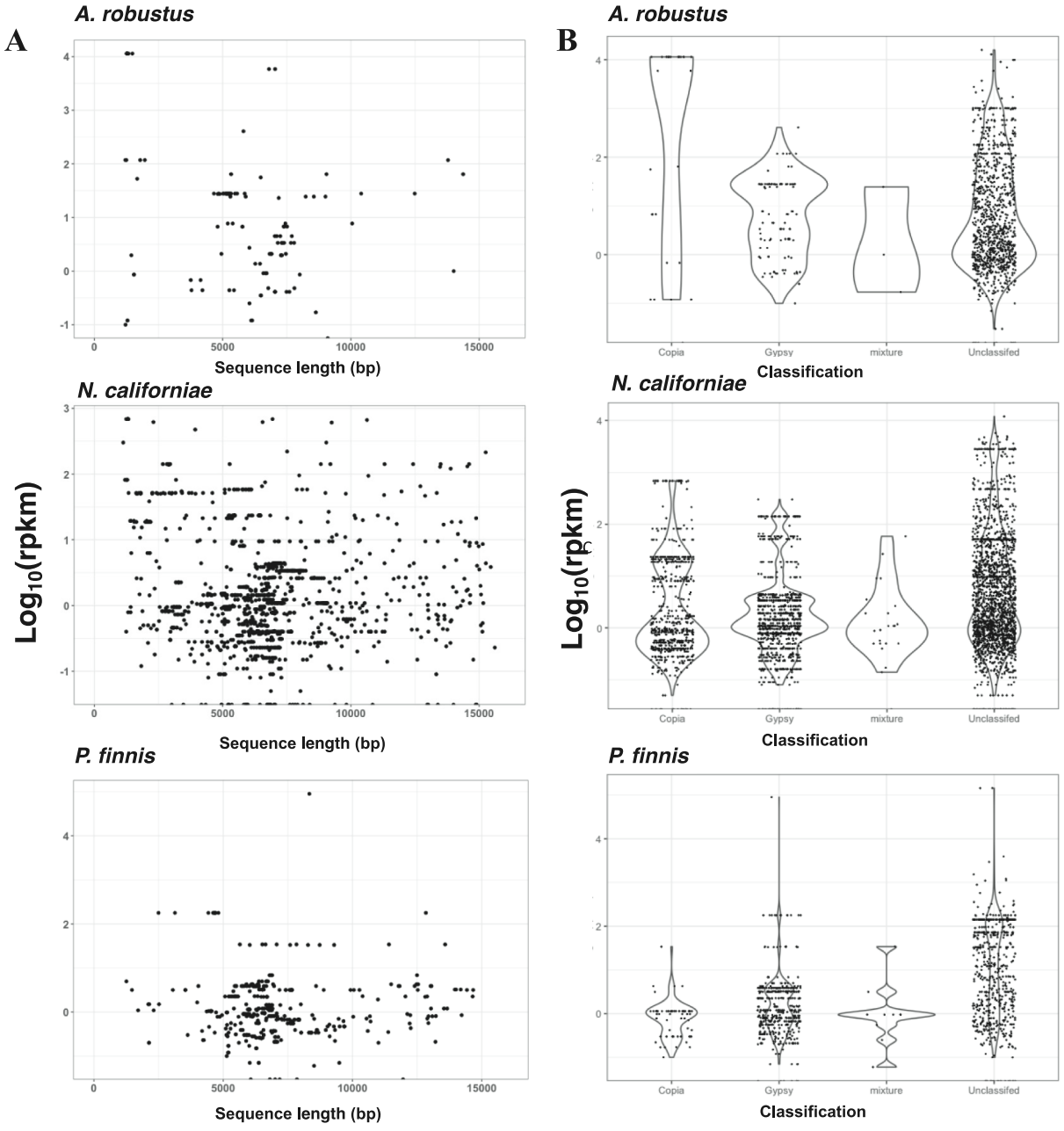

Fig. 5. A: Scatter plots of LTR retrotransposon length vs. transcriptional level in $\log_{10}$reads per kilobase million mapped reads (rpkm) across three isolates of AGF with reported transcriptomic data. There is no apparent correlation between LTR retrotransposon length and transcriptional level, indicating that for the majority of sequences, any inactivation of retrotransposition is non-transcriptional in nature. B: Distributions of transcriptional level for classification types. Patterns of transcription vary across these strains and by type.

narrower distribution (Figure S6). Selected promoter-containing sequences of interest are reported in Table S6.

2.6. Stress (heat shock) increases transcriptional activity for a subset of LTR retrotransposons

Because of the abundance of evidence for stress as a positive regulator of LTR retrotransposon activity in plants (Grandbastien, 2015), specific heat-induced activation of the *Copia* type retrotransposon ONSEN in *Arabidopsis thaliana* (Cavrak et al., 2014), as well as the recent discovery of heat-induced transposon mobility in the human fungal pathogen *Cryptococcus deneoformans* (Gusa et al., 2023), we sought to characterize the transcriptional response of classifiable LTR retrotransposons in *Neocallimastigomycota* to heat shock. We previously reported global changes in gene regulation in *N. californiae* from incubation at elevated temperature (48 °C) with RNA measurements immediately following heat shock (t_0) and in fifteen-minute intervals post-shock (t_{15}, t_{30}, t_{45}, and t_{60}) (Swift et al., 2021b). Here, we report that many LTR retrotransposon clusters show statistically significant upregulation at all post-shock times studied (Fig. 6). Furthermore, the largest LTR retrotransposon cluster (cluster size of 64) showed significantly upregulated expression at t_{60} (Table 2).

2.7. Complex growth substrates reduce LTR retrotransposon activity

Neocallimastigomycota are an herbivore gut-dwelling fungal lineage that contain the largest number of carbohydrate-active genes per genome of any organism (Solomon et al., 2016) and can use a variety of carbon sources to grow, including paper, lignocellulosic grasses, and simple sugars (Teunissen et al., 1991). We characterized the impact on classifiable LTR retrotransposon transcription in *N. californiae* by substrate variation and report log2-fold change values relative to growth on the monosaccharide glucose (Figures S8, S10). Interestingly, we found that the complex carbohydrate sources tested, reed canary grass (RCG) and switchgrass (SG), appear to generally reduce the expression of LTR retrotransposons. However, LTRs that contain CAZymes tend to increase in expression in these complex carbohydrate sources (Table S4).

3. Discussion

3.1. LTR-bounded sequences in anaerobic gut fungi are highly polymorphic

LTR-bounded sequences are widespread in the genomes of anaerobic gut fungi, but we found that most exhibit extensive loss of intervening sequence, leading them to not be classifiable (Fig. 2A). The variation in the intervening sequence loss also resulted in reduced clustering, as clustered LTR-bounded sequences had a much higher proportion that was successfully classified than singleton, unclustered LTR-bounded sequences (Fig. 2B). We posit that loss-of-function mutations are selected for in general in anaerobic fungal LTR retrotransposons, including the observed deletions. Many unclassified LTR-bounded sequences also exhibit extensive ORF fragmentation, leading to a small population of sequences with many ORFs, but a dominant fraction with few ORFs relative to the ORF distribution for classified LTR retrotransposons (Figure S4). Interestingly, many of these unclassified LTRs are still highly transcribed, though at a lower proportion than those that are classifiable, which suggests the presence of at least partially complete promoters within or adjacent to these LTR-bounded sequences. This architecture is consistent with the hypothesis that while these unclassified LTRs with missing intervening sequence are likely not autonomously capable of transposition, they still contain promoters and machinery for transcription. We do not discount that there may also be additional mechanisms of post-transcriptional inactivation taking place, such as siRNA-mediated processes, as has been observed in *Magnaporthe oryzae* (Murata et al., 2007).

While this work examines already-existing LTR retrotransposons in sequenced anaerobic gut fungi genomes, the timescale of retrotransposon replication in these organisms is unknown. Additionally, it is unclear if the high transcriptional activity of some LTR retrotransposons identified here is correlated with functional transposition. As a variety of environmental stressors can trigger TE activity in other organisms (Grandbastien, 2015), it is plausible that laboratory "domestication" of these fungi has led to marked changes in the patterns of LTR retrotransposon activity. To address these questions, re-sequencing of isolates, Southern blot analysis, and/or genomic qPCR approaches are warranted to characterize patterns of active retrotransposition in anaerobic fungi.

3.2. LTR retrotransposons may have played a role in the evolutionary history of anaerobic gut fungi

LTR retrotransposons may also play important roles in the sister phylum *Chytridiomycota.* LTR retrotransposons in the genomes of *B. dendrobatidis* and *B. salamandivorans* have been identified in virulence regions, and duplication or recombination events mediated by LTR retrotransposons could have played a role in the evolution of these fungi becoming amphibian pathogens (Wacker et al., 2023). We report here that diverse carbohydrate-active enzymes are encoded within many LTR-bounded sequences in anaerobic fungi, both within

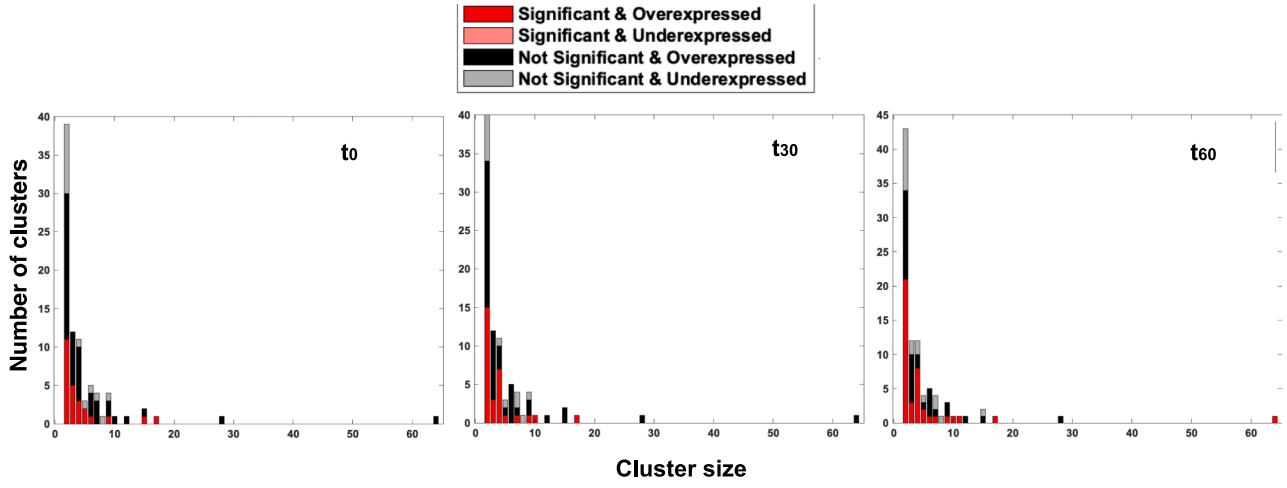

Fig. 6. In N. californiae, many large LTR retrotransposon clusters are upregulated with heat shock. The transcriptional response is shown to heat shock immediately following (t_0), 30 min after (t_{30}), and 60 min after (t_{60}). Each cluster is colored by statistical significance and relative expression (to a pre-heat shock measurement).

Table 2
Top 5 differentially regulated LTR retrotransposons and *large cluster (size 64) representative LTR retrotransposon, with genome location, matching transcript, LTR retrotransposon length, type classification, and expression fold-change 60 min post-heat shock relative to immediately prior to heat shock.

Genome location	Best transcript match	Length	Type	Log_2(fold change)
Scaffold 297; [100438,113767]	Locus25980v1rpkm0.09	13,330	*Gypsy*	9.85
Scaffold 392; [34234,40859]	Locus15888v1rpkm0.97	6626	*Gypsy*	9.16
Scaffold 109; [483971,498622]	Locus13650v1rpkm1.44	14,652	*Gypsy*	8.70
Scaffold 36; [421129,427420]	Locus14645v1rpkm1.20	6292	*Copia*	7.74
Scaffold 211; [71460,85437]	Locus12214v1rpkm1.91	13,978	*Copia*	7.73
Scaffold3; [506073,513097]*	Locus17229v1rpkm0.78	7025	*Gypsy*	2.13

retrotransposons and sequences lacking gag-pol homology (Figure S5). Intriguingly, some of these are further differentially regulated upon heat stress and show predictable, substrate-dependent expression. Accordingly, we propose that the expansion of gene families responsible for the profound biodegradation activities of AGF may be partially a result of LTR retrotransposon activity. Furthermore, we observed that there is little sequence conservation among LTR retrotransposons from different species, implying divergent and relatively recent patterns of LTR retrotransposon acquisition and retrotransposition activity.

3.3. LTR retrotransposons offer a pathway towards genetic transformation of anaerobic gut fungi

The mining of genomes for parts is a critical part of tool development for eukaryotes, which is challenging, owing to the complexity of gene regulation, such as distal promoter sequences. Although it is possible that some transcriptionally active LTR retrotransposons identified here are only transcribed due to insertion downstream of native promoter sequences, intact LTR retrotransposons contain everything necessary for retrotransposition, including promoters to generate RNA intermediates, and a reverse transcriptase and integrase (i.e. they are autonomous). Indeed, MAGGY, an LTR retrotransposon first characterized in the fungus *Magnaporthe grisea* was found to be able to autonomously transpose in heterologous organisms (Nakayashiki et al., 1999), suggesting potential use as a genetic tool. Similarly, *Ty* from *Saccharomyces cerevisiae* has been long used as a genetic tool for genomic integration of foreign DNA. Non-autonomous LTR-bounded sequences should not be excluded from promoter strength characterization though – we show that such sequences that contain putative promoters of known genes in *N. californiae* tend to have higher transcript-matched expression levels than those that lack promoters (Figure S6).

To better understand the structure of promoters and other required machinery for transcription in anaerobic fungi, further work is warranted to characterize interactions between endogenous anaerobic fungal transcription factors and LTR retrotransposons identified here, especially ones that are complete, through experiments like DAP-Seq (Bartlett et al., 2017). Knowledge of promoter sequence conservation, transcriptional factor regulation, and transcriptional activity can inform the development of efficient vectors towards the engineering of anaerobic gut fungi.

4. Methods

4.1. LTR retrotransposon identification

To identify Long Terminal Repeat (LTR) retrotransposons in anaerobic fungal genomes, assembled unmasked scaffolds were downloaded from Mycocosm (Grigoriev et al., 2012) and processed with LTRharvest in GenomeTools (Ellinghaus et al., 2008). The resulting output files contained LTR-bounded sequences along with their associated data, including sequence length, start position, and end position. Default options were used for this initial identification, with the minimum repeat length set at 100 bp, the minimum distance between the two repeats set to 1000 bp, and the similarity threshold between the two repeats set to 85 %.

The LTRharvest output file was then processed with CD-HIT-EST (Fu et al., 2012) for clustering to establish LTR retrotransposon copy number within each species' genome. Options for CD-HIT were set to default except as follows: alignment coverage for the short sequence was set at 0.7, alignment coverage for the long sequence was set at 0.7, the sequence identity threshold was set to 90 %, and "accurate but slow mode" was chosen.

4.2. Assignment of transcriptional activity to each LTR retrotransposon cluster

To establish whether an individual LTR-bounded sequence is transcriptionally active, we used the LTRharvest sequence file as a BLASTn database and used the organism's transcriptome (from Mycocosm) as the query file (Camacho et al., 2009). BLASTn parameter options: query (transcriptome) coverage was set to 70 %, and the percent identity threshold was set to 90 %.

Since multiple transcripts could be BLAST hits to the same LTR-bounded sequence, the transcript with the highest rpkm was assigned. For assignment of rpkm values to a CD-HIT-identified cluster of LTR-bounded sequences, the mode of the rpkm values for each member LTR retrotransposon in a cluster was used.

4.3. LTR sequence classification

To assign classifications to each LTR sequence and its respective cluster, a HMMER search against GyDB was performed using TESorter (Llorens et al., 2011; Zhang et al., 2022) with default options.

4.4. Cross-species LTR sequence comparisons

For comparisons of LTR-bounded sequences between species, we used each species' list of LTR sequences as a BLASTn database and query. BLASTn parameter options: query coverage was set to 70 %.

4.5. DESeq2 Heatshock and substrate analysis

RNAseq reads from a *N. californiae* heat shock experiment, available at National Center for Biotechnology Information (NCBI) BioProject PRJNA665745, were processed (Swift et al., 2021b) and analyzed with DESeq2 (Love et al., 2014) in Bioconductor, implemented in R v.4.3.1, for differential expression, using the "before" read counts as the reference condition. An adjusted p-value of 0.05 was used as a cutoff, and a count threshold of 10 was used to filter out genes with low read counts. RNAseq reads from growth of *N. californiae* on various substrates were processed (Solomon et al., 2016) and RNAseq reads are available at NCBI BioProject PRJNA377241. DESeq2 was used to establish which LTR retrotransposons were differentially expressed under different growth substrate conditions, with the glucose condition being used as the reference condition.

4.6. Open reading frame content analysis

ORFs within LTRs were enumerated using getorf in EMBOSS v. 6.6.0.0 (Rice et al., 2000) using −minsize 150 −find 0 or −minsize 150

−find 1 options for the stop-to-stop or start-to-stop ORF definitions, respectively.

4.7. Promoter analysis

A list of promoters in *N. californiae* was created (neo-sp1_promoter_proteinid.fa) by selecting 1000 bp upstream from the start of genes, or less than 1000 bp (with a minimum length of 50 bp) when the gene is located at a scaffold coordinate of less than1 kb,it. Bidirectionality is indicated if the intergenic length between two adjacent genes is less than 1 kb, and the two genes are on opposite strands. To avoid redundant analysis of promoters adjacent to LTR genes, blat (Kent, 2002) was used with minimum identity parameter set to 90 % to match LTR open reading frames to genes, and promoters for these genes were excluded from the analysis, creating the file promoters_not_for_LTRorfs.fa. This file was used as the query and LTR list was used as database. The fasta files used for promoter analysis are available at https://github.com/O-Malley-Lab/LTR-AGF/tree/main/promoter_analysis.

4.8. Carbohydrate-active enzyme annotation and classification

dbCAN (Zheng et al., 2023) was run on the LTRHarvest output file for each genome, using option −prok which is for nucleotide searches.

For dockerin and scaffoldin annotations, a HMMer search was carried out using the 6-frame translation file generated with getorf (EMBOSS v. 6.6.0.0) with options −minsize 100 using the PF02013.hmm model for dockerin and cohesin3.hmm model for scaffoldin (Haitjema et al., 2017) available in https://github.com/O-Malley-Lab/LTR-AGF/tree/main/CAZymes.

4.9. Visualization

Graphs were generated using MATLAB version R2022b or in RStudio v. 2023.03.0 + 386.

Data availability

All genome sequence files and transcriptome files can be found at mycocosm.org. Code used for LTR retrotransposon identification, classification and transcriptional activity can be found at https://github.com/O-Malley-Lab/LTR-AGF, with specific directories as follows:

Compiled spreadsheets of identified LTR-bounded sequences from anaerobic fungi: https://github.com/O-Malley-Lab/LTR-AGF/tree/main/All%20LTRs%20from%20LTRHarvest/Spreadsheets

N. californiae heat shock LTR transcriptional analysis: https://github.com/O-Malley-Lab/LTR-AGF/tree/main/DESeq2%20Analysis/Heatshock%20Analysis

N. californiae growth substrate LTR transcriptional analysis: https://github.com/O-Malley-Lab/LTR-AGF/tree/main/DESeq2%20Analysis/substrate

N. californiae promoter LTR analysis: https://github.com/O-Malley-Lab/LTR-AGF/tree/main/promoter_analysis

N. californiae carbohydrate-active LTR analysis: https://github.com/O-Malley-Lab/LTR-AGF/tree/main/CAZymes

CRediT authorship contribution statement

Tejas A. Navaratna: Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Nabil Alansari:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation. **Amy R. Eisenberg:** Writing – review & editing, Writing – original draft, Investigation, Formal analysis, Data curation, Conceptualization. **Michelle A. O'Malley:** Writing – review & editing, Supervision, Project administration, Funding acquisition, Formal analysis, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We acknowledge funding support from the Department of Energy, Office of Science (DESC0020420 and DE-SC0022142), the Institute for Collaborative Biotechnologies (W911NF-19-D-0001), and National Science Foundation (2128271). This research was made possible by computational facilities purchased with funds from the National Science Foundation (CNS-1725797) and administered by the Center for Scientific Computing (CSC). The CSC is supported by the California Nano-Systems Institute and the Materials Research Science and Engineering Center (MRSEC; NSF DMR 2308708) at UC Santa Barbara. The authors thank Igor Podolsky, Candice Swift, Stephen Mondo, Ping Navaratna and Asaf Salamov for helpful conversations and advice regarding the data and analyses included in this manuscript. We further thank Kevin Solomon for the *Neocallimastix frontalis var. giraffae* and *Piromyces sp. UH3-1* fungal genomes.

Appendix A. Supplementary material

Supplementary data to this article can be found online at https://doi.org/10.1016/j.fgb.2024.103897.

References

Bartlett, A., O'Malley, R.C., Huang, S.S.C., Galli, M., Nery, J.R., Gallavotti, A., Ecker, J. R., 2017. Mapping genome-wide transcription-factor binding sites using DAP-seq. Nat. Protoc. 12 https://doi.org/10.1038/nprot.2017.055.

Boeke, J.D., Xu, H., Fink, G.R., 1988. A general method for the chromosomal amplification of genes in yeast. Science 80, 239. https://doi.org/10.1126/science.2827308.

Britten, R.J., Kohne, D.E., 1968. Repeated sequences in DNA. Hundreds of thousands of copies of DNA sequences have been incorporated into the genomes of higher organisms. Science 161.

Brown, J.L., Swift, C.L., Mondo, S.J., Seppala, S., Salamov, A., Singan, V., Henrissat, B., Drula, E., Henske, J.K., Lee, S., LaButti, K., He, G., Yan, M., Barry, K., Grigoriev, I.V., O'Malley, M.A., 2021. Co-cultivation of the anaerobic fungus Caecomyces churrovis with Methanobacterium bryantii enhances transcription of carbohydrate binding modules, dockerins, and pyruvate formate lyases on specific substrates. Biotechnol. Biofuels 14. https://doi.org/10.1186/s13068-021-02083-w.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., Madden, T.L., 2009. BLAST+: Architecture and applications. BMC Bioinformatics 10. https://doi.org/10.1186/1471-2105-10-421.

Cavrak, V.V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L.M., Mittelsten Scheid, O., 2014. How a retrotransposon exploits the plant's heat stress response for its activation. PLoS Genet. 10 https://doi.org/10.1371/journal.pgen.1004115.

Dementiev, A., Lillington, S.P., Jin, S., Kim, Y., Jedrzejczak, R., Michalska, K., Joachimiak, A., O'Malley, M.A., 2023. Structure and enzymatic characterization of CelD endoglucanase from the anaerobic fungus piromyces finnis. Appl. Microbiol. Biotechnol. 107 https://doi.org/10.1007/s00253-023-12684-0.

Eickbush, T.H., Malik, H.S., 2002. Origins and evolution of retrotransposons. Mobile DNA II, 1111–1144. https://doi.org/10.1128/9781555817954.ch49.

Ellinghaus, D., Kurtz, S., Willhoeft, U., 2008. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics 9. https://doi.org/10.1186/1471-2105-9-18.

Finnegan, D.J., 1989. Eukaryotic transposable elements and genome evolution. Trends Genet. 5 https://doi.org/10.1016/0168-9525(89)90039-5.

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

Gao, L., McCarthy, E.M., Ganko, E.W., McDonald, J.F., 2004. Evolutionary history of Oryza sativa LTR retrotransposons: a preliminary survey of the rice genome sequences. BMC Genomics 5, 1–18. https://doi.org/10.1186/1471-2164-5-18/FIGURES/10.

Gozashti, L., Roy, S.W., Thornlow, B., Kramer, A., Ares, M., Corbett-Detig, R., 2022. Transposable elements drive intron gain in diverse eukaryotes. Proc. Natl. Acad. Sci. USA 119. https://doi.org/10.1073/pnas.2209766119.

Grandbastien, M.A., 2015. LTR retrotransposons, handy hitchhikers of plant regulation and stress response. Biochim. Biophys. Acta - Gene Regul. Mech. 1849. https://doi.org/10.1016/j.bbagrm.2014.07.017.

Grigoriev, I.V., Nordberg, H., Shabalov, I., Aerts, A., Cantor, M., Goodstein, D., Kuo, A., Minovitsky, S., Nikitin, R., Ohm, R.A., Otillar, R., Poliakov, A., Ratnere, I., Riley, R.,

Smirnova, T., Rokhsar, D., Dubchak, I., 2012. The genome portal of the department of energy joint genome institute. Nucleic Acids Res. 40 https://doi.org/10.1093/nar/gkr947.
Gusa, A., Yadav, V., Roth, C., Williams, J.D., Shouse, E.M., Magwene, P., Heitman, J., Jinks-Robertson, S., 2023. Genome-wide analysis of heat stress-stimulated transposon mobility in the human fungal pathogen Cryptococcus deneoformans. Proc. Natl. Acad. Sci. USA 120. https://doi.org/10.1073/pnas.2209831120.
Hagen, L.H., Brooke, C.G., Shaw, C.A., Norbeck, A.D., Piao, H., Arntzen, M., Olson, H.M., Copeland, A., Isern, N., Shukla, A., Roux, S., Lombard, V., Henrissat, B., O'Malley, M. A., Grigoriev, I.V., Tringe, S.G., Mackie, R.I., Pasa-Tolic, L., Pope, P.B., Hess, M., 2021. Proteome specialization of anaerobic fungi during ruminal degradation of recalcitrant plant fiber. ISME J. 15 https://doi.org/10.1038/s41396-020-00769-x.
Haitjema, C.H., Gilmore, S.P., Henske, J.K., Solomon, K. V., De Groot, R., Kuo, A., Mondo, S.J., Salamov, A.A., LaButti, K., Zhao, Z., Chiniquy, J., Barry, K., Brewer, H. M., Purvine, S.O., Wright, A.T., Hainaut, M., Boxma, B., Van Alen, T., Hackstein, J.H. P., Henrissat, B., Baker, S.E., Grigoriev, I. V., O'Malley, M.A., 2017. A parts list for fungal cellulosomes revealed by comparative genomics. Nat. Microbiol. Doi: 10.1038/nmicrobiol.2017.87.
Hartinger, T., Zebeli, Q., 2021. The present role and new potentials of anaerobic fungi in ruminant nutrition. J. Fungi. https://doi.org/10.3390/jof7030200.
Havecker, E.R., Gao, X., Voytas, D.F., 2004. The diversity of LTR retrotransposons. Genome Biol. 5, 1–6. https://doi.org/10.1186/GB-2004-5-6-225/FIGURES/3.
Henske, J.K., Gilmore, S.P., Knop, D., Cunningham, F.J., Sexton, J.A., Smallwood, C.R., Shutthanandan, V., Evans, J.E., Theodorou, M.K., O'Malley, M.A., 2017. Transcriptomic characterization of Caecomyces churrovis: a novel, non-rhizoid-forming lignocellulolytic anaerobic fungus. Biotechnol. Biofuels 10. https://doi.org/10.1186/s13068-017-0997-4.
Hoffmann, D., Schott, J.W., Geis, F.K., Lange, L., Müller, F.J., Lenz, D., Zychlinski, D., Steinemann, D., Morgan, M., Moritz, T., Schambach, A., 2017. Detailed comparison of retroviral vectors and promoter configurations for stable and high transgene expression in human induced pluripotent stem cells. Gene Ther. 245 (24), 298–307. https://doi.org/10.1038/gt.2017.20.
Hooker, C.A., Hanafy, R., Hillman, E.T., Muñoz Briones, J., Solomon, K.V., 2023. A genetic engineering toolbox for the lignocellulolytic anaerobic gut fungus neocallimastix frontalis. ACS Synth. Biol. 12 https://doi.org/10.1021/acssynbio.2c00502.
Kent, W.J., 2002. BLAT - The BLAST-like alignment tool. Genome Res. 12 https://doi.org/10.1101/gr.229202. Article published online before March 2002.
Kumar, A., Bennetzen, J.L., 2003. Plant Retrotransposons. https://doi.org/10.1146/annurev.genet.33.1.479 33, 479-532.
Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J.M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G.P., Maumus, F., Munoz-Pomer, A., Sempere, J.M., Latorre, A., Moya, A., 2011. The gypsy database (GyDB) of mobile genetic elements: release 2.0. Nucleic Acids Res. 39 https://doi.org/10.1093/nar/gkq1061.
Love, M.I., Huber, W., Anders, S., 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol. https://doi.org/10.1186/s13059-014-0550-8.
Lowe, S.E., Theodorou, M.K., Trinci, A.P.J., 1987. Cellulases and xylanase of an anaerobic rumen fungus grown on wheat straw, wheat straw holocellulose, cellulose, and xylan. Appl. Environ. Microbiol. 53 https://doi.org/10.1128/aem.53.6.1216-1223.1987.
Mountfort, D.O., Asher, R.A., 1989. Production of xylanase by the ruminal anaerobic fungus Neocallimastix frontalis. Appl. Environ. Microbiol. 55 https://doi.org/10.1128/aem.55.4.1016-1022.1989.
Murata, T., Kadotani, N., Yamaguchi, M., Tosa, Y., Mayama, S., Nakayashiki, H., 2007. siRNA-dependent and -independent post-transcriptional cosuppression of the LTR-retrotransposon MAGGY in the phytopathogenic fungus Magnaporthe oryzae. Nucleic Acids Res. 35 https://doi.org/10.1093/nar/gkm646.
Murphy, C.L., Youssef, N.H., Hanafy, R.A., Couger, M.B., Stajich, J.E., Wang, Y., Baker, K., Dagar, S.S., Griffith, G.W., Farag, I.F., Callaghan, T.M., Elshahed, M.S., 2019. Horizontal gene transfer as an indispensable driver for evolution of Neocallimastigomycota into a distinct gutdwelling fungal lineage. Appl. Environ. Microbiol. 85, e00988–e01019. https://doi.org/10.1128/AEM.00988-19.
Muszewska, A., Hoffman-Sommer, M., Grynberg, M., 2011. LTR retrotransposons in fungi. PLoS One 6. https://doi.org/10.1371/journal.pone.0029425.
Muszewska, A., Steczkiewicz, K., Stepniewska-Dziubinska, M., Ginalski, K., 2019. Transposable elements contribute to fungal genes and impact fungal lifestyle. Sci. Reports 91 (9), 1–10. https://doi.org/10.1038/s41598-019-40965-0.
Nakayashiki, H., Kiyotomi, K., Tosa, Y., Mayama, S., 1999. Transposition of the retrotransposon MAGGY in heterologous species of filamentous fungi. Genetics 153. https://doi.org/10.1093/genetics/153.2.693.
Ou, S., Chen, J., Jiang, N., 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI). Nucleic Acids Res. 46 https://doi.org/10.1093/nar/gky730.
Rice, P., Longden, L., Bleasby, A., 2000. EMBOSS: the European molecular biology open software suite. Trends Genet. https://doi.org/10.1016/S0168-9525(00)02024-2.
SanMiguel, P., Gaut, B.S., Tikhonov, A., Nakajima, Y., Bennetzen, J.L., 1998. The paleontology of intergene retrotransposons of maize. Nat. Genet. 20 https://doi.org/10.1038/1695.
Solomon, K.V., Haitjema, C.H., Henske, J.K., Gilmore, S.P., Borges-Rivera, D., Lipzen, A., Brewer, H.M., Purvine, S.O., Wright, A.T., Theodorou, M.K., Grigoriev, I.V., Regev, A., Thompson, D.A., O'Malley, M.A., 2016. Early-branching gut fungi possess large, comprehensive array of biomass-degrading enzymes. Science 80 351, 1192–1195. https://doi.org/10.1126/science.aad1431.
Swift, C.L., Louie, K.B., Bowen, B.P., Olson, H.M., Purvine, S.O., Salamov, A., Mondo, S. J., Solomon, K.V., Wright, A.T., Northen, T.R., Grigoriev, I.V., Keller, N.P., O'Malley, M.A., 2021a. Anaerobic gut fungi are an untapped reservoir of natural products. Proc. Natl. Acad. Sci. USA 118. https://doi.org/10.1073/pnas.2019855118.
Swift, C.L., Malinov, N.G., Mondo, S.J., Salamov, A., Grigoriev, I.V., O'Malley, M.A., 2021b. A genomic catalog of stress response genes in anaerobic fungi for applications in bioproduction. Front. Fungal Biol. 2 https://doi.org/10.3389/ffunb.2021.708358.
Teunissen, M.J., Smits, A.A.M., Op den Camp, H.J.M., Huis in't Veld, J.H.J., Vogels, G.D., 1991. Fermentation of cellulose and production of cellulolytic and xylanolytic enzymes by anaerobic fungi from ruminant and non-ruminant herbivores. Arch. Microbiol. 156. Doi: 10.1007/BF00263000.
van Opijnen, T., Bodi, K.L., Camilli, A., 2009. Tn-seq: High-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms. Nat. Methods 6. https://doi.org/10.1038/nmeth.1377.
Wacker, T., Helmstetter, N., Wilson, D., Fisher, M.C., Studholme, D.J., Farrer, R.A., 2023. Two-speed genome evolution drives pathogenicity in fungal pathogens of animals. Proc. Natl. Acad. Sci. USA 120. https://doi.org/10.1073/pnas.2212633120.
Wang, Y., Youssef, N.H., Couger, M.B., Hanafy, R.A., Elshahed, M.S., Stajich, J.E., 2019. Molecular dating of the emergence of anaerobic rumen fungi and the impact of laterally acquired genes. mSystems 4. https://doi.org/10.1128/msystems.00247-19.
Weber, E.L., Cannon, P.M., 2007. Promoter choice for retroviral vectors: transcriptional strength versus trans-activation potential. Hum. Gene Ther. 18, 849–860. https://doi.org/10.1089/hum.2007.067.
Wessler, S.R., 2006. Transposable elements and the evolution of eukaryotic genomes. Proc. Natl. Acad. Sci. USA. https://doi.org/10.1073/pnas.0607612103.
Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramírez-González, R.H., De Oliveira, R., Mayer, K.F.X., Paux, E., Choulet, F., 2018. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biol. 19, 1–18. https://doi.org/10.1186/S13059-018-1479-0/FIGURES/8.
Wilken, S.E., Monk, J.M., Leggieri, P.A., Lawson, C.E., Lankiewicz, T.S., Seppälä, S., Daum, C.G., Jenkins, J., Lipzen, A.M., Mondo, S.J., Barry, K.W., Grigoriev, I.V., Henske, J.K., Theodorou, M.K., Palsson, B.O., Petzold, L.R., O'Malley, M.A., 2021. Experimentally validated reconstruction and analysis of a genome-scale metabolic model of an anaerobic neocallimastigomycota fungus. mSystems 6. https://doi.org/10.1128/msystems.00002-21.
Youssef, N.H., Couger, M.B., Struchtemeyer, C.G., Liggenstoffer, A.S., Prade, R.A., Najar, F.Z., Atiyeh, H.K., Wilkins, M.R., Elshahed, M.S., 2013. The genome of the anaerobic fungus orpinomyces sp. strain c1a reveals the unique evolutionary history of a remarkable plant biomass degrader. Appl. Environ. Microbiol. https://doi.org/10.1128/AEM.00821-13.
Zhang, R.G., Li, G.Y., Wang, X.L., Dainat, J., Wang, Z.X., Ou, S., Ma, Y., 2022. TEsorter: An accurate and fast method to classify LTR-retrotransposons in plant genomes. Hortic. Res. Doi: 10.1093/hr/uhac017.
Zheng, J., Ge, Q., Yan, Y., Zhang, X., Huang, L., Yin, Y., 2023. dbCAN3: automated carbohydrate-active enzyme and substrate annotation 51, W115–W121.