

MycoMobilome: a community-focused non-redundant database of transposable element consensus sequences for the fungal kingdom

Tobias Baril * and Daniel Croll *

Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, Rue Émile-Argand 11, 2000 Neuchâtel, Switzerland

*To whom correspondence should be addressed. Email: tobias.baril@unine.ch

Correspondence may also be addressed to Daniel Croll. Email: daniel.croll@unine.ch

Abstract

Transposable elements (TEs) are found in nearly all eukaryotic genomes. Despite significant advances in the sequencing of genomes, TE resources remain sparse, leading to a lack of traceability, reproducibility, and duplication of effort when annotating TEs. Here, we focus on the fungal kingdom and present MycoMobilome, a database of TE consensus sequences computationally curated using a set of 4309 genomes covering all major clades. The initial database contains 586 441 consensus sequences after filtering to remove putative host genes and low-quality consensus sequences. We provide a consistent naming convention to surface information on the confidence in the classification, including potential conflicting open reading frame functions, along with metadata to enable evaluation of TEs of interest and to determine whether further curation work is required on a case-by-case basis. Finally, we provide guidelines for community contributions and encourage researchers to deposit new or curated sequences, which will be incorporated into future MycoMobilome releases.

Introduction

Annotation of transposable elements (TEs) is essential for the analysis of genomes, even if the focus is primarily on coding sequences. TE annotation relies on the initial discovery of TE families, termed *de novo* curation, followed by homology-based annotation using tools to recognize all members of each TE family. Various approaches exist for *de novo* curation, with the most common including RepeatModeler2 [1] and TEde novo of REPET [2], whilst there are also databases providing previously curated TE families that can be used to annotate genomes, including Dfam [3, 4], TREP [5], RepBase [6], RepetDB [7], REXdb [8], and Yeast Transposons (<https://github.com/bergmanlab/yeast-transposons>) [9–11]. An exhaustive list of available databases is hosted at <https://tehub.org>.

The fungal kingdom spans almost a billion years of evolution, exhibits remarkable phenotypic and genomic diversity, and contains species found in nearly all ecological niches. The availability of genome resources for species spanning the kingdom is accelerating, partly due to efforts including the 1000 Fungal Genomes Initiative (<https://1000.fungalgenomes.org>) and broader efforts of the Darwin Tree of Life and Earth Biogenome projects [12, 13]. Despite these growing efforts, curated and multi-species TE resources for the fungal kingdom are almost non-existent. For example, in Dfam version 3.9, the only curated TEs are found for *Septoria linicola* and *Zymoseptoria tritici* (<https://dfam.org/>). Dedicated efforts to curate TE libraries for individual species include the plant pathogens *Magnaporthe oryzae* [14] and *Z. tritici* [15]. However, cross-referencing TE annotations across species has not been attempted, leading to reproducibility issues and duplica-

tion of effort as researchers perform annotations largely on a per-species basis.

Here, we present MycoMobilome (<https://github.com/TobyBaril/MycoMobilome>), a database of TE consensus sequences covering species across the fungal kingdom. We created MycoMobilome using all publicly available genome resources for fungi and provide researchers with a systematically generated TE consensus library with a persistent naming scheme to improve consistency and reproducibility. Consensus sequences are provided with labels to show whether their classification is supported by open reading frames (ORFs), and whether these consistently match the provided classification. Applying MycoMobilome classification to primary fungal models, we find that the genome fraction annotated as TEs increases by between +7.20% in *M. oryzae* to +13.41% in *Cryptococcus neoformans* compared to existing community TE annotations.

MycoMobilome is a key resource to facilitate investigations into TEs across the fungal kingdom and provides easy-to-implement methods for researchers to contribute new or updated sequences with attribution. Consistent application of MycoMobilome will improve consistency in TE naming conventions, whilst reducing the duplication of effort often accompanying repeat annotation in new genome assemblies.

Materials and methods

Initial curation for MycoMobilome

All publicly available genomes and associated proteomes for the fungal kingdom were sourced using Mycotoools (version

Received: November 12, 2025. Revised: February 16, 2026. Accepted: February 17, 2026

© The Author(s) 2026. Published by Oxford University Press.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0/>), which permits unrestricted reuse, distribution, and reproduction in any medium, provided the original work is properly cited.

0.32.3) [16]. Genomes were manually filtered to remove those under embargo, resulting in a final set of 4309 genome assemblies. Genome assembly metadata is provided in the MycoMobilome database (<https://doi.org/10.5281/zenodo.18243921>).

For each genome, a library of putative TE consensus sequences was generated using `earlGreyLibconstruct` in Earl Grey (v4.4.0) [17], configured with Dfam curated elements only (v3.7) [3, 4]. All 4309 *de novo* consensus libraries were subsequently combined into a single FASTA file containing 773 843 sequences. To generate the clustered non-redundant library, these sequences were clustered using the `easy-cluster` scalable cascaded clustering approach in MMseqs2 [18] with `--min-seq-id 0.8 -c 0.8 --cov-mode 1 --cluster-reassign` to cluster sequences to the 80-80-80 TE family rule [19], as adopted in very recent large-scale studies [20, 21], resulting in 354 315 non-redundant consensus sequences. The representative sequence for each cluster was extracted and labelled with the source genome (i.e. the genome from which the consensus sequence originated).

Autonomous TEs encode domains for their transposition activity, and the identity of these ORFs can be used to classify TEs [19, 22, 23]. Conversely, multicopy host proteins can erroneously be curated as repetitive sequences by automated TE curation methods, given their repetitive occurrence in host genomes. We translated all six frames of each TE consensus sequence using `transeq -clean -frame 6` in EMBOSS (v6.6.0) [24] and identified matches to known host proteins present in the Fungi partition of RefSeq release 228 [25] using Diamond BLASTp (v 2.1.11) [26] with `--sensitive --matrix BLOSUM62 --evalue 1e-3`. Following this, we detected similarity to characterized TE proteins using two complementary approaches [20, 21]. First, we used HMMscan in HMMER (v3.4) [27] with `-E 10 -noali` to detect homology to known TE protein domain HMM models supplied in ProfilesBank-ForREPET_Pfam35.0_GypsyDB as part of the REPET software suite [2, 28, 29]. Hits were filtered to retain those with $fseq_evalue \leq 0.001$ and $fseq_bitscore \geq 50$. Next, we used BLASTp (v2.14.1+) [30] to detect homology to known TE protein domains found in RepeatPeps.lib, which is part of RepeatMasker (v4.1.9) [31] and is used to classify TEs using the RepeatClassifier module in the initial *de novo* curation step in Earl Grey. A minimum *e*-value of $1e-3$ was defined, and nested hits were removed to retain the highest-quality protein hit for each query, followed by combining of adjacent and overlapping hits. Potential host gene hits were identified using the approach adopted by [20, 21, 32]. A TE consensus sequence was designated as a putative host gene if either: (i) there were hits to RefSeq queries and no hits to known TE queries; (ii) there were hits to both RefSeq queries and known TE queries, but at least 90 residues aligned to a RefSeq query did not overlap with alignments to known TE queries. Consensus sequences with no hits were retained in the TE library as putative non-autonomous TEs. In total, 24 571 consensus sequences were identified as potential host genes and removed.

To further refine the database, poor-quality TE consensus sequences were filtered. We define a poor-quality consensus sequence as <120 bp in length, as these are likely to be incomplete and poor quality. Further, majority-rule consensus sequences can sometimes contain unknown nucleotides, indicated with N in the sequence. We removed all sequences with $\geq 5\%$ N using `seqtk comp` (<https://github.com/lh3/seqtk>). In total, 187 402 consensus sequences were removed from the

unclustered library, and 53 103 consensus sequences were removed from the clustered library for not meeting the quality thresholds, resulting in final database sizes of 586 441 and 276 642 putative TE consensus sequences for the unclustered and clustered databases, respectively.

The majority of TE consensus sequences in publicly available databases remain uncured. Despite this, these sequences remain a widely used resource. To enable the community to evaluate TE annotation quality of newly analysed genomes, we introduce an automated classification of each TE consensus against protein hits to known TEs (tables are provided in the MycoMobilome database). Each TE consensus was labelled with a two-letter code providing the level of evidence supporting the TE classification: `\_PE` for protein evidence that matches the given classification, `\_DA` for protein evidence that contradicts the given classification, and `\_NE` for no protein evidence, which includes non-autonomous TEs.

No genome quality filtering was performed on our genome set to maximize taxonomic breadth. Consequently, some TE consensus sequences could be incomplete or low quality due to their curation in low-quality genomes. Consensus sequences in MycoMobilome that have been curated from genomes with an N50 below 50 kb and BUSCO score <90% complete are labelled with `LQGenome` in the header, enabling these to be identified and filtered if required. In total, 18 768 (3.2%) consensus sequences in the unclustered library were generated from lower quality genomes, whilst 11 970 (4.3%) consensus sequences in the clustered library are represented by consensus sequences from lower-quality genomes.

Finally, each TE consensus sequence in MycoMobilome was given a unique name. MycoMobilome families are defined based on the clustered database, following the 80-80-80 rule: `@[genus species]`. For the unclustered database, the consensus sequences are labelled with a family ID consistent with the clustered database, and an extra member ID to discriminate between consensus sequences that come from the same cluster: `@[genus species]`.

Comparison of existing TE annotation approaches with MycoMobilome

We wanted to assess the extent to which sampling TEs across fungal diversity enables improved detection of TE-derived genome content. To do this, we selected three well-studied organisms and compared TE annotation using MycoMobilome against the traditional approaches employed by the research communities working with each of these organisms. We compared our standardized approach with MycoMobilome against (i) annotation with known fungal TEs from RepBase and Dfam, as used in a recent study on *C. neoformans* [33]; (ii) a community annotation resource for *Candida albicans* (http://www.candidagenome.org/download/gff/C_albicans_SC5314/Assembly22/); (iii) annotation using a curated TE library with RepeatMasker for *M. oryzae* [14].

For *C. neoformans*, we adopted the same approach used in the original study and annotated the genome assembly for isolate CNA3 H99 (GCF_000149245.1) using RepeatMasker (v4.1.9) configured with Dfam (v3.9) and RepBase RepeatMasker Edition (release 20181026) and the options `--species fungi -norna -no\_is`. For *M. oryzae*, we sourced the reference genome assembly MG8 (GCF_000002495.2)

and annotated this with RepeatMasker using the TE consensus library provided with the publication [14]. For *C. albicans*, we obtained the genome assembly in FASTA format and the community feature annotation in GFF format for genome assembly SC5314 Assembly 22 from the Candida Genome portal (http://www.candidagenome.org/download/gff/C_albicans_SC5314/Assembly22/). To obtain a GFF of repeat regions, we filtered the feature GFF for features named `long_terminal_repeat`, `repeat_region`, and `retrotransposon`.

For comparison, we adopted our standardized approach as recommended with MycoMobilome. We annotated all three genome assemblies with `earlGreyAnnotationOnly` using Earl Grey (v6.3.2), providing MycoMobilome as the input library, with all other settings left as default.

Hits that are shared and unique to each methodological approach were identified using BEDTools intersect (v2.31.1) [34], and resultant feature coverage was calculated in R (v4.4.3) [35] with tidyverse [36]. For shared annotations, the width of the annotation in each case was calculated, rather than using a single length of annotation across compared methodologies, to account for variation in annotation size between methods. Venn diagrams were generated using the ggVennDiagram package [37].

Results and discussion

MycoMobilome assesses TE evidence across the fungal kingdom

The MycoMobilome release contains 276 642 consensus sequences, of which 39 265 have classifications supported by protein evidence. This provides a computationally curated resource for the diversity of TEs across the fungal tree of life. In particular, for clades without previous TE annotation efforts, this represents a very substantial expansion of genomic datasets. We performed filtering steps to stringently remove any consensus sequences with the potential to be host genes, as well as poor-quality consensus sequences. Following these efforts, we provide three versions of the database: a full database, a subset containing only TE consensus sequences with protein evidence, and a subset containing only TE consensus sequences lacking protein evidence. The compressed database files are provided via Zenodo (<https://doi.org/10.5281/zenodo.18243921>) in FASTA format, ensuring compatibility with a broad range of bioinformatics approaches. We provide guidance to non-specialists on how to best implement the MycoMobilome dataset into TE annotation pipelines (<https://github.com/TobyBaril/MycoMobilome>). We also provide a record of the genomes used to curate TEs across fungal diversity, and tables of hits to known TE proteins from multiple sources to cross-reference evidence supporting TE classifications for sequences of interest. To connect community-driven efforts and provide a centralized point for the genomics community, the unclustered consensus library has also been submitted to the Dfam consortium (dfam.org) [3, 4], where consensus sequences for species beyond fungi are also available.

A community-based curation effort through continuous submissions

We also provide a MycoMobilome community (<https://zenodo.org/communities/mycomobilome>), hosted on Zenodo,

with the aim of encouraging users to contribute new or improved (i.e. manually curated) TE consensus sequences to future releases. TE curation takes considerable work and expertise, which researchers should be recognized for. By adopting the Zenodo community approach, user contributions are assigned persistent DOIs, enabling contributors to be cited and recognized for their contributions to the wider genomics community. We provide guidance on how to contribute to MycoMobilome on the GitHub page, maximizing ease of use. Further, hosting such a resource on Zenodo reduces risks associated with longevity of the database. The nature of MycoMobilome as a database in FASTA format limits file sizes and enables efficient storage and reuse without extensive infrastructure or resource requirements.

MycoMobilome improves TE detection in important fungal models

By sampling TEs across fungal diversity, we increase the proportion of well-studied host genomes that is attributed to TEs (Fig. 1 and [Supplementary Table S1](#)). For the major rice and wheat pathogen, *M. oryzae*, which has a well-curated TE library and much attention focused on TE dynamics, MycoMobilome increases the proportion of the host genome annotated as TEs by 7.20% (2.95 Mb). The use of Earl Grey to annotate TEs also increases the proportion of the genome annotated as TEs for loci detected both in previous annotations and MycoMobilome by 10.57% (i.e. ‘shared’; 801.1 kb). The improved annotation is likely due to the incorporation of de-fragmentation and overlap resolution steps, consistent with previous observations [17]. To assess whether the increase in annotated TE content was driven by elements typically absent from manually curated databases, we re-annotated *M. oryzae* using only the subset of MycoMobilome containing TEs with detectable protein evidence. This resulted in a substantial reduction in total annotated sequence (54%; ~5.18 Mb), affecting both annotations unique to MycoMobilome (50% reduction; 3.28–1.40 Mb) and those shared with the original annotation approach (45% reduction; 7.58–4.16 Mb). Despite this reduction, the relative composition of annotations changed only modestly, with MycoMobilome-unique annotations comprising 23% of the total compared to 29% in the full database, and shared annotations increasing from 61% to 69% ([Supplementary Fig. S1](#)).

When annotating the genome of *C. neoformans* with MycoMobilome, we find an increase of host genome proportion annotated as TEs of 13.41% (2.54 Mb) compared to the previous annotation, which relied on fungal TEs hosted in Dfam and RepBase (Fig. 1). The large increase observed in *C. neoformans* likely arises from the absence of fungal TE families in existing databases, rather than over-annotation by MycoMobilome. Supporting this, 33.9% of the TE copies annotated using MycoMobilome come from consensus sequences curated in *Cryptococcus* species. In addition, a similar pattern is observed in *C. albicans*, where annotation using MycoMobilome increases the proportion of the genome recognized as TE by 9.82% (2.81 Mb) compared to the current community resource based on within-species TE characterization. The increase in unclassified TEs likely reflects the annotation of previously undetected, lineage-specific TEs lacking curated homologs. The proportion of unclassified TEs will decrease as representation of fungal lineages in curated TE reference databases, including MycoMobilome, increases. Hence,

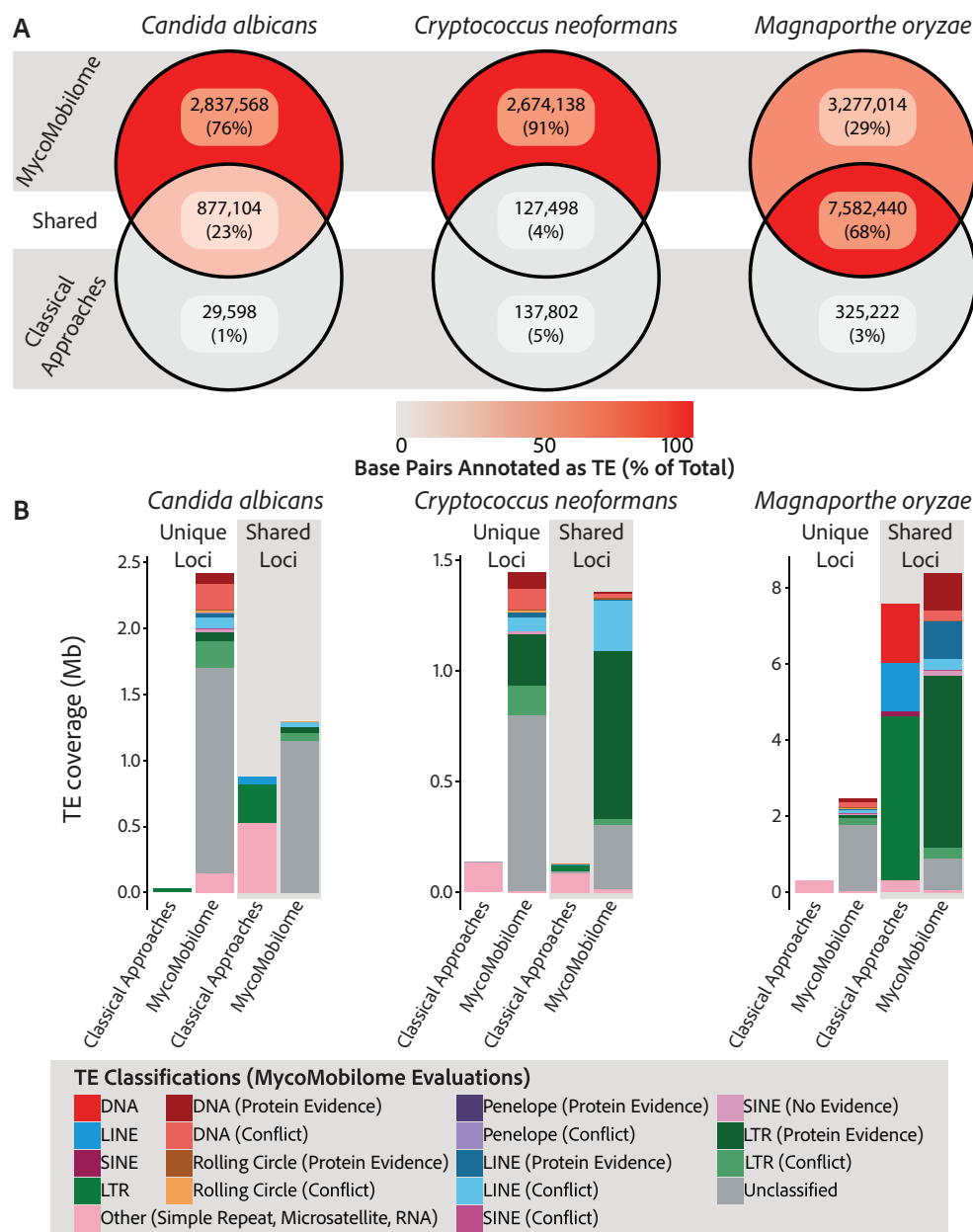


Figure 1. The use of MycoMobilome increases the fraction of genomes annotated as TEs in three key fungal pathogen models. **(A)** Venn diagrams illustrating the number of base pairs for each genome assembly that were annotated as TEs using either classical approaches or MycoMobilome (with Earl Grey). Depth of colour is proportional to the percentage of total base pairs annotated as TEs across all categories. Numbers in brackets show the percentage of all annotated base pairs that are found in each category, as labelled. **(B)** TE annotations split by approaches with which they are identified, and the classification of each annotated TE, as indicated in the key. Discrepancies in TE coverage at shared loci arise due to the defragmentation and overlap resolution steps automatically performed with Earl Grey following TE annotation with MycoMobilome. [Supplementary Fig. S1](#) reports a re-analysis of the genomes using only TE coverage with protein evidence.

MycoMobilome provides a substantial increase in annotated TEs even in comparatively well-studied fungal models, the magnitude of which is consistent with observed increases in annotated TE content in well-studied organisms following the development of Dfam [4], where annotated TE bases increased by 150 Mb (5.1% of total genome size) in the human genome (hg38), 144 Mb (5.5% of total genome size) in the mouse genome (mm10), 6.5 Mb (5.5% of total genome size) in the nematode genome (ce10), and 0.95 Mb (0.7% of total genome size) in the fruit fly genome (dm6).

TEs can persist in host genomes over long evolutionary time, over which they will experience mutation and degrada-

tion, leading to the accumulation of genomic fossils, challenging our ability to detect these sequences as TE-derived. However, in related lineages, these TE families may persist and be retained with higher levels of identity and at higher copy numbers, enabling their detection. By sampling TEs across deep evolutionary time and using this information to generate a non-redundant TE library, we were able to detect homology to TEs that may evade detection using single-species *de novo* approaches, which require high TE copy numbers to generate an initial consensus sequence. We show that the use of MycoMobilome increases the proportion of the host genome annotated as TEs, demonstrating the power of sampling

TEs across the fungal kingdom, and over deep evolutionary time, to be able to detect TE-derived sequences across fungal diversity.

Limitations and future directions

A high proportion of unclassified TEs is expected in fungal genomes when analysed across broad phylogenetic scales, reflecting the uneven and lineage-biased representation of fungi in curated TE reference databases. While manually curated TE libraries exist for several well-studied ascomycete and basidiomycete species, these resources are typically species-specific and capture only a subset of TE diversity when considered across fungal lineages, isolates, and evolutionary timescales. In MycoMobilome, unclassified elements therefore primarily reflect historic bias in TE curation towards focal taxa, rather than limitations of the annotation framework. As curated fungal TE families expand, consensus classifications will be iteratively re-evaluated and updated while maintaining stable MycoMobilome family identifiers to ensure consistency across versions.

A related limitation is that MycoMobilome relies on computationally curated TE consensus sequences, which enables broad taxonomic coverage but may introduce some misannotations. Post-processing steps, stringent quality filtering, and homology-based annotation mitigate this risk; however, future improvements through targeted manual curation of representative lineages and expansion of curated reference libraries will further enhance classification accuracy.

MycoMobilome prioritizes taxonomic breadth over strict genome assembly quality thresholds to capture the widest possible diversity of fungal TEs. Heterogeneity in assembly quality and lineage sampling may result in some fragmented or incomplete consensus sequences. Excluding lower-quality genomes, however, would disproportionately remove early-diverging lineages and poorly sampled clades, reinforcing existing biases in curated TE resources. By retaining these genomes and explicitly flagging sequences derived from lower-quality assemblies, MycoMobilome balances coverage with transparency.

De novo TE discovery pipelines, including RepeatModeler2, can generate truncated or chimeric consensus sequences, particularly in fragmented assemblies or complex repeat landscapes. MycoMobilome implements post-processing steps to reduce these issues. Similarly, clustering consensus sequences using the 80-80-80 rule provides a convenient but imperfect approximation of TE family structure: it may overmerge young families or oversplit ancient or highly diverged families, and performs poorly for some TEs, such as DIRS-like elements and non-LTR retrotransposons. To accommodate these limitations, MycoMobilome provides both clustered and unclustered versions of the database, enabling users to apply alternative thresholds or phylogeny-based approaches.

Automated host-gene removal presents a further trade-off, as fungal genomes contain domesticated transposases and TE-derived genes. MycoMobilome applies conservative filtering criteria to minimize host-gene contamination, which may result in exclusion of some genuine TE-derived sequences. This choice prioritizes annotation specificity while allowing excluded sequences to be reintroduced as curated TE and TE-derived protein resources expand. Consistent with this conservative design, independent similarity searches indicate that retained unclassified elements are not enriched for conserved

host genes ([Supplementary Table S2](#)), but instead predominantly match unannotated genomic regions or uncharacterized transcripts, supporting their interpretation as bona fide yet poorly curated TE-derived sequence.

Finally, many fungal TEs lack detectable ORFs or conserved domains due to rapid degeneration, nested insertions, or high sequence divergence. MycoMobilome preserves all original classifications assigned by RepeatClassifier and does not re-classify sequences based on ORF evidence. ORF and domain detection is provided solely as flags and metadata to indicate supporting evidence, allowing users to assess classification confidence on a per-element basis. As curated fungal TE protein models and reference libraries improve, these flags are expected to become increasingly informative while maintaining reproducible classification assignments.

Collectively, these limitations reflect the current state of fungal genome resources and TE curation. MycoMobilome is designed as a dynamic, community-oriented database: it provides broad taxonomic coverage and transparent annotations while creating a foundation for iterative refinement as genome quality, TE curation, and classification methods continue to improve.

Conclusions

Here, we introduce MycoMobilome, a database of TE consensus sequences for the fungal kingdom. We aim to provide this database as a key starting point to facilitate genomic investigations in lineages spanning the fungal kingdom, compatible with widely used TE annotation tools, including Earl Grey and RepeatMasker. Key features include consistent naming conventions, metadata on the quality of TE classifications, and an easy-to-follow tutorial for the community to process additional genomes. We welcome contributions to the MycoMobilome database and provide citable recognition for submitted TE annotations. We encourage the fungal research community to engage and expand the database scope to the benefit of all.

Acknowledgements

Author contributions: Tobias James Baril (Conceptualization [equal], Formal analysis [lead], Investigation [lead], Visualization [lead], Writing – original draft [lead], Writing – review & editing [equal]) and Daniel Croll (Conceptualization [equal], Funding acquisition [lead], Project administration [lead], Supervision [lead], Writing – review & editing [equal])

Supplementary data

[Supplementary data](#) is available at NAR Genomics & Bioinformatics online.

Conflict of interest

None declared.

Funding

D.C. was supported by the Swiss National Science Foundation (grant 201149).

Data availability

The MycoMobilome database has been deposited in the Zenodo database under DOI 10.5281/zenodo.18243921 (<https://doi.org/10.5281/zenodo.18243921>). User guidance and documentation are hosted on GitHub (<https://github.com/TobyBaril/MycoMobilome>) and under the DOI 10.5281/zenodo.17473060 (<https://doi.org/10.5281/zenodo.17473060>).

References

- Flynn JM, Hubley R, Goubert C. *et al.* RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci USA* 2020;117:9451–7. <https://doi.org/10.1073/pnas.1921046117>
- Flutre T, Duprat E, Feuillet C. *et al.* Considering transposable element diversification in *de novo* annotation approaches. *PLoS One* 2011;6:e16526. <https://doi.org/10.1371/journal.pone.0016526>
- Storer J, Hubley R, Rosen J. *et al.* The Dfam community resource of transposable element families, sequence models, and genome annotations. *Mobile DNA* 2021;12:2. <https://doi.org/10.1186/s13100-020-00230-y>
- Hubley R, Finn RD, Clements J. *et al.* The Dfam database of repetitive DNA families. *Nucleic Acids Res* 2016;44:D81–9. <https://doi.org/10.1093/nar/gkv1272>
- Wicker T, Matthews DE, Keller B. TREP: a database for Triticeae repetitive elements. *Trends Plant Sci* 2002;7:561–2. [https://doi.org/10.1016/S1360-1385\(02\)02372-5](https://doi.org/10.1016/S1360-1385(02)02372-5)
- Jurka J, Kapitonov VV, Pavlicek A. *et al.* Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 2005;110:462–7. <https://doi.org/10.1159/000084979>
- Amselem J, Cornut G, Choise N. *et al.* RepetDB: a unified resource for transposable element references. *Mobile DNA* 2019;10:6. <https://doi.org/10.1186/s13100-019-0150-y>
- Neumann P, Novák P, Hošťáková N. *et al.* Systematic survey of plant LTR-retrotransposons elucidates phylogenetic relationships of their polypeptide domains and provides a reference for element classification. *Mobile DNA* 2019;10:1. <https://doi.org/10.1186/s13100-018-0144-1>
- Nelson MG, Linheiro RS, Bergman CM. McClintock: an integrated pipeline for detecting transposable element insertions in whole-genome shotgun sequencing data. *G3* 2017;7:2763–78. <https://doi.org/10.1534/g3.117.043893>
- Czaja W, Bensasson D, Ahn HW. *et al.* Evolution of Ty1 copy number control in yeast by horizontal transfer and recombination. *PLoS Genet* 2020;16:e1008632. <https://doi.org/10.1371/journal.pgen.1008632>
- Carr M, Bensasson D, Bergman CM. Evolutionary genomics of transposable elements in *Saccharomyces cerevisiae*. *PLoS One* 2012;7:e50978. <https://doi.org/10.1371/journal.pone.0050978>
- Lewin HA, Robinson GE, Kress WJ. *et al.* Earth BioGenome Project: sequencing life for the future of life. *Proc Natl Acad Sci USA* 2018;115:4325–33. <https://doi.org/10.1073/pnas.1720115115>
- Blaxter M, Mieszkowska N, Palma FD. *et al.* Sequence locally, think globally: the Darwin Tree of Life Project. *Proc Natl Acad Sci* 2022;119:e2115642118.
- Lin L, Sun T, Guo J. *et al.* Transposable elements impact the population divergence of rice blast fungus *magnaporthe oryzae*. *mBio* 2024;15:e0008624. <https://doi.org/10.1128/mbio.0008624>
- Baril T, Croll D. A pangenome-guided manually curated library of transposable elements for *Zymoseptoria tritici*. *BMC Res Notes* 2023;16:335. <https://doi.org/10.1186/s13104-023-06613-7>
- Konkel Z, Slot JC. Mycotools: an automated and scalable platform for comparative genomics. *bioRxiv*, <https://doi.org/10.1101/2023.09.08.556886>, 12 September 2023, preprint; not peer reviewed.
- Baril T, Galbraith J, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. *Mol Biol Evol* 2024;41:msae068. <https://doi.org/10.1093/molbev/msae068>
- Steinberger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* 2017;35:1026–8. <https://doi.org/10.1038/nbt.3988>
- Wicker T, Sabot F, Hua-Van A. *et al.* A unified classification system for eukaryotic transposable elements. *Nat Rev Genet* 2007;8:973–82. <https://doi.org/10.1038/nrg2165>
- Muller H, Savaiaar R, Peccoud J. *et al.* Phylogenetic relatedness rather than aquatic habitat fosters horizontal transfer of transposable elements in animals. *Genome Res* 2025;35:2011–22.
- Romeijn J, Bañales I, Seidl MF. Extensive horizontal transfer of transposable elements shapes fungal mobilomes. *Curr Biol* 2025;36:355–69.e4. <https://doi.org/10.1016/j.cub.2025.11.012>
- Wells JN, Feschotte C. A field guide to eukaryotic transposable elements. *Annu Rev Genet* 2020;54:539–61. <https://doi.org/10.1146/annurev-genet-040620-022145>
- Hayward A, Gilbert C. Transposable elements. *Curr Biol* 2022;32:R904–9. <https://doi.org/10.1016/j.cub.2022.07.044>
- Rice P, Longden L, Bleasby A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* 2000;16:276–7. [https://doi.org/10.1016/S0168-9525\(00\)00204-2](https://doi.org/10.1016/S0168-9525(00)00204-2)
- Goldfarb T, Kodali VK, Pujar S. *et al.* NCBI RefSeq: reference sequence standards through 25 years of curation and annotation. *Nucleic Acids Res* 2025;53:D243–57. <https://doi.org/10.1093/nar/gkac1038>
- Buchfink B, Reuter K, Drost H-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 2021;18:366–8. <https://doi.org/10.1038/s41592-021-01101-x>
- Sonnhammer EL, Eddy SR, Birney E. *et al.* Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* 1998;26:320–2. <https://doi.org/10.1093/nar/26.1.320>
- Inizan O. PASTEC: an automatic transposable element classification tool. *PLoS One* 2014;9:e91929.
- Quesneville H, Bergman CM, Andrieu O. *et al.* Combined evidence annotation of transposable elements in genome sequences. *PLoS Comp Biol* 2005;1:166–75. <https://doi.org/10.1371/journal.pcbi.0010022>
- Camacho C, Coulouris G, Avagyan V. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* 2009;10:1–9. <https://doi.org/10.1186/1471-2105-10-421>
- Smit AFA, Hubley RR, Green PR. RepeatMasker Open-4.0. 2013. <http://repeatmasker.org>. (1 November 2025, date last accessed).
- Zhang H-H, Peccoud J, Xu M-R-X. *et al.* Horizontal transfer and evolution of transposable elements in vertebrates. *Nat Commun* 2020;11:1–10.
- Priest SJ, Yadav V, Roth C. *et al.* Uncontrolled transposition following RNAi loss causes hypermutation and antifungal drug resistance in clinical isolates of *Cryptococcus neoformans*. *Nat Microbiol* 2022;7:1239–51. <https://doi.org/10.1038/s41564-022-01183-z>
- Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 2010;26:841–2. <https://doi.org/10.1093/bioinformatics/btq033>
- R Core Team. R: a language and environment for statistical computing. Vienna, Austria, 2025. <https://www.R-project.org/>
- Wickham H, Averick M, Bryan J. *et al.* Welcome to the Tidyverse. *JOSS* 2019;4:1686. <https://doi.org/10.21105/joss.01686>
- Gao C-H, Yu G, Cai P. GgVennDiagram: an intuitive, easy-to-use, and highly customizable R package to generate venn diagram. *Front Genet* 2021;12:706907. <https://doi.org/10.3389/fgene.2021.706907>