BMC Research Notes

DATA NOTE

Open Access

# A pangenome-guided manually curated library of transposable elements for *Zymoseptoria tritici*

Tobias Baril[1] and Daniel Croll[1*]

## Abstract

**Objectives** High-quality species-specific transposable element (TE) libraries are required for studies to elucidate the evolutionary dynamics of TEs and gain an understanding of their impacts on host genomes. Such high-quality TE resources are severely lacking for species in the fungal kingdom. To facilitate future studies on the putative role of TEs in rapid adaptation observed in the fungal wheat pathogen *Zymoseptoria tritici*, we produced a manually curated TE library. This was generated by detecting TEs in 19 reference genome assemblies representing the global diversity of the species supplemented by multiple sister species genomes. Improvements over previous TE libraries have been made on TE boundary resolution, detection of ORFs, TE domains, terminal inverted repeats, and class-specific motifs.

**Data description** A TE consensus library for *Z. tritici* formatted for use with RepeatMasker. This data is relevant to other researchers investigating TE-host evolutionary dynamics in *Z. tritici* or who are interested in comparative studies of the fungal kingdom. Further, this TE library can be used to improve gene annotation. Finally, this TE library increases the number of manually curated TE datasets, providing resources to further our understanding of TE diversity.

**Keywords** Transposable Elements, Zymoseptoria tritici, Transposons, TE library, Manual curation, TE consensus

## Objective

Transposable elements (TEs) are autonomous DNA sequences that can move within the genome. TEs have been implicated in host genome evolution through processes including chromosomal rearrangements, exon shuffling, and donation of coding sequences [1–4]. TEs are highly diverse among eukaryotes and current levels of sampling are insufficient to gain a deep understanding of the evolutionary dynamics of TEs. Compounding this, databases are inundated with putative TE sequences, however only a small fraction of these are curated. For example, in Dfam release 3.7, only 19,730 families (0.57%) are curated out of a total 3,437,876 families [5, 6].

To facilitate evolutionary studies, species-specific TE libraries are needed as TE content can vary significantly, even within a single genus [7]. Thus, TE libraries for even closely related species are not sufficient to accurately characterize the TE content of a genome. Further, TE resources for Fungi are lacking and this impedes studies focusing on genome evolution in this extremely diverse kingdom. *Zymoseptoria tritici* is a fungal wheat pathogen with extensive standing genetic variation within and among distinct populations across the globe [8, 9]. Consequently, parallel evolution across geographic regions has enabled the pathogen to rapidly overcome host resistance and tolerate fungicides in extremely short timeframes [10]. This rapid adaptation, combined with variable TE loads within and among populations [8],

*Correspondence:
Daniel Croll
daniel.croll@unine.ch
[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchatel, Rue -Argand 11, 2000 Neuchatel, Switzerland

**Table 1** Overview of data files/data sets

| Label | Name of data file/data set | File types (file extension) | Data repository and identifier (DOI or accession number) |
|---|---|---|---|
| Data set 1 | ZymTri_2023.manCurTE.v1_0.fasta | FASTA (.fasta) | Zenodo (https://doi.org/10.5281/zenodo.8379981) [38] |
| Data set 2 | *.TE_2023.gff | General Feature Format (GFF) | Zenodo (https://doi.org/10.5281/zenodo.8390461) [39] |

makes *Z. tritici* a fascinating system for investigations into TE-host evolutionary dynamics.

To enable studies on evolutionary dynamics of TEs in *Z. tritici*, we present an improved manually curated TE consensus library constructed from a 19-genome reference panel and the sister species of *Z. tritici*. Improvements have been made on TE boundary resolution, detection of ORFs, TE domains, terminal inverted repeats, and class-specific motifs. We have also reduced redundancy in the library.

## Data description

Putative TE consensus sequences were first obtained by annotating all 23 reference-quality genome assemblies [11–25] with Earl Grey (v3.0; https://github.com/TobyBaril/EarlGrey) [26, 27]. Consensus sequences generated from each reference genome were clustered using CD-Hit-Est (v4.8.1) [28, 29] to group sequences with 90% similarity across 80% of the longer sequence length to reduce redundancy whilst preventing the collapsing of chimeric sequences. Each consensus sequence was then subject to manual curation as described by Goubert et al. (2022) [30]. Briefly, genomic copies of each TE were obtained using a "BLAST, Extract, Align, Trim" process to recover genomic copies from each of the 23 reference genome assemblies with 1000 flanking bases at either end [30, 31]. For families with > 100 BLASTN hits, the 25 longest hits were selected, along with 75 random hits. Multiple alignments were generated for each putative TE family using MAFFT (v7.505) with the –auto flag [32]. Columns composed of > = 80% gaps were removed with T-COFFEE (v13.45.0.4846264) [33]. Subsequently, all sequence alignments were manually curated to define TE boundaries and remove regions of low conservation and rare insertions. Following manual curation, new majority-rule consensus sequences were generated with EMBOSS (v6.6.0.0) cons [34]. TE-Aid (https://github.com/clemgoub/TE-Aid/) was used to aid visual inspection and to identify diagnostic features for classification of extended consensus sequences. Following this, TIRs were recorded (if present), and nhmmscan (HMMER v3.3.2) [35] was used to identify homology to known curated elements in Dfam (v3.7). Combining this information, each TE consensus sequence was manually classified using available information following the naming convention '>ZymTri_2023_family_[n]#[Classification]/[Family]' for compatibility with RepeatMasker [36]. Consensus sequences classified with low confidence have a '?' added to the name, as well as the string '_LowConf'. To reduce redundancy in the final TE library, sequences were clustered to the family-level using the '80–80-80 rule' (*i.e.* ≥ 80% identity, ≥ 80% length, ≥ 80 bp) [30, 37] implemented in CD-hit-est. The representative sequence for each cluster was manually selected to retrieve the sequence with the highest classification confidence, also defined as the 'most intact consensus'. Chimeric sequences erroneously clustered were manually separated to retain sequences for the chimeric TE and the individual elements that generated the chimera.

In total, we curated 331 distinct consensus sequences for the final TE library (Table 1). Of these, 199 could be confidently classified and 105 consensus sequences remain putative TEs labelled in the library as 'Unclassified'. The 27 remaining TE consensus sequences are classified with low confidence. TE families from all major classifications are present: 92 DNA transposons, 22 long interspersed nuclear elements (LINEs), 65 long terminal repeat retrotransposons (LTRs), 31 miniature inverted terminal repeat elements (MITEs), 11 rolling circles (also known as helitrons), 1 short interspersed nuclear element (SINE), 1 terminal-repeat retrotransposon in miniature (TRIM), and 105 unclassified elements. The TE consensus library in FASTA format is supplied in data set 1 (Table 1) and a Tar archive containing the annotation of the 19 reference genomes in GFF format is supplied in data set 2 (Table 1).

## Limitations

Whilst we made use of large public databases and an extensive set of genomes, 105 TE consensus remain unclassified and an additional 27 are classified with low confidence. Further manual curation efforts following sampling of more genome assemblies might aid in the classification of these by providing additional diagnostic features.

Limited knowledge on the diversity of TEs across the fungal kingdom may have impacted our ability to classify sequences to family-level. We anticipate this limitation will become less significant as genomic sampling and TE curation across the kingdom expands.

The integration of nearly two dozen reference-quality genomes significantly improved our ability to identify even low-copy TEs in the species. However, the dynamic nature of TE activation and repression within the species [8, 40] poses a significant challenge to capture the full TE content of the species. Hence, some recently reactivated TEs or very low-copy number TEs might have evaded detection, and so be missing from the final library.

**Abbreviations**

| | |
|---|---|
| TEs | Transposable elements |
| LINE | Long Interspersed Nuclear Element |
| LTR | Long Terminal Repeat |
| MITE | Miniature Inverted Terminal repeat Element |
| SINE | Short Interspersed Nuclear Element |
| TRIM | Terminal Repeat In Miniature |

**Acknowledgements**
Not applicable.

**Author contributions**
TB and DC conceived the study. Analyses were performed by TB. TB wrote the manuscript with input from DC. All authors read and approved the final manuscript.

**Funding**
TB was supported by a grant from the Swiss National Science Foundation awarded to DC (Grant Number: 201149).

**Availability of data and materials**
The manually-curated TE library for *Zymoseptoria tritici* can be freely accessed on Zenodo at https://doi.org/10.5281/zenodo.8379981 [38]. The TE annotations for the 19 *Zymoseptoria tritici* reference genomes can be freely accessed on Zenodo at https://doi.org/10.5281/zenodo.8390461 [39].

## Declarations

**Ethics approval and consent to participate**
Not applicable.

**Consent for publication**
Not applicable.

**Competing interests**
The authors declare that they have no competing interests.

Received: 29 September 2023 Accepted: 3 November 2023
Published online: 16 November 2023

**References**

1. Feschotte C, Pritham EJ. DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet. 2007;41:331–68.
2. Mita P, Boeke JD. How retrotransposons shape genome regulation. Curr Opin Genet Dev. 2016;37:90–100.
3. Bennetzen JL. Transposable element contributions to plant genome evolution. Plant Mol Biol. 2000;42:251–69.
4. Coates BS, Hellmich RL, Grant DM, Abel CA. Mobilizing the genome of lepidoptera through novel sequence gains and end creation by non-autonomous Lep1 Helitrons. DNA Res. 2012;19:11–21.
5. Hubley R, Finn RD, Clements J, Eddy SR, Jones TA, Bao W, et al. The Dfam database of repetitive DNA families. Nucleic Acids Res. 2016;44:D81–9.
6. Storer J, Hubley R, Rosen J, Wheeler TJ, Smit AF. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mob DNA. 2021;12:2.
7. Baril T, Hayward A. Migrators within migrators: exploring transposable element dynamics in the monarch butterfly. Danaus plexippus Mob DNA. 2022;13:5.
8. Feurtey A, Lorrain C, McDonald MC, Milgate A, Solomon PS, Warren R, et al. A thousand-genome panel retraces the global spread and adaptation of a major fungal crop pathogen. Nat Commun. 2023;14:1059.
9. Singh NK, Karisto P, Croll D. Population-level deep sequencing reveals the interplay of clonal and sexual reproduction in the fungal wheat pathogen Zymoseptoria tritici. Microb Genom. 2021. https://doi.org/10.1099/mgen.0.000678.
10. Petit-Houdenot Y, Lebrun M-H, Scalliet G. Understanding plant-pathogen interactions in Septoria tritici blotch infection of cereals. In: Oliver R, editor. Achieving durable disease resistance in cereals. London: Burleigh Dodds Science Publishing; 2021. p. 263–302.
11. Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biol. 2020;18:12.
12. Goodwin SB, M'barek SB, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 2011;7:e1002070.
13. Plissonneau C, Hartmann FE, Croll D. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 2018;16:5.
14. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Genome compartmentalization predates species divergence in the plant pathogen genus Zymoseptoria. BMC Genomics. 2020;21:588.
15. Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. Zymoseptoria tritici Global Population Raw RNA sequence reads. 2020. European Nucleotide Archive. https://identifiers.org/bioproject:PRJNA559981.
16. Goodwin SB, M'barek SB, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Zymoseptoria tritici IPO323, whole genome shotgun sequencing project. 2011. European Nucleotide Archive. https://www.ebi.ac.uk/ena/browser/view/ACPE01000000.
17. Plissonneau C, Hartmann FE, Croll D. A small secreted protein in Zymoseptoria tritici is the avirulence factor for the major wheat resistance gene Stb6. 2016. European Nucleotide Archive. https://identifiers.org/bioproject:PRJEB15648.
18. Plissonneau C, Hartmann FE, Croll D. Complete genome assembly of the Zymoseptoria tritici isolate ST99CH_1E4. 2017. European Nucleotide Archive. https://identifiers.org/bioproject:PRJEB20900.
19. Plissonneau C, Hartmann FE, Croll D. Complete genome assembly of the Zymoseptoria tritici isolate ST99CH_3D1. 2017. European Nucleotide Archive. https://identifiers.org/bioproject:PRJEB20899.
20. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Long-read sequencing of Zymoseptoria passerinii, strain Zpa63. 2020. European Nucleotide Archive. https://identifiers.org/bioproject:PRJNA638605.
21. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Transcriptomics of Zymoseptoria passerinii, Zpa63, in axenic growth. 2020. European Nucleotide Archive. https://identifiers.org/bioproject:PRJNA639021.
22. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Long-read sequencing of Zymoseptoria brevis, Zb87. 2020. European Nucleotide Archive. https://identifiers.org/bioproject:PRJNA638553.
23. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Long-read sequencing of Zymoseptoria pseudotritici, Zp13. 2020. European Nucleotide Archive. https://identifiers.org/bioproject:PRJNA638515.
24. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. Long-read sequencing of Zymoseptoria ardabiliae, Za17. 2020. European Nucleotide Archive. https://identifiers.org/bioproject:PRJNA638382.
25. Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, et al. De novo genome assemblies of Zymoseptoria tritici natural isolates Assembled using PacBio. 2018. https://identifiers.org/bioproject:PRJNA414407.
26. Baril T, Imrie RM, Hayward A. Earl Grey: a fully automated user-friendly transposable element annotation and analysis pipeline. BioRxiv. 2022;4:1686.

27. Baril T, Galbraith J, Hayward A. Earl Grey. 2023. Zenodo. https://doi.org/10.5281/zenodo.5654615.
28. Li W, Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics. 2006;22:1658–9.
29. Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28:3150–2.
30. Goubert C, Craig RJ, Bilat AF, Peona V, Vogan AA, Protasio AV. A beginner's guide to manual curation of transposable elements. Mob DNA. 2022;13:7.
31. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinform. 2009;10:1–9.
32. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013;30:772–80.
33. Notredame C, Higgins DG, Heringa J. T-coffee: a novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302:205–17.
34. Rice P, Longden L, Bleasby A. EMBOSS: the European molecular biology open software suite. Trends Genet. 2000;16:276–7.
35. Wheeler TJ, Eddy SR. nhmmer: DNA homology search with profile HMMs. Bioinformatics. 2013;29:2487–9.
36. Smit AFA, Hubley RR, Green PR. RepeatMasker Open-4.0. http://repeatmasker.org. 2013.
37. Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. Nat Rev Genet. 2007;8:973–82.
38. Baril T, Croll D. Zymoseptoria tritici pangenome-guided TE consensus library. 2023. Zenodo. https://doi.org/10.5281/zenodo.8379981.
39. Laboratory of Evolutionary Genetics @ UNINE. Zymoseptoria tritici pangenome resources. 2023. Zenodo. https://doi.org/10.5281/zenodo.8390461.
40. Badet T, Feurtey A, Croll D. Recent reactivation of a pathogenicity-associated transposable element triggers major chromosomal rearrangements in a fungal wheat pathogen. bioRxiv. 2023;105:11845.


## Publisher's Note