THE ROLE OF TRANSPOSABLE ELEMENTS IN THE ADAPTIVE EVOLUTION OF FUNGI

A thesis submitted to

UNIVERSITY OF NEUCHÂTEL

For the degree of

DOCTOR OF SCIENCE

Presented by

URSULA OGGENFUSS

Accepted on the recommendation of

Professor Dr Daniel Croll, Université de Neuchâtel (thesis director)

Professor Dr Pilar Junier, Université de Neuchâtel (internal expert)

Professor Dr Anne Roulin, University of Zurich (external expert)

Professor Dr Anna Selmecki, University of Minnesota (external expert)

Defended on

15.03.2022

Imprimatur

UNIVERSITÉ DE NEUCHÂTEL

Faculté des Sciences
Secrétariat-décanat de Faculté
Rue Emile-Argand 11
2000 Neuchâtel – Suisse
Tél : + 41 (0)32 718 21 00
E-mail : secretariat.sciences@unine.ch

IMPRIMATUR POUR THESE DE DOCTORAT

La Faculté des sciences de l'Université de Neuchâtel autorise l'impression de la présente thèse soutenue par

Madame Ursula OGGENFUSS

Titre:

"The role of transposable elements in the adaptative evolution of fungi"

sur le rapport des membres du jury composé comme suit:

- Prof. Daniel Croll, directeur de thèse, Université de Neuchâtel, Suisse
- Prof. Pilar Junier, Université de Neuchâtel, Suisse
- Ass. Prof. Anna Selmecki, University of Minnesota, USA
- Prof. ass. Anne Roulin, Université de Zürich

Neuchâtel, le 24 mars 2022

Le Doyen, Prof. A. Bangerter

Summary

Transposable elements (TEs) are repetitive and mobile genetic elements. Autonomous TEs contain all coding regions needed for excision, duplication and insertion. TEs can be considered to be selfish elements or genomic parasites, with the potential to dramatically disturb the genome by gene disruption, deletion of regions or change in gene expression. However, dramatic distortion also carries the potential to provide beneficial new functions. An increasing body of research highlights beneficial impacts of TE insertions that increase adaptability to new environmental conditions or toxic impacts, or become part of the proteome. In this thesis, we studied how TEs influence fungal species from a local and short-term impact to deeper evolutionary time scales, and how TEs themselves are evolving. In the first chapter, we studied population dynamics and the impact of TEs on genome size evolution in a fungal plant pathogen. Despite a strong impact of purifying selection, we detected an increase in TE copies from the population of origin to more recently established populations, with a strong recent burst in North America. Increase in TE copy numbers is strongly correlated with an increase in genome size as well. Yet, which TEs lead to a copy number increase, and where active copies insert was not clear. We therefore studied expansion routes of TEs in a number of high-quality genomes in the second chapter. We found that older elements are accumulated in regions with low gene contents and high indication of an ascomycete-specific defense mechanisms against TEs. Insertions that are part of a recent burst are generally closer located to genes and not yet affected by defense mechanisms. To study the impact of TEs on the proteome evolution, we analyzed species covering the fungal kingdom in the third chapter. We screened for proteins with indications to be host-TE fusions. We found an accumulation of host-TE fusions in Saccharomycotina that generally have a lower gene- and TE content. The majority of host-TE fusions has a helicase TE domain, indicating a strong impact of DNA binding functions. Furthermore, we found several host-TE fusions that are potentially involved in defense mechanisms against TEs.

Keywords: Transposable elements, Population dynamics, Genome evolution

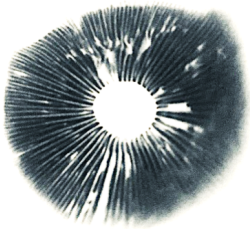
Résumé

Les éléments transposables (ETs) sont des éléments génétiques répétitifs et mobiles. Les ET autonomes contiennent toutes les séquences codantes nécessaires à leur excision, duplication et insertion. Les ET peuvent être considérés comme des éléments égoïstes ou des parasites génomiques, ayant le potentiel de dramatiquement perturber le génome par l'interruption de gènes, la délétion de séquences ou la modification de l'expression des gènes. Cependant, ces perturbations peuvent également offrir de nouvelles fonctions avantageuses. Des recherches récentes ont souligné les fonctions bénéfiques de certaines insertions d'ET, d'augmenter l'adaptabilité à de nouvelles conditions environnementales ou à des impacts toxiques, ou prenant part au protéome. Dans cette thèse, nous avons étudié comment les ET influencent les espèces fongiques, d'un impact local et à court terme à des échelles de temps évolutives plus profondes, et comment les ETs eux-mêmes évoluent sous des processus neutres et sélectifs. Dans le premier chapitre, nous avons étudié la dynamique des populations et l'impact des ETs sur l'évolution de la taille du génome chez un champignon phytopathogène. Malgré un fort impact de la sélection négative sur les ET, nous avons détecté une augmentation du nombre de copies entre la population d'origine et les populations plus récemment établies, avec une forte augmentation récente en Amérique du Nord. L'augmentation du nombre de copies de ET est également fortement corrélée à une augmentation de la taille du génome. Cependant, il n'est pas clair quels ETs conduisent à une augmentation de leur nombre de copies, et où les copies actives s'insèrent. Dans un deuxième chapitre, nous avons donc étudié les voies d'expansion des ETs dans plusieurs génomes de haute qualité. Nous avons constaté que les ET les plus anciens s'accumulent dans les régions avec peu de gènes et une forte indication de mécanismes de défense contre les ET. À l'inverse, les insertions les plus récentes sont généralement situées plus près de gènes et ne sont pas encore affectées par les mécanismes de défense. Pour étudier l'impact des ETs sur l'évolution du protéome, nous avons dans un troisième chapitre porté nos études sur un ensemble d'espèces couvrant le règne des champignons. Nous avons identifié les protéines qui chez ces différentes espèces présentent les indications d'une fusion entre protéines hôte et ET. Nous avons trouvé une accumulation de fusions hôte-ET chez la sous-classe des Saccharomycotina, qui de façon générale ont un faible nombre de gènes et d'ETs. La majorité des fusions hôte-ET possède un domaine hélicase, indiquant un fort impact des fonctions de liaison à l'ADN. De plus, nous avons trouvé plusieurs fusions hôte-ET qui sont potentiellement impliquées dans des mécanismes de défense contre les ET.

Mots-clés : Éléments transposables, Dynamique des populations, Évolution du génome.

Table of Contents

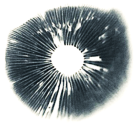
Scientific output



Talks and posters

- Biology22, University of Basel, Switzerland, 2022 (talk)
- 5th Uppsala Transposon Symposium, 2021 (online lightning talk & poster)
- Annual PhD students meeting, Neuchâtel, Switzerland, 2021 (talk)
- EMBO workshop “The mobile Genome”, virtual meeting, 2021 (poster)
- CSHL Biology of Genomes 2021 virtual meeting, 2021 (poster)
- 4th Uppsala Transposon Symposium, 2020 (online lightning talk & poster)
- #TEjournalClub, April 2020 (virtual talk)
- Biology20, Fribourg, Switzerland, 2020 (talk)
- Zurich Mycology Symposium, Agroscope Wädenswil, Switzerland, 2020 (talk)
- 3rd Uppsala Transposon Symposium, Uppsala, Sweden, 2019 (talk)
- Host-Microbes Genomics meeting, Zurich, Switzerland, 2019 (talk)
- Ecology & Evolution Days, Fribourg, Switzerland, 2019 (speed talk & poster)
- Host-Microbes Genomics 2018, ETH Zurich, Switzerland, 2018 (talk)
- Zymoseptoria tritici community meeting, ETH Zurich, Switzerland, 2018 (talk)
- Evolution 2018: joint congress on evolutionary biology, Montpellier, France, 2018 (talk)
- Biology18, Neuchâtel, Switzerland, 2018 (talk)
- Rigi-workshop 2018, Rigi, Switzerland, 2018 (poster & speed talk)
- Zurich Mycology Symposium, Agroscope Reckenholz, Switzerland, 2018 (talk)
- Host-Microbes Genomics 2017, Neuchâtel, Switzerland, 2017 (speed talk)
- Zymoseptoria tritici community meeting, Kiel, Germany, 2017 (poster)
- Annual PhD students meeting, Neuchâtel, Switzerland, March 2017 (poster)

Publications



González-Sayer S, **Oggenfuss U**, García I, Aristizabal F, Croll D, Riaño-Pachon DM, High-quality genome assembly of Pseudocercospora ulei the main threat of natural rubber trees. 2022. Genetics and Molecular Biology 45:1–5

Fouché S, **Oggenfuss U**, Chanclud E, Croll D. 2021. A devil's bargain with transposable elements in plant pathogens. Trends in Genetics (*published ahead of print*)

Oggenfuss U, Badet T, Wicker T, Hartmann FE, Singh NK, Abraham N, Karisto P, Vonlanthen T, Mundt CC, McDonald BA, Croll, D. 2021. A population-level invasion by transposable elements in a fungal pathogen. eLife 10: 1–25.

Pereira D, **Oggenfuss U**, McDonald BA, Croll D. 2021. The population genomics of transposable element activation in the highly repressive genome of an agricultural pathogen. Microbial Genomics 7: 1–15.

Lorrain C, **Oggenfuss U**, Croll D, Duplessis S, Stukenbrock E. 2021. Transposable Elements in Fungi: Coevolution With the Host Genome Shapes, Genome Architecture, Plasticity and Adaptation. *In: Encyclopedia of Microbiology*

Torres DE*, **Oggenfuss U***, Croll D, Seidl MF. 2020. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biology Reviews 34: 136–143. (* shared first)

Badet T, **Oggenfuss U**, Abraham L, McDonald BA, Croll D. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biology 18: 1–18.

Fouché S, Badet T, **Oggenfuss U**, Plissonneau C, Francisco CS, Croll D. 2020. Stress-Driven Transposable Element de-repression Dynamics and Virulence Evolution in a Fungal Pathogen. Molecular biology and evolution 37: 221–239.

Tralamazza SM, Rocha LO, **Oggenfuss U**, Corrêa B, Croll D. 2019. Complex evolutionary origins of specialized metabolite gene cluster diversity among the plant pathogenic fungi of the Fusarium graminearum species complex. Genome Biology and Evolution 11: 3106–3122.

Unpublished manuscripts

Stalder L, **Oggenfuss U**, Mohd-Assaad N, Croll D. The population genetics of adaptation through copy-number variation in a fungal plant pathogen (*in revision in Molecular Ecology*)

Oggenfuss U, Croll D. Expansion routes of transposable elements across the genome

Oggenfuss U, Badet T, Croll D. Co-option of transposable elements across the fungal kingdom

Fouché S, Zala M, **Oggenfuss U**, McDonald B. A, Croll D. Repeated chromosomal degeneration through repeat-induced chromosomal rearrangements in a fungal pathogen

Abraham N, **Oggenfuss U**, Croll D. Locus specific expression of TE in a pathogen population

González-Sayer S, **Oggenfuss U**, García I, Aristizabal F, Riaño-Pachon DM, Croll. Characterization of the transposable element composition in the *P. ulei* genome.

Introduction

Transposable elements (TEs) are genetic sequences able to autonomously or non-autonomously proliferate in their host's genome. Autonomous TEs encode the proteins necessary for both self-replication and mobility. After the loss of coding regions, TEs can still be active as non-autonomous parasites of the corresponding autonomous element. TE insertions induce structural variant mutations that strongly impact genome integrity, often generating deleterious impacts. A disregard of the host genome integrity, a gene set only providing replication of themselves and then no clear visible phenotypic benefit coined terms like "junk DNA" or "selfish elements" (Doolittle & Sapienza, 1980). Yet, if the impact of TEs would be solely deleterious, one would expect TEs to be purged from most genomes. Indeed, recent evidence shows TEs to be major players in local adaptation, gene evolution and even to impact speciation and general evolution over longer periods of time.

Barbara McClintock and the discovery of transposable elements

"I was just so interested in what I was doing, I could hardly wait to get up in the morning and get at it. One of my friends, a geneticist, said I was a child, because only children can't wait to get up in the morning to get at what they want to do."
Barbara McClintock

Figure 1: Barbara McClintock, 1983. Photography by David Miklos, Cold Spring Harbor Laboratory Archives.

No research on TEs would be possible without the sharp thinking and meticulous experiments of the agricultural scientist, cytologist and geneticist Barbara McClintock (Kass, 2003). During her pioneering studies on maize ears in the 1950s, McClintock predicted the presence of mobile "controlling elements" as being responsible for color differences in the triploid aleurone layer of maize kernels (McClintock, 1953; Fedoroff *et al.*, 1983). To consider the mobile property, controlling elements are also called "jumping genes" or "transposable elements". Barbara McClintock not only predicted the existence of TEs, but also clearly understood the impact

such elements might have on the genome at large, including changing the expression of nearby genes as well as chromosomal rearrangements (McClintock, 1951). Despite her outstanding experiments, McClintock's findings were largely dismissed by the scientific community for decades and their implications were not fully understood and did not match with existing ideas about genetics (Ravindran, 2012). It was only after the presence of TEs was confirmed by other scientists and in species outside of maize in the 1960s, that McClintock's findings became more generally accepted (Ravindran, 2012). With time, their importance became undeniable, leading to McClintock being awarded a Nobel Prize in Physiology or Medicine in 1983. For decades, transposable elements were still routinely masked and ignored as "junk DNA" in whole-genome studies. Yet, improvements in sequencing techniques and better understanding of the impact of TEs led to a strongly increasing interest in the field of TEs.

TE classification

The classification of TEs is not an easy task. The widely used classification system is based on the division of TEs into classes, orders and superfamilies that are present in all species or widespread in a kingdom (Wicker *et al.*, 2007). Still, the discovery of new TE orders and superfamilies is ongoing. TEs that assumingly are hybrids between known elements and non-autonomous elements complicate classification (Wicker *et al.*, 2007). The currently accepted approach for classifying TEs bases nomenclature on sequence similarity and shared structures and mechanisms (Jurka *et al.*, 2005; Wicker *et al.*, 2007; Piegu *et al.*, 2015; Arkhipova, 2017; Wells & Feschotte, 2020; Storer *et al.*, 2021; Figure 2). Generally, TEs are grouped into two classes, Retrotransposons and DNA transposons (**Figure 2**). Retrotransposons use an RNA intermediate and proliferate via a copy-and-paste mechanism. Consequently, the original copy remains, while a new copy is created and inserted into a new genetic locus. Retrotransposons can further be divided into the orders of LTR (long terminal repeat) that contain long terminal repeats on both sides of the sequence, DIRS (*Dictyostelium* intermediate repeat sequence), Penelope and LINE (long interspersed nuclear element) (Wicker *et al.*, 2007). DNA transposons have a DNA intermediate and move by excision and insertion, or a cut-and-paste mechanism. A first subclass consists of the orders TIR (terminal inverted repeat) and Crypton. TIR DNA transposons encode for a transposase that binds to TIR structures of the element, excise the complete element and insert it in a different location in the genome (for

visualizations of the mechanisms, see the Bachelor project in Knowledge Visualization at ZHdK by Tina Schwendener: https://diplome.kvis.zhdk.ch/Tina-Schwendener). Insertion occurs via double-strand break and leads to a short target site duplication that can remain even after the element is excised (Linheiro & Bergman, 2012). Cryptons do not contain TIRs and code for a tyrosine recombinase. A second subclass of DNA transposons, the Helitrons and Maverick use different mechanisms. Helitrons proliferate by a copy-and-paste mechanism, using a rolling-circle mechanism. Maverick (also called Polinton) are large TEs with similarities to retroviruses that replicate through an unknown mechanism (Pritham *et al.*, 2007).

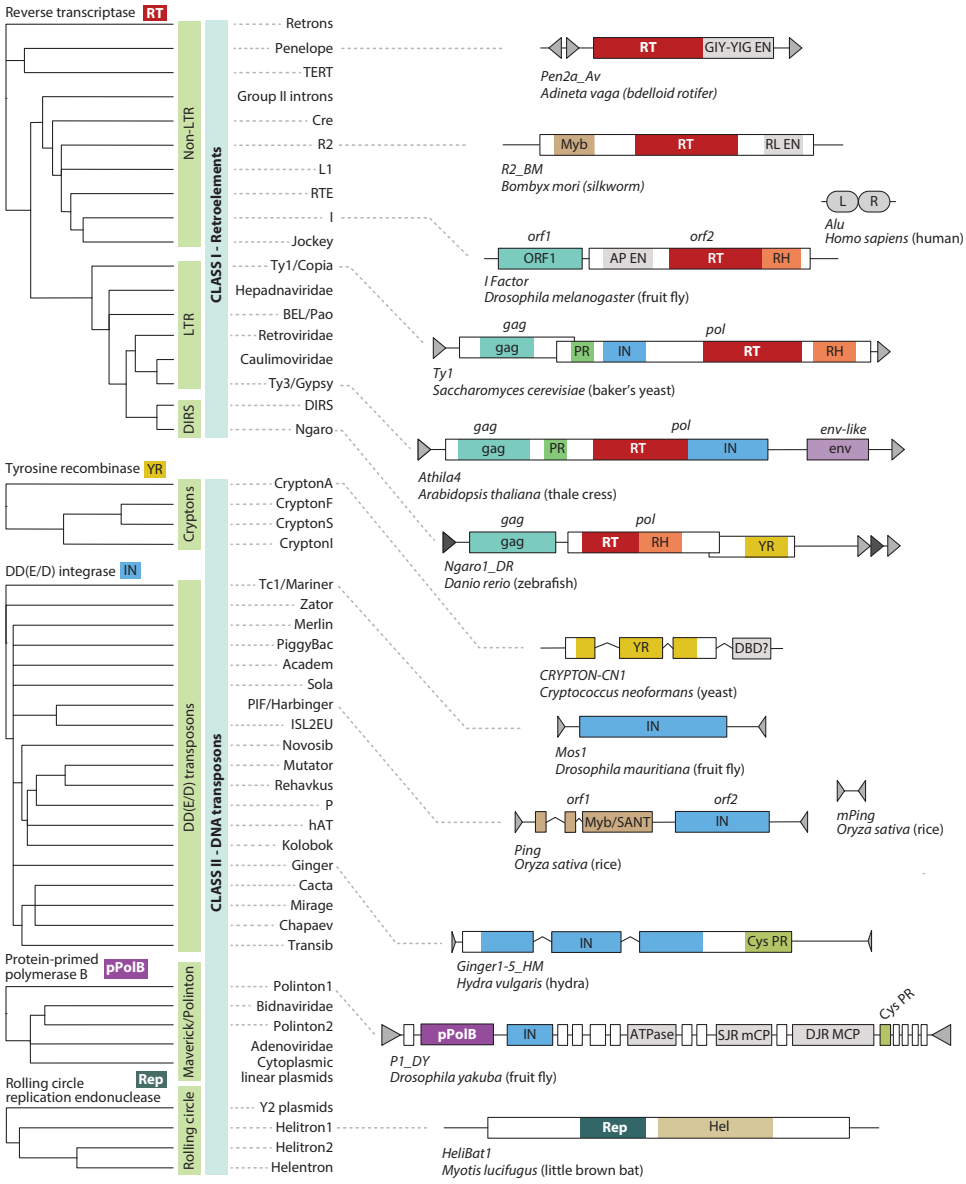


Figure 2: Current TE classification (from: Wells & Feschotte, 2020)

A subset of TEs have lost some or all coding regions, and depend on proteins from autonomous elements. Non-autonomous elements can still be highly active, which indicates that the loss of coding regions might be evolutionary beneficial and can make them parasites of complete

elements (Hua-Van *et al.*, 2005; Lu *et al.*, 2012; Suh, 2019). Non-autonomous elements include MITES (miniature inverted repeat transposable elements; derivates of TIRs) (Feschotte *et al.*, 2002), SINE (short interspersed nuclear element; retrotransposition of Pol III) and TRIM and LARD (terminal repeat retrotransposons in miniature; large retrotransposon derivatives; derivates of Retrotransposons/LTR). Conventional nomenclature systems based on TE class, order and superfamily assign a three-letter code for each TE family (Wicker *et al.*, 2007). Most TE superfamilies are present in all species. Superfamilies are defined by the presence and sequential order of coding regions. Yet, superfamilies still evolve after species separation, leading to distinctive families or subfamilies that are at most shared between very closely related species.

TE curation

TEs diverge and evolve independently in species, but can also be gained by horizontal TE transfer. Consequently, individual TE curation has to be done for each species or clade. Repeatmodeler, a tool based on Repbase, can give a first insight in the number and diversity of TE families (Jurka *et al.*, 2005; Flynn *et al.*, 2020). Repeatmodeler scans the genome for repeated sequences that show high sequence similarity to known elements via Hidden Markov Models and creates consensus sequences for each predicted TE family. The flaking regions are not always precise, and consensus sequences can be based on fragments instead of full-length copies, especially when nested insertions are frequent. For more in-depth analyses, consensus sequences have to be manually curated, which is often time consuming. WickerSoft is a tool based solely on manual curation leading to highly accurate consensus sequences (Breen *et al.*, 2010): For each TE superfamily, the genome is screened for sequence similarity to the coding region of a known element, the results are visually inspected with dot plots, multiple sequence alignments are used to define the correct start and end and are used to create TE family consensus sequences (for more details on the methods see Badet *et al.*, 2020). Several tools have been developed to detect non-autonomous families that are missing the typical coding regions, including LTR-finder, MITE-Tracker and Sine-Scan (for an up-to-date list of all tools visit the TE Hub page: https://tehub.org/en/resources/repeat_tools) (Xu & Wang, 2007; Mao & Wang, 2017; Crescente *et al.*, 2018; Elliott *et al.*, 2021).

Impacts of TEs on the genome

In many species, TEs make up a large part of the genome and TE activity can lead to immense changes in genome size (Wells & Feschotte, 2020). Yet, TEs are not just useless "junk DNA". TEs can disrupt coding and regulatory regions, influence gene regulation, gene evolution and genome integrity.

TE activity, silencing and de-repression

Uncontrolled TE proliferation would have a wide range of negative impacts on the host, thus most TE copies remain epigenetically silenced via histone modification (Slotkin & Martienssen, 2007; He *et al.*, 2019). Silencing is reversible, and under some conditions, TEs can be de-repressed (Miousse *et al.*, 2015). De-repressed TEs are able to create new copies in the genome, which can lead to proliferation bursts (Figure 3). Stress-induced TE de-repression can have negative impacts on host fitness and is associated with many diseases in humans, including some types of cancer, Autism, and Alzheimer's (Guffanti *et al.*, 2014; Goke & Ng, 2016; Horváth *et al.*, 2017; Pavliv *et al.*, 2017; Lapp & Hunter, 2019). Interestingly, controlled TE activity is also associated with a healthy development of the human brain, while TEs are silenced in other tissues (Ahmadi *et al.*, 2020). TEs activity is part of a regulatory network for fine-tuned and cell specific epigenetic expressions of neurons in the hippocampus (Mustafin & Khusnutdinova, 2020). Similar expression profiles of TEs and effector genes during the invasion of the host indicates stress-induced co-regulation in the fungal plant pathogen *Zymoseptoria tritici* (Fouché *et al.*, 2020).

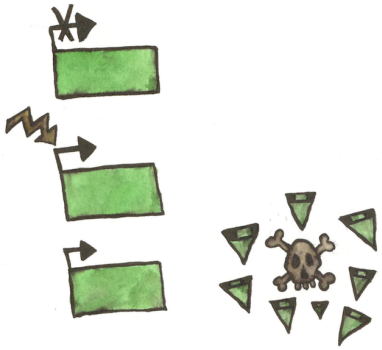


Figure 3: Silenced TEs can become activated during stress conditions and proliferate uncontrolled in the genome

Impact on gene sequences and gene expression

TEs that insert into new locations not only have the potential to be disruptive, but they can also impact the expression level of neighboring genes. Changes to gene expression in turn can impact phenotypic expression. Many examples of TE-induced expression change have recently been described, and the list is growing. Blood oranges and red apples have different types of TEs or TE fragments (*i.e.*, solo-LTRs) in the promoter region of transcriptional activators of anthocyanin pigment production, that are overexpressed in the cold and lead to a red coloration of the pulp or skin (Butelli *et al.*, 2012; Lisch, 2013; Zhang *et al.*, 2019). During the Industrial Revolution, air pollution darkened the barks of birch trees and killed bright lichens, leading to strong exposure of the bright colored peppered moth *Biston betularia* (Cook & Turner, 2008). A darker, better camouflaged phenotype emerged and quickly increased in frequency in the populations. Responsible for the darker phenotype is a TE insertion in the intron of the *cortex* gene, which leads to overexpression in skin tissue and the increased production of dark pigments (van't Hof *et al.*, 2016; Figure 4). With the improvement of air quality, and the change of the environment, a bright phenotype was more preferential again, and the frequency of dark phenotypes subsequently decreased (Cook & Turner, 2008). Through changing gene expression profiles, TEs are important drivers of fast local adaptations and allow species to react quickly to changing environments.

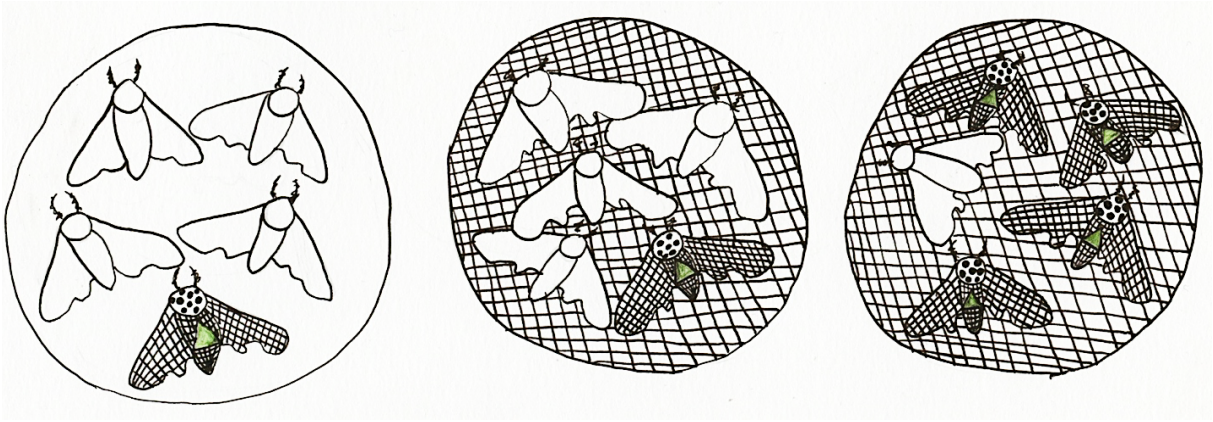


Figure 4: Population dynamics *Biston betularia* in different environments. The green triangle represents the TE insertion.

Impact on chromosomal rearrangements and genome size evolution

TEs also play an important role in large scale rearrangements of the genome. First, TEs will disrupt contiguous stretches of DNA and often increase the distance between genes (Wicker *et al.*, 2018). Generally, uncontrolled TE proliferation will lead to an increase in genome size. Some species, including maize or the fungal species *Cenococcum geophilum*, *Blumeria graminis* and *Peudocercospora ulei* have very large genomes with TE sequences making up

more than 80% of the whole genome (Peter *et al.*, 2016; Müller *et al.*, 2019; González-Sayer *et al.*, 2021; Stitzer *et al.*, 2021). Ectopic recombination either between repetitive elements in the TE (leaving back solo-LTRs) or between two TE copies (leaving back one TE copy while deleting the region between the two elements) can counter-balance genome expansion to some extent (Devos *et al.*, 2002; Petrov *et al.*, 2003; Figure 5).

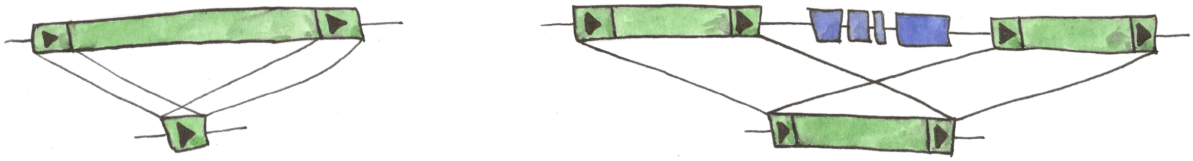


Figure 5: Ectopic recombination between the LTRs of one TE copy, leading to a solo-LTR or between two copies of the same TE family leading to a single copy and the deletion of the region between the copies respectively.

Solo-LTRs can still carry regulatory elements and impact the gene expression, as mentioned in the blood orange example above (Butelli *et al.*, 2012). Ectopic recombination between TEs likely led to the deletion of a larger, TE rich region that contained an already recognized effector gene in the fungal pathogen *Z. tritici*, leading to individuals that were not recognized by the host anymore (Hartmann *et al.*, 2017). TE-derived rearrangements led to a large-scale reorder and change in orientation of genes in the *Orp* regions between sorghum and rice and between rice cultivars (Ma *et al.*, 2005). TEs are also involved in large scale chromosomal rearrangements, creating both breakpoints and anchor points in *Z. tritici* (Mieczkowski *et al.*, 2006; Zhang *et al.*, 2011; Fouché, 2020).

TE co-option

Over time, some TEs became integral parts of the genome, by co-option (also called domestication or exaptation). TEs can capture genes or fuse with gene fragments, creating *de novo* genes, or adding new function by exon shuffling of existing genes (Bennetzen, 2005; Morgante *et al.*, 2005; Fouché *et al.*, 2018; Figure 6).

TEs can either provide genes or regulatory regions, for example promotors or enhancers in the LTR sequences (Rebollo *et al.*, 2012). Most TEs will lose their ability of self-replication and mobility, will become a part of the gene and evolve in a gene like fashion (Hoen & Bureau, 2015; Chuong *et al.*, 2017). Host-TE fusions can still be present in the genome, even when the original TE family is extinct, and can remain in species, clades or even kingdoms over long evolutionary time-scales when they provide a benefit (Jangam *et al.*, 2017; Cosby *et al.*, 2019). Copies of an old family of SINE-like TEs are acting as conserved and tissue-specific neocortex enhancers in tetrapods (Notwell *et al.*, 2015). Fusions with TE-derived transposases play an

important part in transcriptional regulation in vertebrates (Cosby *et al.*, 2021). Co-opted TEs induced many new functions and play an important part in the evolution of proteomes.

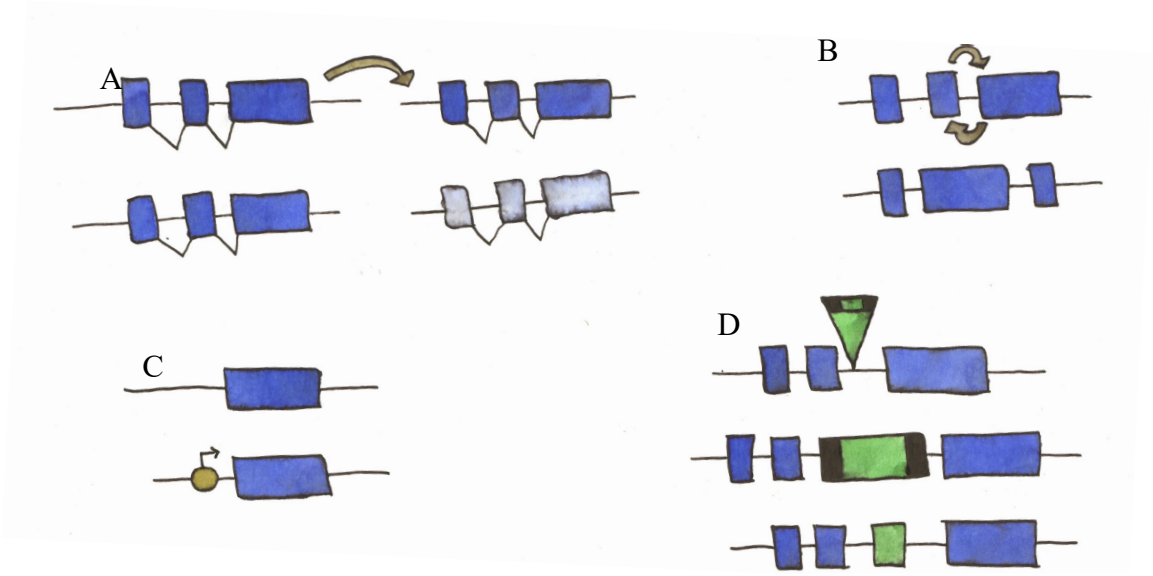


Figure 6: Gene evolution. A) Duplication of a gene, with followed divergence of one copy. B) Exon shuffling. C) insertion of a transcription factor binding site in front of a non-coding region. D) Insertion, partial deletion and co-option of a TE into a gene

Population dynamics of TEs

TE distribution in genomes is a product of TE activity, selection, drift and defense mechanisms against TEs. TEs undergo birth (duplications and transpositions), death (excision and deletion) and resurrection (de-repression) (Blumenstiel, 2019). The birth-death rates differ between TE families and host species (Petrov *et al.*, 2011; Drakos & Wahl, 2015). An increase in copy numbers can be counterbalanced by defense mechanisms and purifying selection (Charlesworth & Charlesworth, 1983). Comparisons of the TE composition showed a high variability of TE composition between individuals of *Drosophila melanogaster* or *Z. tritici* (Barrón *et al.*, 2014; Plissonneau *et al.*, 2016; Badet *et al.*, 2020). Especially in asexual species, TE activity is an important driver of recombination in the absence of meiotic recombination (Schmidt *et al.*, 2013). Petrov and colleagues (2011) suspect strongly differing transposition rates between TE families. Allele frequencies give an indication of the impact of a TE. Most TE insertions are rare and remain at low frequencies in the population (Petrov *et al.*, 2011). Low frequency indicates a slightly negative or neutral impact (Petrov *et al.*, 2011). Beneficial insertions are selected for, and will increase in frequency in the population at a fast rate (Barrón *et al.*, 2014). Yet, a high allele frequency is not a direct indication for positive selection, as

genetic drift can randomly lead to the increase of neutral and slightly deleterious insertions. Additionally, purifying selection seems to have a higher impact on longer TEs and on TEs located in regions with higher recombination rates (Petrov *et al.*, 2011). TEs can also be lost by deletion through ectopic recombination. Elaborate defense mechanisms prevent the spread of TEs, either by inducing a higher mutation rate, or by epigenetically silencing TEs (Slotkin & Martienssen, 2007). Some filamentous ascomycetes contain a control mechanism called repeat induced point mutations (RIP) that limits the activity of TEs (Clutterbuck, 2011). During the sexual cycle (diploid phase), duplications are detected and CpG -> TpA mutations are introduced (Galagan & Selker, 2004), leading to missense or nonsense sequences. RIP-marked mutations often are targets for methylation and thus silencing (Galagan & Selker, 2004).

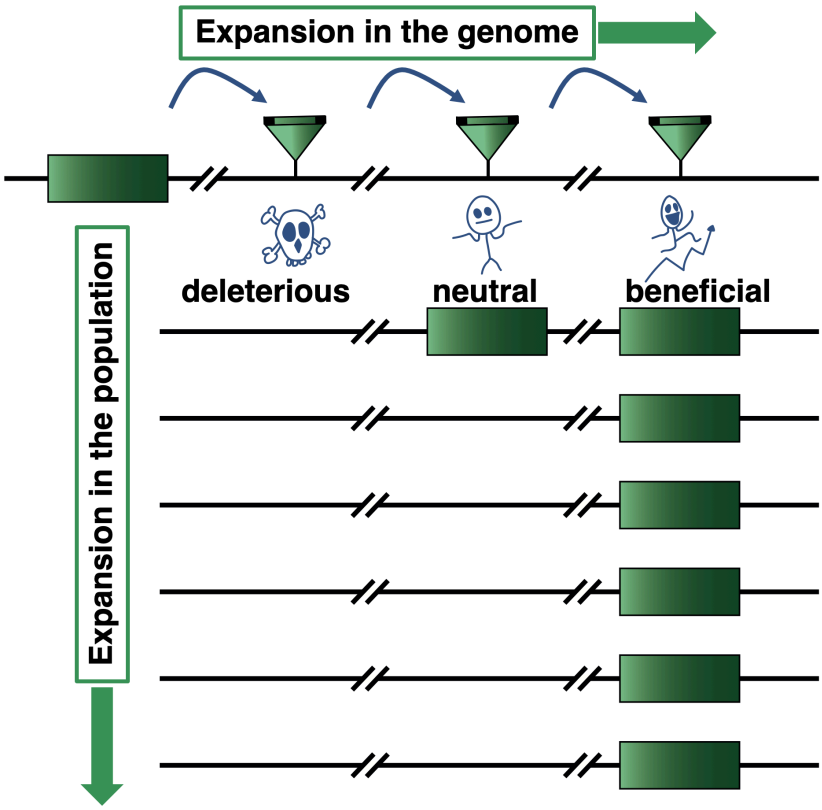


Figure 7: Allele frequency changes in populations of deleterious, neutral and beneficial TE insertions.

TEs in fungi

Fungi are a very diverse group of organisms, that range from small single celled organisms to the largest known living species on earth (Sipos *et al.*, 2017). Fungi provide many important functions, as mutualists with land plants and quasi-domesticated strains for food and pharmaceutical production (Rokas, 2009; Peter *et al.*, 2016). On the other hand, pathogenic fungi pose an important threat to human health, especially for immunosuppressed patients. A gain of resistance against antifungal drugs and the slow development of new treatments remains an important challenge (Martins *et al.*, 2014; Casadevall, 2018). Likewise, fungal plant pathogens are threatening important local plant species and degrease food production dramatically (McKinney *et al.*, 2014; Stukenbrock & Croll, 2014). A recent study estimated the fungal kingdom to consist of at least 12 million species (Wu *et al.*, 2019). High quality reference genomes exist for ~1,300 fungal species (Figure 8). New sequencing methods including high-throughput amplicon sequencing, shotgun metagenomics and single-cell genomics will speed up genomic and TE research in the future (Wu *et al.*, 2019). TEs have been shown to lead to multidrug fungicide resistance and other adaptations (Omrane *et al.*, 2015, 2017; Gusa *et al.*, 2020; Mäe *et al.*, 2020). The best-studied model organisms of the fungal kingdom, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Candida albicans* and *Cryptococcus neoformans*, show some of the lowest known TE contents (3.5%, 1.1%, 0.8%, 6.6% of the whole genome respectively) (Maxwell, 2020). However, a low TE content is not a typical fungal characteristic, as many fungal genomes are made of more than 80 % TEs (Peter *et al.*, 2016; Müller *et al.*, 2019; González-Sayer *et al.*, 2021). TEs are significant drivers of diversification and adaptation in many fungi, including the model organisms (Raffaele & Kamoun, 2012; Gusa *et al.*, 2020). We recently hypothesized three different modes of TE proliferation in fungal genomes to explain why the TE content can vary dramatically even between closely related species (Fouché *et al.*, 2021): A) First, genomes can be “resistant” to uncontrolled burst of TEs, keeping the TE content constantly low, indicating strong purifying selection or defense mechanisms against TEs. An example of a “resistant” species is *C. albicans* with a low TE content but indication of ongoing TE activity (Holton *et al.*, 2001).

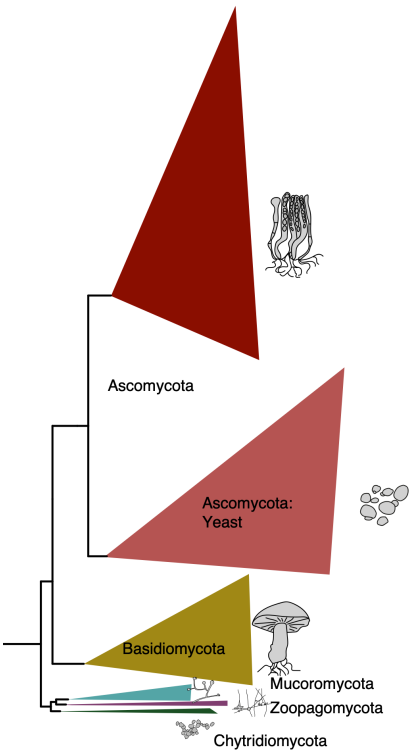


Figure 8: Simplified phylogenetic tree of the fungal kingdom.

B) "Semi-tolerance", where the genome is separated into TE-rich/gene-poor and gene-rich/TE-poor compartments (Figure 9). Strong purifying selection will prevent most TE insertions in the gene-rich compartments to remain in the population, but TE insertions into TE-rich compartments will be tolerated, creating nested insertions (Rouxel & Balesdent, 2017). Fast-evolving genes, e.g. effectors involved in the interaction with the host, are often located in TE-rich compartments, which led to the formulation of the multiple-compartment or two-speed-genome hypothesis (Dong *et al.*, 2015; Torres *et al.*, 2020). C) Finally, some fungal species with high TE activity and strong bursts of proliferation might have become "tolerant" to new TE insertions. In the TE-bloated genomes *C. geophilum* and *P. ulei* with over 80% TE content, the likelihood of a TE to insert into a coding or regulatory region is very low, as these genomes are virtually made of TEs with some sparse genes (Peter *et al.*, 2016; González-Sayer *et al.*, 2021).

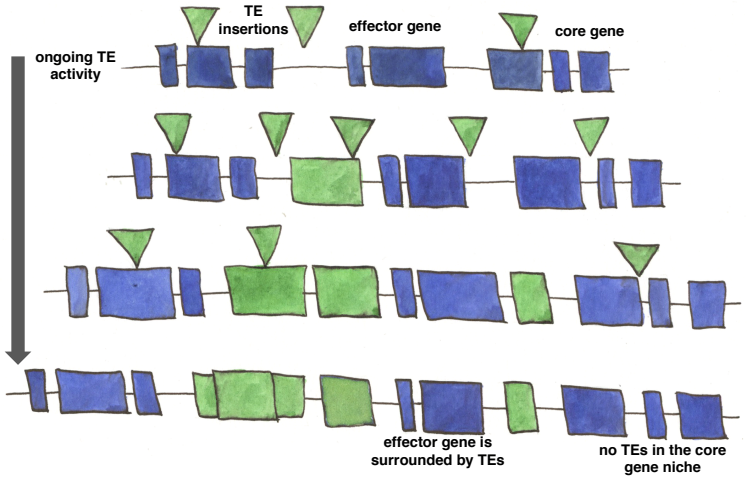


Figure 9: Compartmentalization. TE insertions in regions with core genes are less likely to remain in the population, while TE insertions close to effector genes can remain.

Zymoseptoria tritici *as an important pathogen of wheat*

Zymoseptoria tritici is one of the most important plant pathogens on *Triticum aestivum* (wheat) and a threat to food security (Jorgensen *et al.*, 2014). *Z. tritici* split an estimated 11,000 years ago from its closest sister species *Z. pseudotritici* and co-evolved during the domestication of wheat (Stukenbrock *et al.*, 2007; Stukenbrock & McDonald, 2008). The infection cycle of the fungus starts with hyphal grow through the stomata into the plant intercellular space, without penetration of the plant cell (Figure 10). After a longer latency phase, sexual or asexual fruiting bodies are formed (Kema *et al.*, 1996; Steinberg, 2015). Clonally produced spores are distributed by water splash over short distances and stay on the same plant or the neighboring plants (Singh *et al.*, 2020; Karisto *et al.*, 2021). Sexual spore production is triggered under changing environmental conditions and in the presence of a high density of compatible mates, mostly towards the end of the growing season. Sexually produced spores are dispersed by wind over larger distances (Steinberg, 2015) and are able to infect new fields. *Z. tritici* strains shows a great variability in virulence towards different wheat cultivars. Hartmann and colleagues (2017) showed a complex cause of virulence in *Z. tritici* with at last 25 distinct genomic regions involved, including genes coding for plant cell wall degradation, cell transport and fungal metabolism. *Z. tritici* can be considered as a model organism to study the co-evolution between filamentous plant pathogens and their hosts. *Z. tritici* has a highly plastic genome and a moderate but dynamic TE content of 16.5-24 % (Badet *et al.*, 2020).

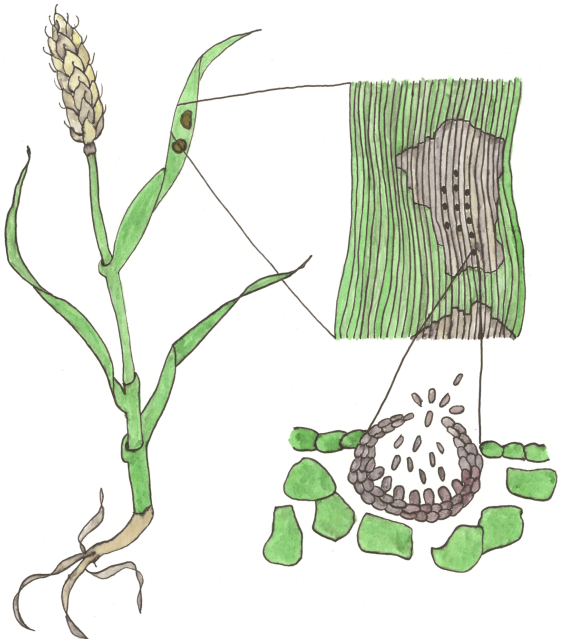


Figure 10: Zymoseptoria tritici disease. Lesions and pycnidia (asexual fruit bodies) on wheat leaves.

The pathogen has adaptive phenotypic traits (*e.g.,* virulence), very successful infection and reproduction strategies, and is easy to grow *in planta* and *in vitro*. The *Z. tritici* genome is composed of 13 core and up to 8 accessory chromosomes (Goodwin *et al.*, 2011). In comparison to other known filamentous plant pathogens, accessory chromosomes do not show significantly higher amounts of virulence or effector genes than core chromosomes (Raffaele & Kamoun, 2012). TEs can be located anywhere in the genome, but are significantly enriched in heterochromatic, subtelomeric and gene poor regions, and are often closer to genes are not shared by all individuals (Raffaele & Kamoun, 2012; Dong *et al.*, 2015; Grandaubert *et al.*, 2015). TEs are especially closer to genes that encode proteins involved in host infection processes than to genes coding for more conserved metabolic functions (Plissonneau *et al.*, 2016; Hartmann & Croll, 2017). Grandaubert and colleagues (2015) found species-specific TE compositions in *Z. tritici* compared to its sister species, with a reduced TE diversity in *Z. tritici* (Hartmann & Croll, 2017). TEs have been shown to impact virulence and resistance to fungicides (Omrane *et al.*, 2015, 2017; Hartmann *et al.*, 2017; Wang *et al.*, 2021).

Objectives

Transposable elements (TEs) are able to replicate within a genome and are major drivers of genome instability and epigenetic changes (Eichler & Sankoff, 2003; Slotkin & Martienssen,

2007). The disruptive nature of TEs is not necessarily deleterious in each insertion. TEs can influence the expression of nearby genes, leading to local adaptation (Lisch, 2009). In fungal plant pathogens, TEs can delete or mutate recognized effector genes, leading to resistance to the host plant (Hartmann *et al.*, 2017; Torres *et al.*, 2020). In asexual species, TE-derived chromosomal rearrangements are an important source of diversity (de Jonge *et al.*, 2013). TEs play an important role in the evolution and adaptation of fungal species. However, most work on TEs in fungal species is focused on individual insertions, and a global view on TE dynamics is still missing. Understanding how TEs influence the evolution of a species can explain adaptation events, where the sole presence of a gene or a single nucleotide polymorphism cannot fully explain a changed phenotype. In this thesis, I focus on the impact of TEs on long- and short-term adaptation of fungal species. I highlight the importance of studying TE dynamics and evolution in terms of TE activity, defense mechanisms against TEs, selection and genetic drift.

We show how TEs impact the evolution of species, how TEs evolve, and how TEs can be co-opted to become essential parts of the genome.

In the **first chapter**, we aimed to describe the TE diversity and to understand how TE activity can lead to local adaptation despite the negative effect of genome size expansions. We scanned genome-wide population data in the fungal plant pathogen *Z. tritici* for TE presence and absence. This was the first population dynamics study of TEs in a fungal species to our knowledge.

In the **second chapter**, we focused more on the evolution and mobility of TEs themselves. We aimed to understand the expansion routes of TEs: in which genomic niches can TEs become active, where will new TEs insert and what is the impact of selection and defense mechanisms? We described genomic niches in *Z. tritici* and scanned for signals of defense against TEs both in the genomic niche and in the TE itself. We used phylogenetic tools to define the expansion routes of TE families.

In the **third chapter**, we were focusing on co-opted TEs, that provided a new function to already existing genes in the fungal kingdom and beyond. We wanted to know which TE and gene functions are more likely involved in host-TE fusions. Additionally, we wanted to know how widespread and old such fusion candidates are in the fungal kingdom, indicating a positive function or adaptation.

Ahmadi Amirhossein, De Toma Ilario, Vilor-Tejedor Natàlia, Eftekhariyan Ghamsari Mohammad Reza, Sadeghi Iman. 2020. Transposable elements in brain health and disease. Ageing Research Reviews 64.

Arkhipova Irina R. 2017. Using bioinformatic and phylogenetic approaches to classify transposable elements and understand their complex evolutionary histories. Mobile DNA 8: 1–14.

Badet Thomas, Oggenfuss Ursula, Abraham Leen, McDonald Bruce A, Croll Daniel. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biology 18: 12.

Barrón Maite G, Fiston-Lavier Anna-Sophie, Petrov Dmitri A, González J. 2014. Population Genomics of Transposable Elements in Drosophila (BL Bassler, Ed.). Annual Review of Genetics 48: 561–581.

Bennetzen Jeffrey L. 2005. Transposable elements, gene creation and genome rearrangement in flowering plants. Current Opinion in Genetics and Development 15: 621–627.

Blumenstiel Justin P. 2019. Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation. Genes 10.

Breen James, Li Dora, Dunn David S, Békés Ferenc, Kong Xiuying, Zhang Juncheng, Jia Jizeng, Wicker Thomas, Mago Rohit, Ma Wujun, Bellgard Matthew, Appels Rudi. 2010. Wheat beta-expansin (EXPB11) genes: Identification of the expressed gene on chromosome 3BS carrying a pollen allergen domain. BMC Plant Biology 10.

Butelli Eugenio, Licciardello Concetta, Zhang Yang, Liu Jianjun, Mackay Steve, Bailey Paul, Reforgiato-Recupero Giuseppe, Martin Cathie. 2012. Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. Plant Cell 24: 1242–1255.

Casadevall Arturo. 2018. Fungal diseases in the 21st century: The near and far horizons. Pathogens and Immunity 3: 183–196.

Charlesworth Brian, Charlesworth Deborah. 1983. The population dynamics of transposable elements. Genetical Research 42: 1–27.

Chuong EB, Elde NC, Feschotte Cédric. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics 18: 71–86.

Clutterbuck AJ. 2011. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. Fungal Genet Biol 48: 306–326.

Cook LM, Turner JRG. 2008. Decline of melanism in two British moths: Spatial, temporal and inter-specific variation. Heredity 101: 483–489.

Cosby Rachel L, Chang Ni Chen, Feschotte Cédric. 2019. Host–transposon interactions: Conflict, cooperation, and cooption. Genes and Development 33: 1098–1116.

Cosby Rachel L, Judd Julius, Zhang Ruiling, Zhong Alan, Garry Nathaniel, Pritham Ellen J, Feschotte Cédric. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. Science 371: eabc6405.

Crescente Juan Manuel, Zavallo Diego, Helguera Marcelo, Vanzetti Leonardo Sebastián. 2018. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics 19: 1–10.

de Jonge R, Bolton MD, Kombrink A, van den Berg GCM, Yadeta KA, Thomma Bphj. 2013. Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. Genome Research 23: 1271–1282.

Devos Katrien M, Brown James KM, Bennetzen Jeffrey L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Research 12: 1075–1079.

Dong S, Raffaele S, Kamoun S. 2015. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev 35: 57–65.

Doolittle WF, Sapienza C. 1980. Selfish genes, the phenotype paradigm and genome evolution. Nature 284.

Drakos NE, Wahl LM. 2015. Extinction probabilities and stationary distributions of mobile genetic elements in prokaryotes: The birth-death-diversification model. Theoretical Population Biology 106: 22–31.

Eichler Evan E, Sankoff David. 2003. Structural dynamics of eukaryotic chromosome evolution. Science 301: 793–797.

Elliott Tyler A, Heitkam Tony, Hubley Robert, Quesneville Hadi, Suh Alexander, Wheeler Travis J. 2021. TE Hub: A community-oriented space for sharing and connecting tools, data, resources, and methods for transposable element annotation. Mobile DNA 12: 1–5.

Fedoroff N, Wessler S, Shure M. 1983. Isolation of the transposable maize controlling elements Ac and Ds. Cell 35: 235–242.

Feschotte Cédric, Jiang Ning, Wessler Susan R. 2002. Plant transposable elements: Where genetics meets genomics. Nature Reviews Genetics 3: 329–341.

Flynn Jullien M, Hubley Robert, Goubert Clément, Rosen Jeb, Clark Andrew G, Feschotte Cédric, Smit Arian F. 2020. RepeatModeler2 for automated genomic discovery of transposable element families. Proceedings of the National Academy of Sciences of the United States of America 117: 9451–9457.

Fouché Simone. 2020. Drivers of genome evolution in a fungal pathogen of wheat.

Fouché Simone, Badet Thomas, Oggenfuss Ursula, Plissonneau Clémence, Francisco Carolina Sardinha, Croll Daniel. 2020. Stress-Driven Transposable Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. Molecular biology and evolution 37: 221–239.

Fouché Simone, Oggenfuss Ursula, Chanclud Emilie, Croll Daniel. 2021. A devil's bargain with transposable elements in plant pathogens. Trends in Genetics: 1–9.

Fouché Simone, Plissonneau Clémence, Croll Daniel. 2018. The birth and death of effectors in rapidly evolving filamentous pathogen genomes. Current Opinion in Microbiology 46: 34–42.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends in Genetics 20: 417–423.

Goke J, Ng HH. 2016. CTRL plus INSERT: retrotransposons and their contribution to regulation and innovation of the transcriptome. Embo Reports 17: 1131–1144.

González-Sayer Sandra, Oggenfuss Ursula, García Ibonne, Aristizabal Fabio. 2021. High-quality genome assembly of Pseudocercospora ulei the main threat to natural rubber trees. Genetics and Molecular Biology: 0–1.

Goodwin Stephen B, Ben M'Barek Sarrah, Dhillon Braham, Wittenberg Alexander HJ, Crane Charles F, Hane James K, Foster Andrew J, Van der Lee Theo AJ, Grimwood Jane, Aerts Andrea, Antoniw John, Bailey Andy, Bluhm Burt, Bowler Judith, Bristow Jim, van der Burgt Ate, Canto-Canché Blondy, Churchill Alice CL, Conde-Ferràez Laura, et al. 2011. Finished Genome of the Fungal Wheat Pathogen Mycosphaerella graminicola Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis (Harmit S Malik, Ed.). PLOS Genetics 7: e1002070.

Grandaubert J, Bhattacharyya A, Stukenbrock EH. 2015. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3-Genes Genomes Genetics 5: 1323–1333.

Guffanti Guia, Gaudi Simona, Fallon James H, Sobell Janet, Potkin Steven G, Pato Carlos, Macciardi Fabio. 2014. Transposable elements and psychiatric disorders. American Journal of Medical Genetics, Part B: Neuropsychiatric Genetics 165: 201–216.

Gusa Asiya, Williams Jonathan D, Cho Jang Eun, Averette Anna Floyd, Sun Sheng, Shouse Eva Mei, Heitman Joseph, Alspaugh J Andrew, Jinks-Robertson Sue. 2020. Transposon mobilization in the human fungal pathogen Cryptococcus is mutagenic during infection and promotes drug resistance in vitro. Proceedings of the National Academy of Sciences of the United States of America 117: 9973–9980.

Hartmann FE, Croll D. 2017. Distinct Trajectories of Massive Recent Gene Gains and Losses in Populations of a Microbial Eukaryotic Pathogen. Molecular Biology and Evolution.

Hartmann Fanny E, Sánchez-Vallet Andrea, McDonald Bruce A, Croll Daniel. 2017. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. The ISME Journal 11: 1189–1204.

He Jiangping, Fu Xiuling, Zhang Meng, He Fangfang, Li Wenjuan, Abdul Mazid, Zhou Jianguo, Sun Li, Chang Chen, Li Yuhao, Liu He, Wu Kaixin, Babarinde Isaac A, Zhuang Qiang, Loh Yuin-han, Chen Jiekai, Esteban Miguel A, Hutchins Andrew P. 2019. Transposable elements are regulated by context- specific patterns of chromatin marks in mouse embryonic stem cells.

Hoen Douglas R, Bureau Thomas E. 2015. Discovery of novel genes derived from transposable elements using integrative genomic analysis. Molecular Biology and Evolution 32: 1487–1506.

Holton Nicholas J, Goodwin Timothy JD, Butler Margaret I, Poulter RTM. 2001. An active retrotransposon in Candida albicans. Nucleic Acids Research 29: 4014–4024.

Horváth Vivien, Merenciano Miriam, González J. 2017. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends in Genetics 33: 832–841.

Hua-Van A, Le Rouzic A, Maisonhaute C, Capy P. 2005. Abundance, distribution and dynamics of retrotransposable elements and transposons: Similarities and differences. Cytogenetic and Genome Research 110: 426–440.

Jangam Diwash, Feschotte Cédric, Betrán Esther. 2017. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends in Genetics 33: 817–831.

Jorgensen LN, Hovmoller MS, Hansen JG, Lassen P, Clark B, Bayles R, Rodemann B, Flath K, Jahn M, Goral T, Czembor J, Cheyron P, Maumene C, De Pope C, Ban R, Nielsen GC, Berg G. 2014. IPM Strategies and Their Dilemmas Including an Introduction to www.eurowheat.org. Journal of Integrative Agriculture 13: 265–281.

Jurka J, Kapitonov V V., Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. 2005. Repbase Update, a database of eukaryotic repetitive elements. Cytogenetic and Genome Research 110: 462–467.

Karisto Petteri, Suffert Frédéric, Mikaberidze Alexey. 2021. Measuring splash-dispersal of a major wheat pathogen in the field. PhytoFrontiersTM.

Kass Lee B. 2003. Records and recollections: A new look at Barbara McClintock, Nobel-Prize-winning geneticist. Genetics 164: 1251–1260.

Kema GHJ, Yu DZ, Rijkenberg FHJ, Shaw MW, Baayen RP. 1996. Histology of the pathogenesis of Mycosphaerella graminicola in wheat. Phytopathology 86: 777–786.

Lapp Hannah E, Hunter Richard G. 2019. Early life exposures, neurodevelopmental disorders, and transposable elements. Neurobiology of Stress 11: 100174.

Linheiro Raquel S, Bergman Casey M. 2012. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster (Jason E Stajich, Ed.). PLOS ONE 7: e30008.

Lisch Damon. 2009. Epigenetic regulation of transposable elements in plants. Annual Review of Plant Biology 60: 43–66.

Lisch D. 2013. How important are transposons for plant evolution? Nature Review Genetics 14: 49–61.

Lu Chen, Chen Jiongjiong, Zhang Yu, Hu Qun, Su Wenqing, Kuang Hanhui. 2012. Miniature inverted-repeat transposable elements (MITEs) have been accumulated through amplification bursts and play important roles in gene expression and species diversity in oryza sativa. Molecular Biology and Evolution 29: 1005–1017.

Ma Jianxin, Sanmiguel Phillip, Lai Jinsheng, Messing Joachim, Bennetzen Jeffrey L. 2005. DNA rearrangement in orthologous Orp regions of the maize, rice and sorghum genomes. Genetics 170: 1209–1220.

Mäe Andres, Fillinger Sabine, Sooväli Pille, Heick Thies Marten. 2020. Fungicide Sensitivity Shifting of Zymoseptoria tritici in the Finnish-Baltic Region and a Novel Insertion in the MFS1 Promoter. Frontiers in Plant Science 11: 1–10.

Mao Hongliang, Wang Hao. 2017. SINE-scan: An efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics 33: 743–745.

Martins Natália, Ferreira Isabel CFR, Barros Lillian, Silva Sónia, Henriques Mariana. 2014. Candidiasis: Predisposing Factors, Prevention, Diagnosis and Alternative Treatment. Mycopathologia 177: 223–240.

Maxwell Patrick H. 2020. Diverse transposable element landscapes in pathogenic and nonpathogenic yeast models: The value of a comparative perspective. Mobile DNA 11: 1–26.

McClintock Barbara. 1951. Chromosome organization and genic expression. Cold Spring Harb Symp Quant Biol 1: 13–47.

McClintock Barbara. 1953. Induction of Instability at Selected Loci in Maize. Genetics 38: 579–99.

McKinney L V, Nielsen LR, Collinge DB, Thomsen IM, Hansen JK, Kjaer ED. 2014. The ash dieback crisis: genetic variation in resistance can prove a long-term solution. Plant Pathology 63: 485–499.

Mieczkowski Piotr A, Lemoine Francene J, Petes Thomas D. 2006. Recombination between retrotransposons as a source of chromosome rearrangements in the yeast Saccharomyces cerevisiae. DNA Repair 5: 1010–1020.

Miousse Isabelle R, Chalbot Marie Cecile G, Lumen Annie, Ferguson Alesia, Kavouras Ilias G, Koturbash Igor. 2015. Response of transposable elements to environmental stressors. Mutation Research - Reviews in Mutation Research 765: 19–39.

Morgante Michele, Brunner Stephan, Pea Giorgio, Fengler Kevin, Zuccolo Andrea, Rafalski Antoni. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nature Genetics 37: 997–1002.

Müller Marion C, Praz Coraline R, Sotiropoulos Alexandros G, Menardo Fabrizio, Kunz Lukas, Schudel Seraina, Oberhänsli Simone, Poretti Manuel, Wehrli Andreas, Bourras Salim, Keller Beat, Wicker Thomas. 2019. A chromosome-scale genome assembly reveals a highly dynamic effector repertoire of wheat powdery mildew. New Phytologist 221: 2176–2189.

Mustafin RN, Khusnutdinova EK. 2020. Involvement of transposable elements in neurogenesis. Vavilovskii Zhurnal Genetiki i Selektsii 24: 209–218.

Notwell James H, Chung Tisha, Heavner Whitney, Bejerano Gill. 2015. A family of transposable elements co-opted into developmental enhancers in the mouse neocortex. Nature Communications 6: 1–7.

Omrane Selim, Audéon Colette, Ignace Amandine, Duplaix Clémentine, Aouini Lamia, Kema Gert, Walker Anne-Sophie, Fillinger Sabine. 2017. Plasticity of the MFS1 promoter leads to multi drug resistance in the wheat pathogen Zymoseptoria tritici. mSphere: 1–42.

Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, Fillinger S. 2015. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. Environmental Microbiology 17: 2805–2823.

Pavliv Oleksandra, Shpyleva Svitlana, Pogribny Igor, Melnyk Stepan, Jill James S. 2017. Overexpression of LINE-1 Retrotransposons in Autism Brain. Molecular Neurobiology 55: 1740–1749.

Peter Martina, Kohler Annegret, Ohm Robin A, Kuo Alan, Krützmann Jennifer, Morin Emmanuelle, Arend Matthias, Barry Kerrie W, Binder Manfred, Choi Cindy, Clum Alicia, Copeland Alex, Grisel Nadine, Haridas Sajeet, Kipfer Tabea, LaButti Kurt, Lindquist Erika, Lipzen Anna, Maire Renaud, et al. 2016. Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus Cenococcum geophilum. Nature Communications 7: 1–15.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. Molecular Biology and Evolution 20: 880–892.

Petrov DA, Fiston-Lavier AS, Lipatov M, Lenkov K, González J. 2011. Population Genomics of Transposable Elements in Drosophila melanogaster. Molecular Biology and Evolution 28: 1633–1644.

Piegu B, Bire S, Arensburger P, Bigot Y. 2015. A survey of transposable element classification systems - A call for a fundamental update to meet the challenge of their diversity and complexity. Molecular Phylogenetics and Evolution 86: 90–109.

Plissonneau Clémence, Stürchler Alessandra, Croll Daniel. 2016. The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat. mBio 7: 1–13.

Pritham Ellen J, Putliwala Tasneem, Feschotte Cédric. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390: 3–17.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology 10: 417–430.

Ravindran Sandeep. 2012. Barbara McClintock and the discovery of jumping genes. Proceedings of the National Academy of Sciences of the United States of America 109: 20198–20199.

Rebollo Rita, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. In: Bassler BL, ed. Annual Review of Genetics, Vol 46. 21–42.

Rokas Antonis. 2009. The effect of domestication on the fungal proteome. Trends in Genetics 25: 60–63.

Rouxel Thierry, Balesdent Marie Hélène. 2017. Life, death and rebirth of avirulence effectors in a fungal pathogen of Brassica crops, Leptosphaeria maculans. New Phytologist 214: 526–532.

Schmidt Sarah M, Houterman Petra M, Schreiver Ines, Ma Lisong, Amyotte Stefan, Chellappan Biju, Boeren Sjef, Takken Frank LW, Rep Martijn. 2013. MITEs in the promoters of effector genes allow prediction of novel virulence genes in Fusarium oxysporum. BMC genomics 14.

Singh Nikhil Kumar, Chanclud Emilie, Croll Daniel. 2020. Population-level deep sequencing reveals the interplay of clonal and sexual reproduction in the fungal wheat pathogen Zymoseptoria tritici.

Sipos György, Prasanna Arun N, Walter Mathias C, O'Connor Eoin, Bálint Balázs, Krizsán Krisztina, Kiss Brigitta, Hess Jaqueline, Varga Torda, Slot Jason, Riley Robert, Bóka Bettina, Rigling Daniel, Barry Kerrie, Lee Juna, Mihaltcheva Sirma, Labutti Kurt, Lipzen Anna, Waldron Rose, et al. 2017. Genome expansion and lineage-specific genetic innovations in the forest pathogenic fungi Armillaria. Nature Ecology and Evolution 1: 1931–1941.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272–285.

Steinberg Gero. 2015. Cell biology of Zymoseptoria tritici: Pathogen cell organization and wheat infection. Fungal Genetics and Biology 79: 17–23.

Stitzer Michelle C, Anderson Sarah N, Springer Nathan M, Ross-Ibarra Jeffrey. 2021. The genomic ecosystem of transposable elements in maize. PLOS Genetics 17: e1009768.

Storer Jessica, Hubley Robert, Rosen Jeb, Wheeler Travis J, Smit Arian F. 2021. The Dfam community resource of transposable element families, sequence models, and genome annotations. Mobile DNA 12: 1–14.

Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. 2007. Origin and domestication of the fungal wheat pathogen Mycosphaerella graminicola via sympatric speciation. Molecular Biology and Evolution 24: 398–411.

Stukenbrock EH, Croll Daniel. 2014. The evolving fungal genome. Fungal Biology Reviews 28: 1–12.

Stukenbrock EH, McDonald BA. 2008. The origins of plant pathogens in agro-ecosystems. In: Annual Review of Phytopathology. 75–100.

Suh Alexander. 2019. Genome Size Evolution: Small Transposons with Large Consequences. Current Biology 29: R241–R243.

Torres David E, Oggenfuss Ursula, Croll Daniel, Seidl Michael F. 2020. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biology Reviews 34: 136–143.

van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534: 102–105.

Wang Chen, Milgate Andrew W, Solomon Peter S, McDonald Megan C. 2021. The identification of a transposon affecting the asexual reproduction of the wheat pathogen Zymoseptoria tritici. Molecular Plant Pathology 22: 800–816.

Wells Jonathan N, Feschotte Cédric. 2020. A Field Guide to Transposable Elements. Annual Review of Genetics 54: 7–34.

Wicker Thomas, Gundlach Heidrun, Spannagl Manuel, Uauy Cristobal, Borrill Philippa, Ramírez-González Ricardo H, De Oliveira Romain, Mayer Klaus FX, Paux Etienne, Choulet Frédéric. 2018. Impact of transposable elements on genome structure and evolution in bread wheat. Genome Biology 19: 1–18.

Wicker Thomas, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982.

Wu Bing, Hussain Muzammil, Zhang Weiwei, Stadler Marc, Liu Xingzhong, Xiang Meichun. 2019. Current insights into fungal species diversity and perspective on naming the environmental DNA sequences of fungi. Mycology 10: 127–140.

Xu Zhao, Wang Hao. 2007. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35: 265–268.

Zhang Liyi, Hu Jiang, Han Xiaolei, Li Jingjing, Gao Yuan, Richards Christopher M, Zhang Caixia, Tian Yi, Liu Guiming, Gul Hera, Wang Dajiang, Tian Yu, Yang Chuanxin, Meng Minghui, Yuan Gaopeng, Kang Guodong, Wu Yonglong, Wang Kun, Zhang Hengtao, et al. 2019. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nature Communications 10: 1494.

Zhang Jianbo, Yu Chuanhe, Krishnaswamy Lakshminarasimhan, Peterson Thomas. 2011. Transposable elements as catalysts for chromosome rearrangements. Methods in molecular biology (Clifton, N.J.) 701: 315–326.

Chapter 1: A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen

Ursula Oggenfuss[1], Thomas Badet[1], Thomas Wicker[2], Fanny E. Hartmann[3,4], Nikhil K. Singh[1], Leen N. Abraham[1], Petteri Karisto[4,6], Tiziana Vonlanthen[4], Christopher C. Mundt[5], Bruce A. McDonald[4], Daniel Croll[1,*]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland
[2] Institute for Plant and Microbial Biology, University of Zurich, Zurich, Switzerland
[3] Ecologie Systématique Evolution, Bâtiment 360, Univ. Paris-Sud, AgroParisTech, CNRS, Université Paris-Saclay, 91400 Orsay, France
[4] Plant Pathology, Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland
[5] Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331-2902, USA
[6] Department of Evolutionary Biology and Environmental Studies, University of Zurich, Zurich, Switzerland

Author contributions: UO, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft; TB, Formal analysis; TW, Data curation, Supervision; FEH, Data curation, Resources, Supervision; NKS, LNA, PK, CCM, BAM, Resources; TV, Investigation; DC, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - original draft, Writing - review and editing

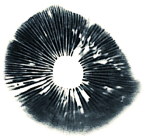
Published in eLife, 2021, Volume 10

"I'm not a monster. I'm a shark!"
Noelle Stevenson (Nimona)

Abstract

Genome evolution is driven by the activity of transposable elements (TEs). The spread of TEs can have deleterious effects including the destabilization of genome integrity and expansions. However, the precise triggers of genome expansions remain poorly understood because genome size evolution is typically investigated only among deeply divergent lineages. Here, we use a large population genomics dataset of 284 individuals from populations across the globe of *Zymoseptoria tritici*, a major fungal wheat pathogen. We built a robust map of genome-wide TE insertions and deletions to track a total of 2,456 polymorphic loci within the species. We show that purifying selection substantially depressed TE frequencies in most populations but some rare TEs have recently risen in frequency and likely confer benefits. We found that specific TE families have undergone a substantial genome-wide expansion from the pathogen's center of origin to more recently founded populations. The most dramatic increase in TE insertions occurred between a pair of North American populations collected in the same field at an interval of 25 years. We find that both genome-wide counts of TE insertions and genome size have increased with colonization bottlenecks. Hence, the demographic history likely played a major role in shaping genome evolution within the species. We show that both the activation of specific TEs and relaxed purifying selection underpin this incipient expansion of the genome. Our study establishes a model to recapitulate TE-driven genome evolution over deeper evolutionary timescales.

Introduction



Transposable elements (TEs) are mobile repetitive DNA sequences with the ability to independently insert into new regions of the genome. TEs are major drivers of genome instability and epigenetic change (Eichler & Sankoff, 2003). Insertion of TEs can disrupt coding sequences, trigger chromosomal rearrangements, or alter expression profiles of adjacent genes (Lim, 1988; Petrov *et al.*, 2003; Slotkin & Martienssen, 2007; Hollister & Gaut, 2009; Oliver *et al.*, 2013). Hence, TE activity can have phenotypic consequences and impact host fitness. While TE insertion dynamics are driven by the selfish interest for proliferation, the impact on the host can range from beneficial to highly deleterious. The most dramatic examples of TE insertions underpinned rapid adaptation of populations or species (Feschotte, 2008; Chuong *et al.*, 2017), particularly following environmental change or colonization events.

Beneficial TE insertions are expected to experience strong positive selection and rapid fixation in populations. However, most TE insertions have neutral or deleterious effects upon insertions. Purifying selection is expected to rapidly eliminate deleterious insertions from populations unless constrained by genetic drift (Walser *et al.*, 2006; Baucom *et al.*, 2008; Cridland *et al.*, 2013; Stuart *et al.*, 2016; Lai *et al.*, 2017; Stritt *et al.*, 2017). Additionally, genomic defense mechanisms can disable transposition activity. Across eukaryotes, epigenetic silencing is a shared defense mechanism against TEs (Slotkin & Martienssen, 2007). Fungi evolved an additional and highly specific defense system introducing repeat-induced point (RIP) mutations into any nearly identical set of sequences. The relative importance of demography, selection and genomic defenses determining the fate of TEs in populations remain poorly understood.

A crucial property predicting the invasion success of TEs in a genome is the transposition rate. TEs tend to expand through family-specific bursts of transposition followed by prolonged phases of transposition inactivity. Bursts of insertions of different retrotransposon families were observed across eukaryotic lineages including *Homo sapiens*, *Zea mays*, *Oryza sativa* and *Blumeria graminis* (Shen *et al.*, 1991; SanMiguel *et al.*, 1998; Eichler & Sankoff, 2003; Piegu *et al.*, 2006; Lu *et al.*, 2017; Frantzeskakis *et al.*, 2018). Prolonged bursts without effective counter-selection are thought to underpin genome expansions. In the symbiotic fungus *Cenococcum geophilum*, the burst of TEs resulted in a dramatically expanded genome compared to closely related species (Peter *et al.*, 2016). Similarly, a burst of a TE family in brown hydras led to an approximately three-fold increase of the genome size compared to related hydras (Wong *et al.*, 2019). Across the tree of life, genome sizes vary by orders of magnitude and enlarged genomes invariably show hallmarks of historic TE invasions (Kidwell, 2002). Population size variation is among the few correlates of genome size across major groups, suggesting that the efficacy of selection plays an important role in controlling TE activity (Lynch, 2007). Reduced selection efficacy against deleterious TE insertions is expected to lead to a ratchet-like increase in genome size. In fungi, TE-rich genomes often show an isochore structure alternating gene-rich and TE-rich compartments (Rouxel *et al.*, 2011). TE-rich compartments often harbor rapidly evolving genes such as effector genes in pathogens or resistance genes in plants (Raffaele & Kamoun, 2012; Jiao & Schneeberger, 2019). Taken together, incipient genome expansions are likely driven by population-level TE insertion dynamics.

The fungal wheat pathogen *Zymoseptoria tritici* is one of the most important pathogens on crops, causing high yield losses in many years (Torriani *et al.*, 2015). *Z. tritici* emerged during the domestication of wheat in the Fertile Crescent where the species retained high levels of genetic variation (Zhan *et al.*, 2005; Stukenbrock *et al.*, 2011). The pathogen migrated to all temperate zones where wheat is currently grown and underwent multiple migration bottlenecks, in particular when colonizing Oceania and North America (Zhan *et al.*, 2005; Estep *et al.*, 2015). The genome is completely assembled and shows size variation between individuals sampled across the global distribution range (Feurtey *et al.*, 2020; Badet *et al.*, 2020) (Goodwin *et al.*, 2011). The TE content of the genome shows a striking variation of 17-24% variation among individuals (Badet *et al.*, 2020). *Z. tritici* recently gained major TE-mediated adaptations to colonize host plants and tolerate environmental stress (Omrane *et al.*, 2015, 2017; Krishnan *et al.*, 2018; Meile *et al.*, 2018). Clusters of TEs are often associated with genes encoding important pathogenicity functions (*i.e.* effectors), recent gene gains or losses (Hartmann & Croll, 2017), and major chromosomal rearrangements (Croll *et al.*, 2013; Plissonneau *et al.*, 2016). Transposition activity of TEs also had a genome-wide impact on gene expression profiles during infection (Fouché *et al.*, 2019). The well-characterized demographic history of the pathogen and evidence for recent TE-mediated adaptations make *Z. tritici* an ideal model to recapitulate the process of TE insertion dynamics, adaptive evolution and changes in genome size at the population level.

Here, we retrace the population-level context of TE insertion dynamics and genome size changes across the species range by analyzing populations sampled on four continents for a total of 284 genomes. We developed a robust pipeline to detect newly inserted TEs using short read sequencing datasets. Combining analyses of selection and knowledge of the colonization history of the pathogen, we tested whether population bottlenecks were associated with substantial changes in the TE content and the size of genomes.

Results

A Dynamic TE landscape shaped by strong purifying selection

We detected 4,753 TE copies, grouped into 30 families with highly variable copy numbers in the reference genome IPO323 (Figure 2-source data 1 and Figure 2-figure supplement 1A). To establish a comprehensive picture of within-species TE dynamics, we analyzed 295 genomes from a worldwide set of six populations spanning the distribution range of the wheat pathogen *Z. tritici*. To ascertain the presence or absence of TEs across the genome, we developed a robust pipeline (Figure 1A). In summary, we called TE insertions by identifying reads mapping both to a TE sequence and a specific location in the reference genome. Then, we assessed the minimum sequencing coverage to reliably recover TE insertions and removed 11 genomes with an average read depth below 15X (Figure 1B). We tested for evidence of TEs using read depth at target site duplications (Figure 1C) and scanned the genome for mapped reads indicating gaps at TE loci (Figure 1D). We found robust evidence for a total of 18,864 TE insertions grouping into 2,465 individual loci. Of these loci, 35.5% (n = 876) have singleton TEs (*i.e.*, this locus is only present in one isolate: Figure 2A, Figure 2-source data 3). An overwhelming proportion of loci (2,345 loci or 95.1%) have a TE frequency below 1%. Singleton TE insertions in particular can be the product of spurious Illumina read mapping errors (Nakamura *et al.*, 2011). To assess the reliability of the detected singletons, we focused on seven isolates for which PacBio long-read data was available (Badet *et al.*, 2020). Aligned PacBio reads confirmed the exact location of 71% (22 out of 31 singleton insertions among seven isolates; see Methods for further details). We found no significant difference in read coverage between confirmed and unconfirmed singleton insertions (Figure 2 - figure supplement 1C-B and Figure 2-source data 2).

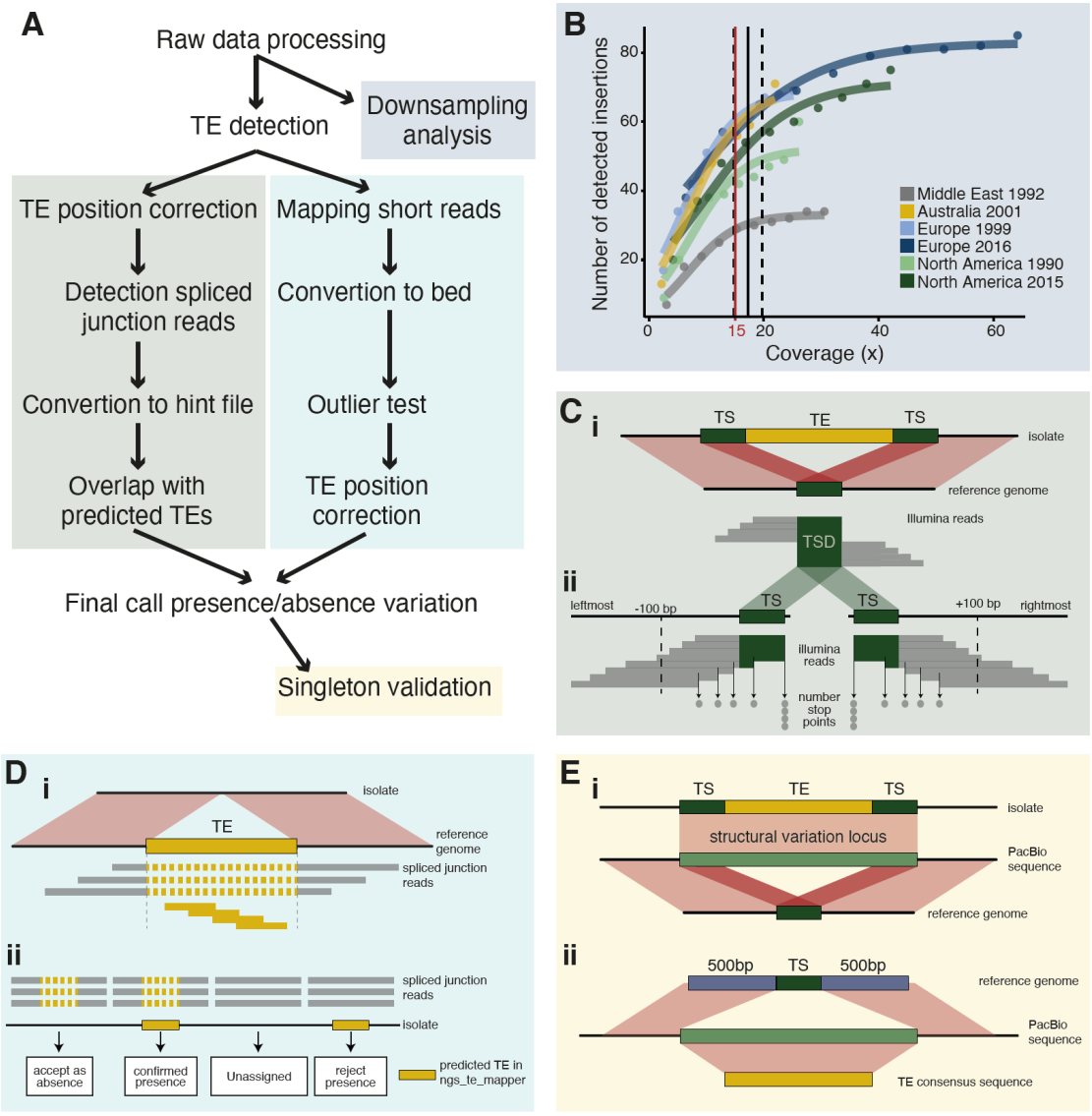


Figure 1: Robust discovery and validation of transposable element (TE) insertions: (A) General analysis pipeline. (B) Read depth down-sampling analysis for one isolate per population with an average coverage of the population. The vertical black line indicates the coverage at which on average 90% of the maximally detectable variants were recovered. Dashed black lines indicate the standard error. The threshold for a minimal mean coverage was set at 15X (red line). (C) Validation of insertions absent in the reference genome. (i) TE insertions that are not present in the reference genome show a duplication of the target site and the part of the reads that covers the TE will not be mapped against the reference genome. We thus expect reads to map to the TE surrounding region and the target site duplication but not the TE itself. At the target site, a local duplication of read depth is expected. (ii) We selected all reads in an interval of 100 bp up- and downstream including the target site duplication to detect deviations in the number of reads terminating near the target site duplication. (D) Validation of insertions present in the reference genome. (i) Analyses read coverage at target site duplications. (ii) Decision map if a TE should be kept as a true insertion or rejected as a false positive. Only predicted TE insertions that overlap evidence of split reads were kept as TE insertions in downstream analyses. (E) Singleton validation using long-read PacBio sequencing. (i) Analysis if TE insertions overlap with a detected insertion/deletion locus (Badet *et al*, 2021). (ii) Homology search of the TE insertion flanking sequences based on the reference genome against PacBio reads. In addition, the consensus sequence of the inserted TE was used for matches between the flanks. **Figure supplement 1.** Validation of transposable element (TE) insertion predictions. (A) TEs not present in the reference genome: distribution of additional TE hits found per locus after the outlier test. Color indicates superfamilies. (B) TEs not present in the reference genome: distribution of additional TE hits found per population after the outlier test. Colors indicate populations. (C) TEs present in the reference genome: distribution of missing data per locus after the validation with spliced junction reads. Missing data indicates that the TE was not predicted with ngs_te_mapper and that there was no indication of spliced reads. The red line (=20 %) indicates the threshold for missing data. TE loci with an amount of missing data > 20 % were completely excluded from further analyses. Color indicates superfamily. (D) TEs present in the reference genome: detection of strong outlier isolates with a high number of split reads. Color indicates the population. **Figure supplement 2.** Establishment of transposable element (TE) loci with differing start and end positions in the isolates. Distribution of length of distance for start position, end position and both start and end combined after the correction. **Figure supplement 3.** Bias for reads with a GC content lower than 30 % per population. Red lines indicate the mean. **Source data 1.** TE insertion validations for non-reference copies. **Source data 2.** TE consensus sequences.

The abundance of singleton TE insertions strongly supports the idea that TEs actively copy into new locations but also indicates that strong purifying selection maintains nearly all TEs at low frequency (Figure 2A). The density of TE loci on accessory chromosomes, which are not shared among all isolates of the species, is almost twice the density found on core chromosomes (102 *versus* 58 TEs per Mb; Figure 2B and Figure 2-figure supplement 2A). This suggests relaxed selection against TE insertion on the functionally dispensable and gene-poor accessory chromosomes. We found no difference in TE allele frequency distribution between recombination hotspots and the rest of the genome (Figure 2-figure supplement 2B). Similarly, the TE density and the number of insertions did not vary between recombination hotspots and the genomic background (Figure 2-figure supplement 2C).

TEs grouped into 23 families and 11 superfamilies, with 88.2% of all copies belonging to class I/retrotransposons (n = 2175; Figure 2C and Figure 2-figure supplement 3A-B). RLG/*Gypsy* (n = 1,483) and RLC/*Copia* (n = 623) elements constitute the largest long terminal repeats (LTR) superfamilies. Class II/DNA transposons are dominated by DHH/*Helitron* (n = 249). As expected, TE families shared among fewer isolates tend to show also lower global copy numbers (*i.e.*, all isolates combined), while TE families that are present in all isolates generally have high global copy numbers (Figure 2D).

We detected 153 loci with TEs inserted into genes with most of the insertions being singletons (44.7%; n = 68) or of very low frequency (Figure 2E). Overall, TE insertions into exonic sequences were less frequent than expected compared to insertions into up- and downstream regions, which is consistent with effective purifying selection (Figure 2F). Insertions into introns were also strongly under-represented, likely due to the small size of most fungal introns (~ 50-100 bp) and the high probability of disrupting splicing or adjacent coding sequences. We also found that insertions 800-1000 bp away from coding sequences of a focal gene were under-represented. Given the high gene density, with an average spacing between genes of 1.744 kb, TE insertions within 800-1,000 bp of a coding gene tend to be near adjacent genes already. Taken together, TEs in the species show a high degree of transposition activity and are subject to strong purifying selection.

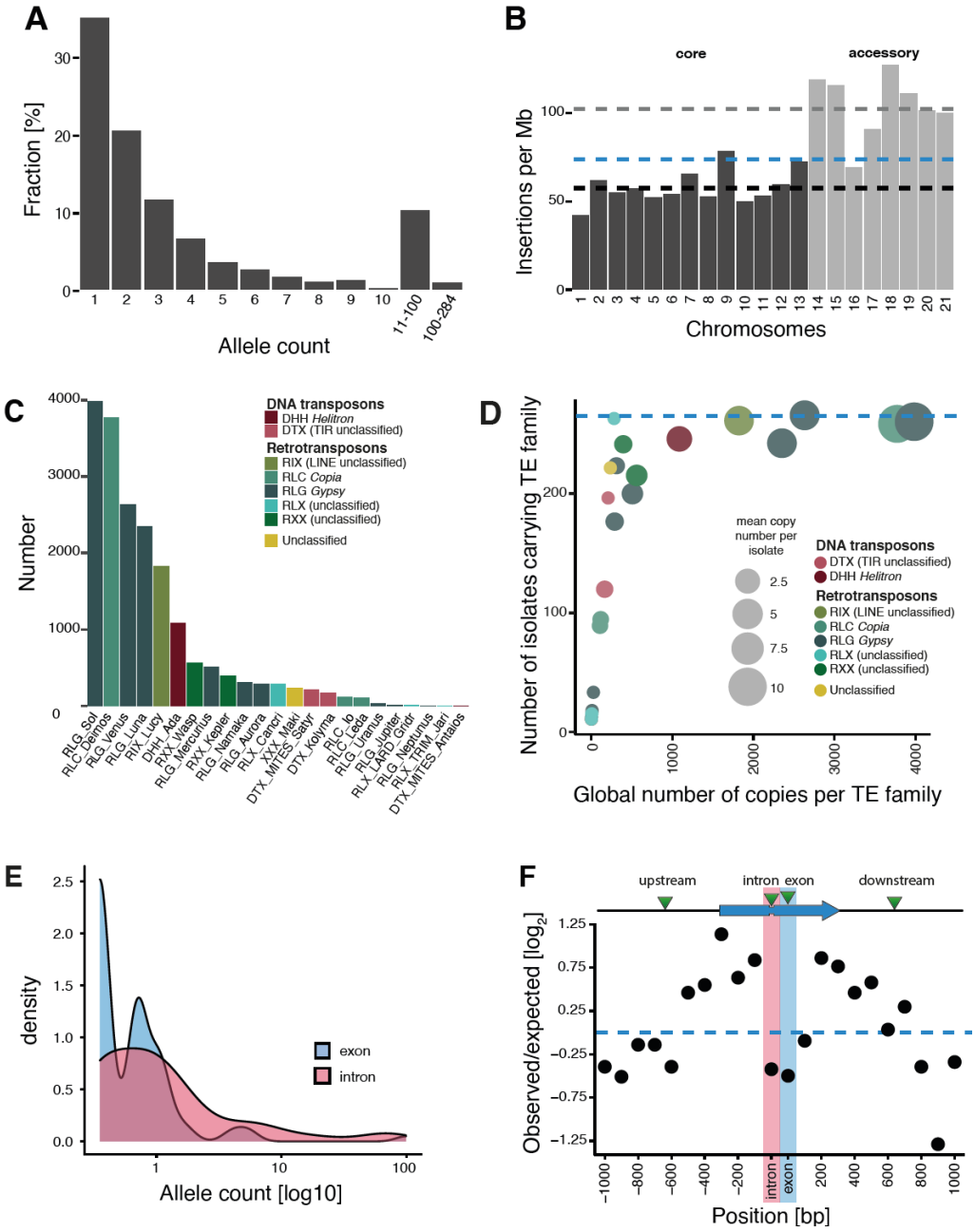


Figure 2: Transposable element (TE) landscape across populations. (A) Allele frequencies of the TE insertions across all isolates. (B) TE insertions per Mb on core chromosomes (dark) and accessory chromosomes (light). Dashed lines represent mean values. Blue: global mean of 75.65 insertions/Mb, dark: core chromosome mean of 58 TEs/Mb, light: accessory chromosome mean of 102.24 insertions/Mb). (C) Number of TE insertions per family. (D) TE frequencies among isolates and copy numbers across the genome. The blue line indicates the maximum number of isolates (n = 284). (E) Allele frequency distribution of TE insertions into introns and exons. (F) Number of TE insertions within 1 kb up- and downstream of genes on core chromosomes including introns and exons (100 bp windows). The blue arrow indicates a gene schematic with exons and an intron, the green triangles indicate TE insertions. The dotted blue line indicates no deviation from the expected value (*i.e.,* mean number of TEs per window). **Figure supplement 1.** Validation of singleton insertions detected by mapped Illumina reads using PacBio read alignments for confirmation. (A) Comparison of TE family copy numbers per isolate to the number of copies found in the reference genome (IPO323). The color is indicating superfamilies. This figure includes only TE families that were detected in any of the isolates used for validation. (B) Confirmation of singleton TE insertions detected in the isolates CH99_SW5, CH99_SW39, CH99_3D7, CH99_3D1, ISR92_Ar_4f, AUS01_1H8 and ORE90Ste_4A10 using aligned PacBio reads. Confirmed/not confirmed TE insertions are shown by TE family. (C) PacBio read coverage (in 500 bp window) at singleton loci. **Figure supplement 2.** TE insertion loci characteristics. (A) Number of TE insertions and density (insertions per Mb) in accessory and core genes. (B) Allele frequencies of TEs genome-wide and restricted to recombination hotspots. (C) TE insertion density and TE copy numbers within and outside of recombination hotspots. **Figure supplement 3.** Hierarchy superfamilies. (A) Number of transposable element (TE) insertions per superfamily. Colors indicate the superfamily. (B) Number of TE loci and classification hierarchy. (C) Comparison of mean genome sequencing coverage and the number of detected TEs with ngs_te_mapper in isolates of the Middle East population. Dots indicate the coverage and colors indicate the superfamily. **Source data 1.** TEs in reference. **Source data 2.** Presence absence matrix TE loci. **Source data 3 .** Singletons.

Detection of candidate TE Loci underlying recent adaptation

The TE transposition activity can generate adaptive genetic variation. To identify the most likely candidate loci, we analyzed insertion frequency variation among populations as an indicator for recent selection. Across all populations, the insertion frequencies differed only weakly with a strong skew towards extremely low F_{ST} values (mean = 0.0163; Figure 3A-B and Figure 3-figure supplement 1). To further analyze evidence for TE-mediated adaptive evolution, we screened a genome-wide SNP dataset for evidence of selective sweeps using selection scans. We found 16.5 % of all TE loci located in regions of selective sweep. Given our population sampling of two population pairs, we tested for adaptive TE insertions in selective sweep regions either in the North American or European population pairs. Hence, we selected loci having low TE insertion frequencies (< 5%) in all populations except either the recent North American or European population (> 20%) (Figure 3B). Based on these criteria, we obtained 7 candidate loci possibly underlying local adaptation (6 in North America, one in Europe; Figure 4A and Figure 4-source data 1). All loci carry inserted retrotransposons with 4 RLG_Luna, one RLG_Mercurius and one RLC_Deimos.

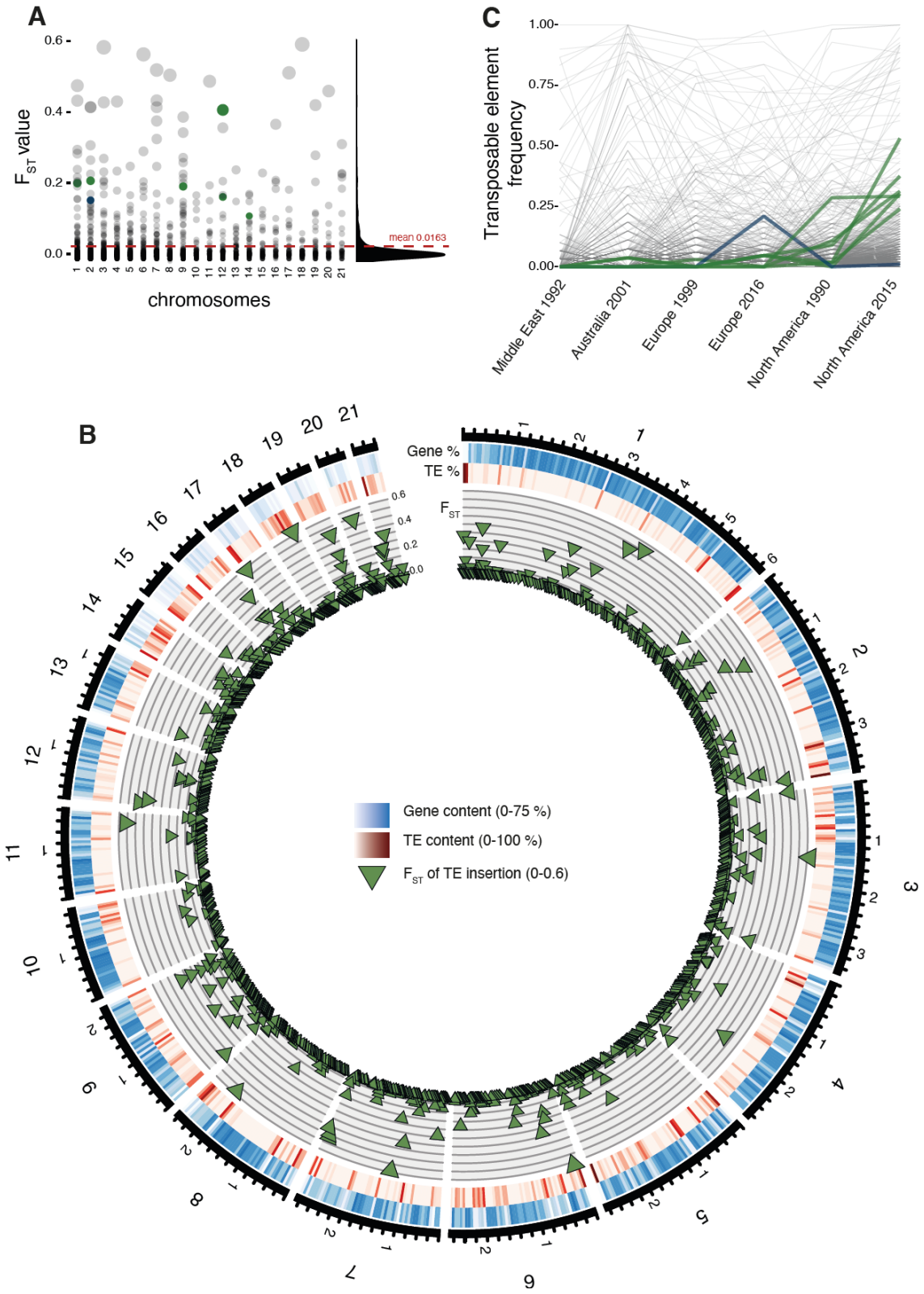


Figure 3: Differentiation in transposable element insertions frequencies across the genome. (A) Global pairwise F_{ST} distributions shown across the 21 chromosomes. The red horizontal line indicates the mean F_{ST} (= 0.0163). TEs with a strong local short-term frequency difference among populations are highlighted (blue: increase in Europe; green: increase in North America). (B) Allele frequency changes between the populations. The same TE loci as in panel A are highlighted. (C) Circos plot describing from the outside to the inside: The black line indicates chromosomal position in Mb. Blue bars indicate the gene density in windows of 100 kb with darker blue representing higher gene density. Red bars indicate the TE density in windows of 100 kb with a darker red representing higher TE density. Green triangles indicate positions of TE insertions with among population F_{ST} value shown on the y-axis. **Figure supplement 1.** Global pairwise FST distributions shown separately for the 21 chromosomes. The red horizontal line indicates the mean FST = 0.0163. Colors are according to the three main superfamilies (RLG, RLC, DHH).

One TE insertion is 3,815 bp downstream of a gene encoding an RTA1-like protein, which can function as transporters with a transmembrane domain and have been associated with resistance against several antifungal compounds (Soustre *et al.*, 1996). The insertion is also 5785 bp upstream of a gene encoding a protein kinase domain (Figure 4B). The TE insertion was not detected in the Middle East or the two European populations, and was at low frequencies in the Australian (3.7%) and North American 1990 (1.7%) populations, but increased to 53% of all isolates in the North American 2015 population (fixation index F_{ST} = 0.42; Figure 4-source data 1). Isolates that carry the insertion show a significantly higher resistance to azole antifungal compounds (Figure 4C). The TE is in the subtelomeric region of chromosome 12, with a moderate GC content, a low TE and a high gene density (Figure 4D). The TE belongs to the family RLG_Luna, which shows a substantial burst across different chromosomes within the species (Figures 4E-F). We found no association between the phylogenetic relationships among isolates based on the two closest genes and the presence or absence of the TE insertion (Figure 4G). A second candidate adaptive TE insertion belongs to the RLG_Mercurius family and is located between two genes of unknown function (Figure 4-figure supplement 1). A third potentially adaptive TE insertion of a RLC_Deimos is 229 bp upstream of a gene encoding a SNARE domain protein and 286 bp upstream of a gene encoding a flavin amine oxidoreductase. Furthermore, the TE is inserted in a selective sweep region (Figure 4-figure supplement 1). SNARE domains play a role in vesicular transport and membrane fusion (Bonifacino & Glick, 2004). An additional four candidates for adaptive TE insertions belong to RLG_Luna and were located distantly to genes (Figure 4-figure supplement 1). We experimentally tested whether the TE insertions in proximity to genes were associated with higher levels of fungicide resistance. For this, we measured growth rates of the fungal isolates in the presence or absence of an azole fungicide widely deployed against the pathogen. We found that the insertion of TEs at two loci was positively associated with higher levels of fungicide resistance, suggesting that the adaptation was mediated by the TE (Figure 4C and Figure 4-figure supplement 1).

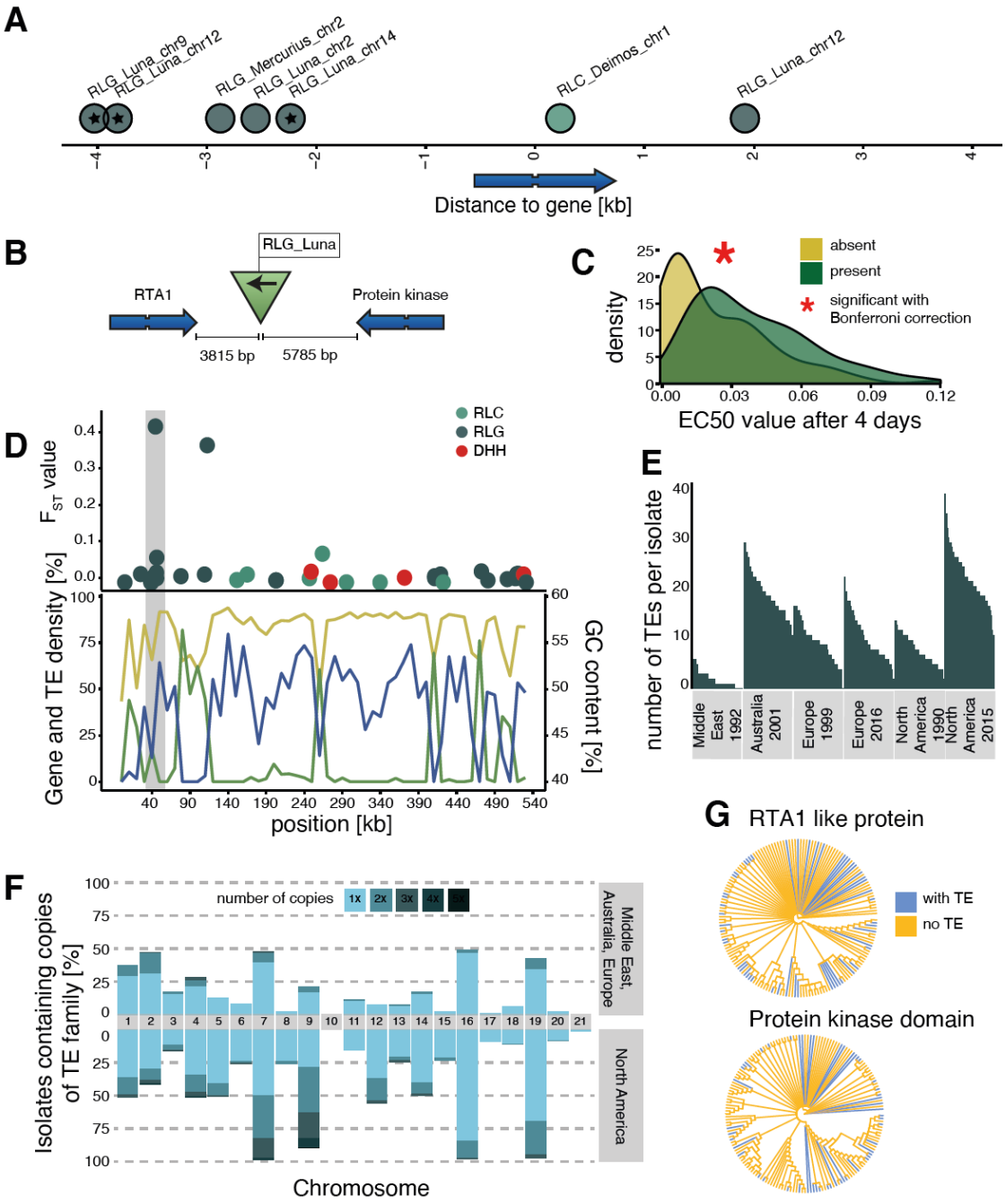


Figure 4: Candidate adaptive transposable element (TE) insertions. (A) Distribution of all extremely differentiated TEs and their distance to the closest gene. Color indicates the superfamily. The stars indicate TE insertions not found in the reference genome. (B) Location of the RLG_Luna TE insertion on chromosome 12 corresponding to its two closest genes. (C) Resistance against azole fungicides among isolates as a function of TE presence or absence. (D) Genomic niche of the RLG_Luna TE insertion on chromosome 12: F_{ST} values for each TE insertion, gene content (blue), TE content (green) and GC content (yellow). The grey section highlights the insertion site. (E) Number of RLG_Luna copies per isolate and population. (F) Frequency changes of RLG_Luna between the two North American populations compared to the other populations. Colors indicate the number of copies per chromosome. (G) Phylogenetic trees of the coding sequences of either the gene encoding the RTA1-like protein or the protein kinase domain. Isolates of the two North American populations and an additional 11 isolates from other populations not carrying the insertion are shown. Blue color indicates TE presence, yellow indicates TE absence. **Figure supplement 1.** Additional top loci. Six additional candidate adaptive transposable element (TE) insertions. Each row corresponds to a candidate, with the first five being candidates detected in the North American populations and the last one in the European populations. For each candidate, the direction of the TE and the direction, function and distance of the closest two genes are indicated. The middle column indicates the location of the TE in the genomic niche, with TE content, gene content and GC content for the surrounding windows. The third column indicates resistance levels towards azole antifungals for isolates with and without the TE insertion. **Source data 1.** Top loci information.

Population-level expansions in TE content

If TE insertion dynamics are largely neutral across populations, TE frequencies across loci should reflect neutral population structure. To test this, we performed a principal component analysis based on a set of six populations on four continents that represent the global genetic diversity of the pathogen (Figure 5A) and 900,193 genome-wide SNPs (Figure 5B). The population structure reflected the demographic history of the pathogen with clear continental differentiation and only minor within-site differentiation. To account for the lower number of TE loci, we performed an additional principal component analysis using a random SNP set of similar size to the number of TE loci. The reduced SNP set retained the geographic signal of the broader set of SNPs (Figure 5C). In stark contrast, TE frequencies across loci showed only weak clustering by geographic origin with the Australian population being the most distinct (Figure 5D). We found a surprisingly strong differentiation of the two North American populations sampled at a 25-year interval in the same field in Oregon.

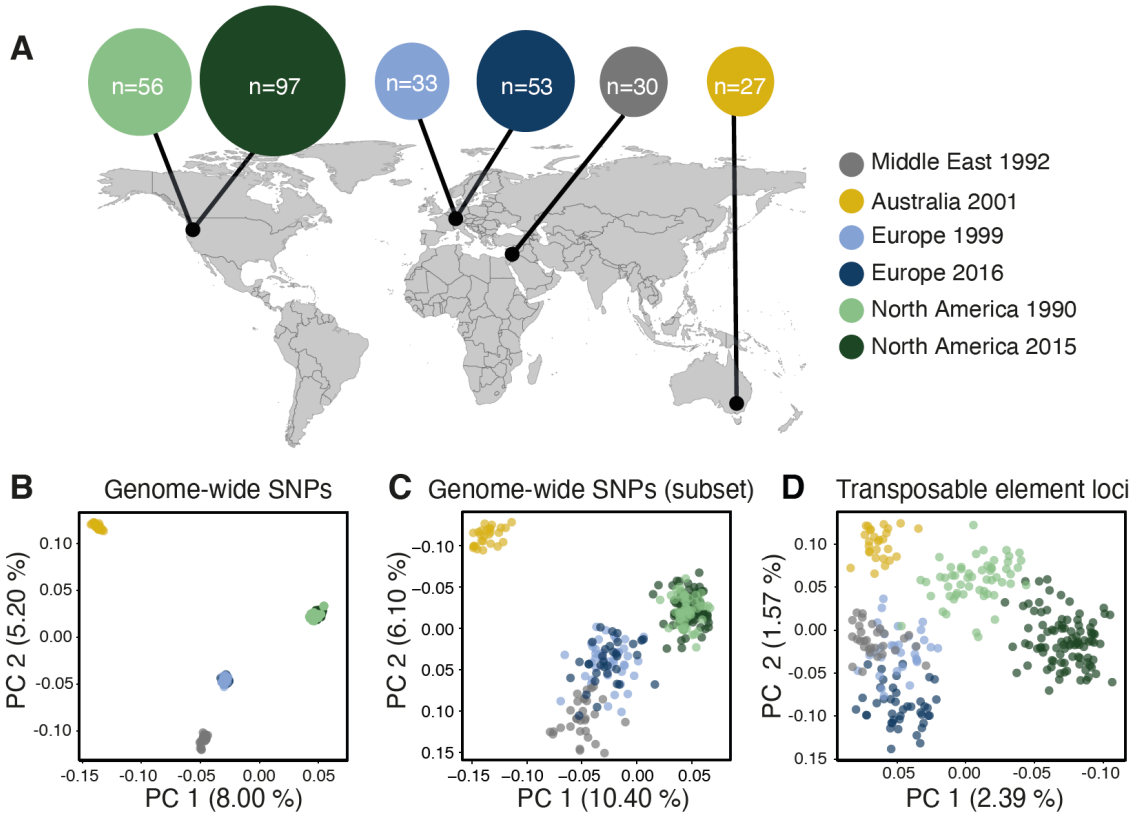


Figure 5: Population differentiation at transposable element (TE) and genome-wide SNP loci. (A) Sampling locations of the six populations. Middle East represents the region of origin of the pathogen. In North America, the two populations were collected at an interval of 25 years in the same field in Oregon. In Europe, two populations were collected at an interval of 17 years from two fields in Switzerland <20 km apart. Dark arrows indicate the historic colonization routes of the pathogen. (B) Principal component analysis (PCA) of 284 *Zymoseptoria tritici* isolates, based on 900,193 genome-wide SNPs. (C) PCA of a reduced SNP data set with randomly selected 203 SNPs matching approximately the number of analyzed TE loci. (D) PCA based on 193 TE insertion loci. Loci with allele frequency < 5% are excluded. **Source data 1.** Isolates.

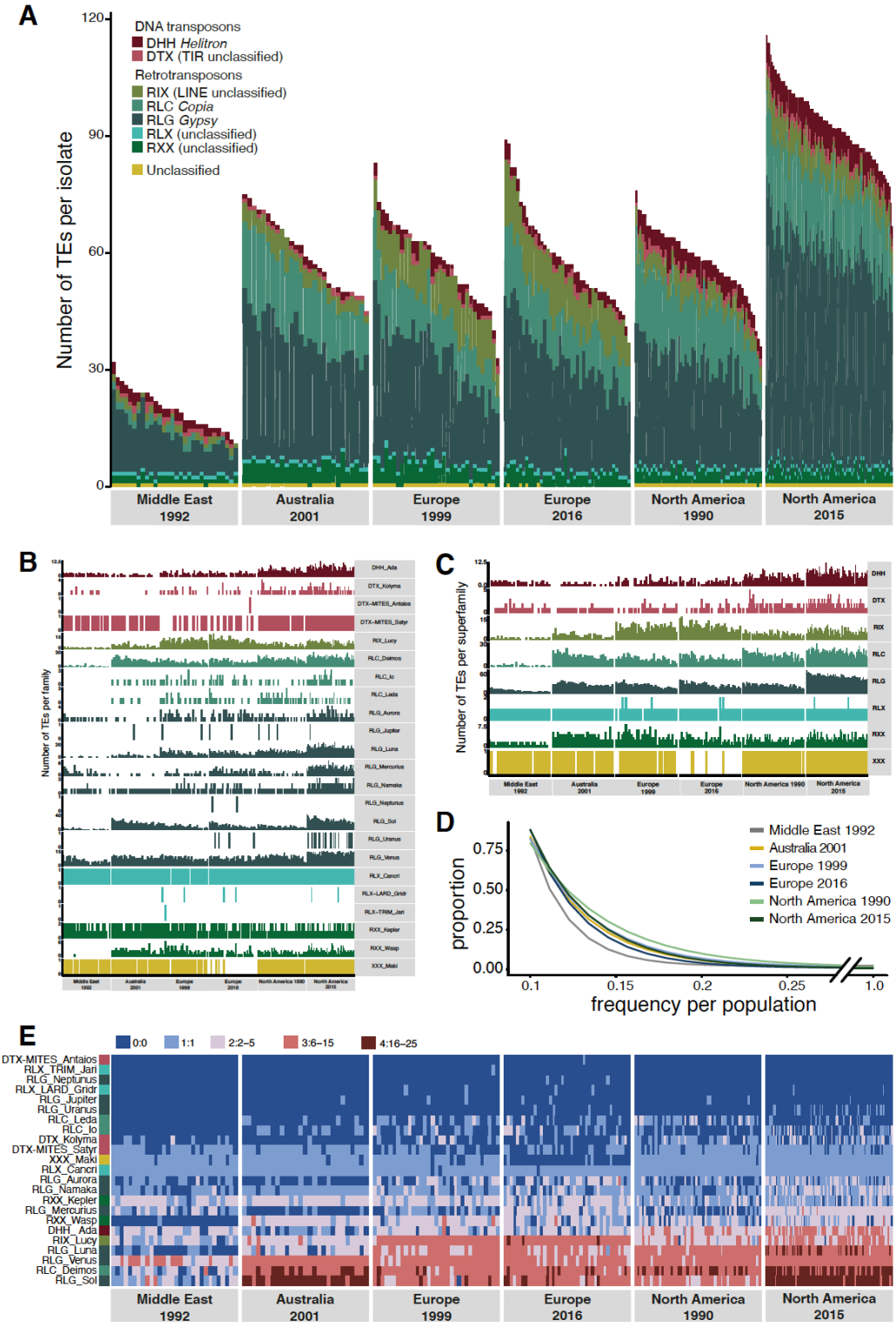


Figure 6: Global population structure of transposable element (TE) insertion polymorphism. (A) Total TE copies per isolate. Colors identify TE superfamilies. (B) TE copies per family and (C) superfamily. (D) TE insertion frequency spectrum per population. The curve fitting was performed with a self-starting Nls asymptomatic regression model (E). TE family copy numbers per isolate. **Figure supplement 1**. Population changes additional. Variation in transposable element (TE) content per isolate across populations. (A) Total TE copies per superfamily (colored) and per isolate only including LTR (long terminal repeat) TEs Copia and Gypsy. Color indicates the family. (B) Total TE copies per superfamily (colored) and per isolate only on the core chromosomes. (C) Total TE copies per superfamily (colored) and per isolate only on the accessory chromosomes. **Figure supplement 2**. Heatmap loci. (A) Presence (blue) and absence (yellow) matrix for all transposable element (TE) loci in all isolates per population. Colors on the left side indicate the superfamily. (B) Comparison of different genomic regions with and without TE insertions in IPO323. **Source data 1**. Kolmogorof-Smirnov.

Unusual patterns in population differentiation at TE loci suggests that TE activity may substantially vary across populations (Figure 6, Figure 4-source data 1). To analyze this, we first identified the total TE content across all loci per isolate. We found generally lower TE numbers in the Middle Eastern population from Israel (Figure 6A-C, and Figure 6-figure supplement 1), which is close to the pathogen's center of origin (Stukenbrock *et al.*, 2007). Populations that underwent at least one migration bottleneck showed a substantial burst of TEs across all major superfamilies. These populations included the two populations from Europe collected in 1999 and 2016 and the North American population from 1990, as well as the Australian population. We found a second stark increase in TE content in the North American population sampled in 2015 at the same site as the population from 1990. Strikingly, the isolate with the lowest number of analyzed TEs collected in 2015 was comparable to the isolate with the highest number of TEs at the same site in 1990. We tested whether sequencing coverage could explain variation in the detected TEs across isolates, but we found no meaningful association (Figure 2-figure supplement 3C). We analyzed whether the population-specific expansions were correlated with shifts in the frequency spectrum of TEs in the populations (Figure 6D). We found that the first step of expansions observed in Europe compared to the Middle East (Israel) was associated with an upwards shift in allele frequencies. This is consistent with transposition activity creating new copies in the genomes and stronger purifying selection in the Middle East. Similarly, the North American populations showed also signatures consistent with relaxation of selection against TEs (*i.e.*, fewer low frequency TEs). We found a significant difference (Two-sample Kolmogorov-Smirnov test, two-sided) in the curve shapes between the population from the Middle East and North America 2015 (Figure 6-source data 1). We analyzed variation in TE copy numbers across families and found that the expansions were mostly driven by RLG elements including the families Luna, Sol and Venus, the RLC family Deimos and the LINE family Lucy (Figure 6E and Figure 6-figure supplement 2). We also found a North American specific burst in DHH elements of the family Ada (increase from 4.6 to 6.1 copies on average per isolate), an increase specific to Swiss populations in LINE elements, and an increase in RLC elements in the Australian and the two North American populations. Analyses of complete *Z. tritici* reference-quality genomes that include isolates from the Israel, Australia, Switzerland (1999) and North American (1990) population revealed high TE contents in Australia and North America (Oregon 1990) (Badet *et al.*, 2020). The reference-quality genomes confirmed also that the increase in TEs was driven by LINE, RLG and RLC families in Australia and DHH, RLG and RLC families in North America (Badet *et al.*, 2020).

TE-mediated genome size expansions

The combined effects of actively copying TE families and relaxed purifying selection leads to an accumulation of new TE insertions in populations. Consequently, mean genome sizes in populations should increase over generations. We estimated the cumulative length of TE insertions based on the length of the corresponding TE consensus sequences and found a strong increase in the total TE length in populations outside the Middle East center of origin, and a second increase between the two North American populations (Figure 7-figure supplement 1A). To test for incipient genome expansions within the species, we first assembled genomes of all 284 isolates included in the study. Given the limitations of short-read assemblies, we implemented corrective measures to compensate for potential variation in assembly qualities. We corrected for variation in the GC content of different sequencing datasets by downsampling reads to generate balanced sequencing read sets prior to assembly (see Methods). We also excluded all reads mapping to accessory chromosomes because different isolates are known to differ in the number of these chromosomes. Genome assemblies were checked for completeness by retrieving the phylogenetically conserved BUSCO genes (Figure 7A). Genome assemblies across different populations carry generally >99% complete BUSCO gene sets, matching the completeness of reference-quality genomes of the same species (Badet *et al.*, 2020). The completeness of the assemblies showed no correlation with either TE or GC content of the genomes. GC content was inversely correlated with genome size consistent with the expansion of repetitive regions having generally low GC content (Figure 7B). We found that the core genome size varied substantially among populations with the Middle East, Australia as well as the two older European and North American populations having the smallest core genome sizes (Figure 7C). We found a notable increase in core genome size in both the more recent European and North American populations. The increase in core genome size is positively correlated with the count and cumulative length of all inserted TEs (Figure 7D, 7E and 7G) and negatively correlated with the genome-wide GC content (Figure 7F and 7G). Hence, core genome size shows substantial variation within the species matching the recent expansion in TEs across continents. We found the most variable genome sizes in the more recent North American population (Figure 7-figure supplement 1B). Finally, we contrasted variation in genome size with the detected TE insertion dynamics. For this, we assessed the variable genome segment as the difference between the smallest and largest analyzed core genome. To reflect TE dynamics, we calculated the cumulative length of all detected TE insertions in any given genome. We found that the cumulative length of inserted

TEs represents between 4.8 and 184 % of the variable genome segment defined for the species or 0.2-2.6% of the estimated genome size per isolate (Figure 7-figure supplement 1C-D).

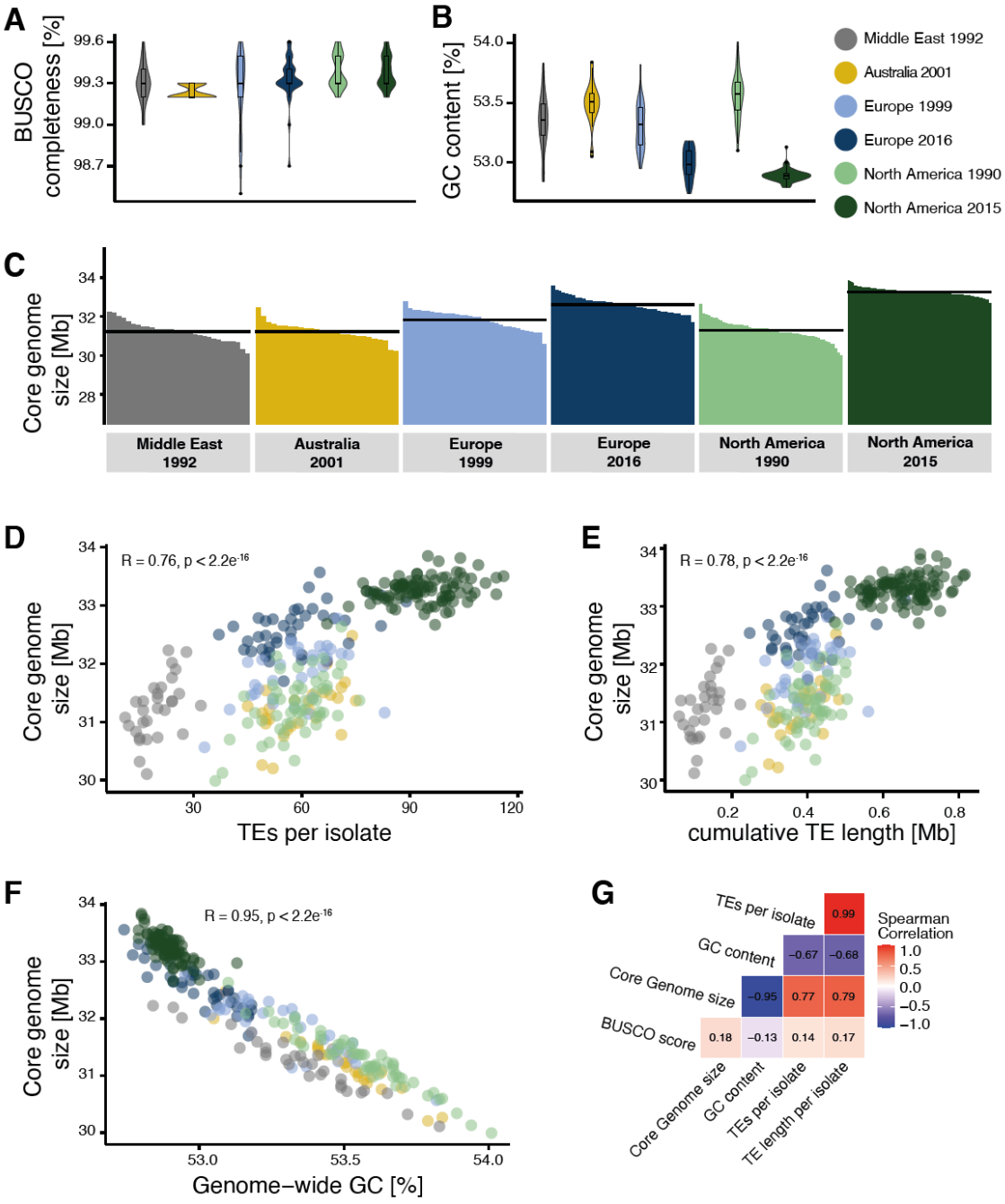


Figure 7: Core genome size and transposable element (TE) evolution across populations. (A) BUSCO completeness variation among genome assemblies. Black lines indicate the mean genome size per population. (B) Genome-wide GC content variation. (C) Core genome size variation among the isolates of the populations (excluding accessory chromosomes). (D) Correlation of core genome size and number of detected TEs. (E) Correlation of core genome size and the cumulative length of all TEs detected as inserted. (F) Correlation of core genome size and genome-wide GC content. (G) Spearman correlation matrix of BUSCO completeness, core genome size, number of detected TEs and genome-wide GC content. **Figure supplement 1.** Genome size expansion. (A) Estimated length of TE insertions per isolate and population. (B) Genome size variation per population. (C) Percentage of TEs content variation compared to the variation in genome size. (D) TE contributions to genome size variation compared to full genome size.

Discussion

TEs play a crucial role in generating adaptive genetic variation within species but are also drivers of deleterious genome expansions. We analyzed the interplay of TEs with selective and neutral processes including population differentiation and incipient genome expansions. TEs have substantial transposition activity in the genome but are strongly counter-selected and are maintained at low frequency. TE dynamics showed distinct trajectories across populations with more recently established populations having higher TE content and a concurrent expansion of the genome.

Recent selection acting on TE insertions

TE frequencies in the species show a strong skew towards singleton insertions across populations. However, our short-read based analyses are possibly skewed towards over-counting singletons as indicated by independent long-read mapping evaluations. Nevertheless, the skew towards low frequency TE insertions indicates both that TEs are undergoing transposition and that purifying selection maintains frequencies at a low level. Similar effects of selection on active TEs were observed across plants and animals, including *Drosophila melanogaster* and *Brachypodium distachyon* (Cridland *et al.*, 2013; Stritt *et al.*, 2017; Luo *et al.*, 2020). TE insertions were under-represented in or near coding regions, showing a stronger purifying selection against TEs inserting into genes. Coding sequences in the *Z. tritici* genome are densely packed with an average distance of only ~1 kb (Goodwin *et al.*, 2011). Consistent with this high gene density, TE insertions were most frequent at a distance of 200-400 bp away from coding sequences. A rapid decay in linkage disequilibrium in the *Z. tritici* populations (Croll *et al.*, 2015; Hartmann *et al.*, 2018) likely contributed to the efficiency of removing deleterious insertions. Some TE superfamilies have preferred insertion sites in coding regions and transcription start sites (Miyao *et al.*, 2003; Fu *et al.*, 2013; Gilly *et al.*, 2014; Quadrana *et al.*, 2016). Hence, some heterogeneity in the observed insertion site distribution across the genome is likely due to insertion preferences of individual TEs. We also found evidence for positive selection acting on TEs with the strongest candidate locus being a TE insertion on chromosome 12. This locus showed a frequency increase only in the more recent North American population, which experienced the first systematic fungicide applications and

subsequent emergence of fungicide resistance in the decade prior to the last sampling (Estep *et al.*, 2015). The nearest gene encodes a RTA1-like protein, a transmembrane exporter which is associated with resistance towards different stressors, including antifungal compounds, and shows strong copy number variation in several fungi (Soustre *et al.*, 1996; Rogers & Barker, 2003; Sirisattha *et al.*, 2004; Ali *et al.*, 2014; Yew *et al.*, 2016; Liang *et al.*, 2018). Hence, the TE insertion may have positively modulated RTA1 expression to resist antifungals.

Transposition activity in a genome and counter-acting purifying selection are expected to establish an equilibrium over evolutionary time (Charlesworth & Charlesworth, 1983). However, temporal bursts of TE families and changes in population size due to bottlenecks or founder events are likely to shift the equilibrium. Despite purifying selection, we were able to detect signatures of positive selection by scanning for short-term population frequency shifts. Population genomic datasets can be used to identify the most likely candidate loci underlying recent adaptation. The shallow genome-wide differentiation of *Z. tritici* populations provides a powerful background to test for outlier loci (Hartmann *et al.*, 2018). We found the same TE families to have experienced genome-wide copy number expansions, suggesting that the availability of adaptive TE insertions may be a by-product of TE bursts in individual populations.

Population-level TE invasions and relaxed selection

Across the surveyed populations from four continents, we identified substantial variation in TE counts per genome. The increase in TEs matches the global colonization history of the pathogen with an increase in TE copies in more recently established populations (Zhan *et al.*, 2003; Stukenbrock *et al.*, 2007). Compared to the Israeli population located nearest the center of origin in the Middle East, the European populations showed a three-fold increase in TE counts. The Australian and North American populations established from European descendants retained high TE counts. We identified a second increase at the North American site where TE counts nearly doubled again over a 25-year period. Compared to the broader increase in TEs from the Middle East, the second expansion at the North American site was driven by a small subset of TE families alone. Analyses of completely assembled reference-quality genomes from the same populations confirmed that genome expansions were primarily driven by the same TE families belonging to the RLG, RLC and DHH superfamilies (Badet *et al.*, 2020). Consistent with the contributions from individual TEs, we found that the first expansion in Europe led to an increase in low-frequency variants, suggesting higher transposition activity of many TEs in conjunction with strong purifying selection. The second expansion at the North

American site shifted TE frequencies upwards, suggesting relaxed selection against TEs. The population-level context of TEs in *Z. tritici* shows how heterogeneity in TE control interacts with demography to determine extant levels of TE content and, ultimately, genome size.

TE invasion dynamics underpins genome size expansions

The number of detected TEs was closely correlated with core genome size, hence genome size expansions were at least partly caused by the very recent proliferation of TEs. Genome assemblies of large eukaryotic genomes based on short read sequencing are often fragmented and contain chimeric sequences (Nagarajan & Pop, 2013). Focusing on the less repetitive core chromosomes in the genome of *Z. tritici* reduces such artefacts substantially. Because genome assemblies are the least complete in the most repetitive regions, any underrepresented sequences may rather underestimate than overestimate within-species variation in genome size. Hence, we consider the assembly sizes to be a robust correlate of total genome size. The core genome size differences observed across the species range match genome size variation typically observed among closely related species. Among primates, genome size varies by ~70% with ~10% between humans and chimpanzees (Rogers & Gibbs, 2014; Miga *et al.*, 2020). In fungi, genome size varies by several orders of magnitude within phyla but is often highly similar among closely related species (Raffaele & Kamoun, 2012). Interestingly, drastic changes in genome size have been observed in the *Blumeria* and *Pseudocercospora* genera where genome size changed by 35-130% between the closest known species (Frantzeskakis *et al.*, 2018; González-Sayer *et al.*, 2021). Beyond analyses of TE content variation correlating with genome size evolution, proximate mechanisms driving genome expansions are poorly understood. By establishing large population genetic datasets, such as those possible for crop pathogens, analyses of genome size evolution become tractable at the population level.

TEs might not only contribute to genome expansion directly by adding length through additional copies, but also by increasing the rate of chromosomal rearrangements and ectopic recombination (Bourque *et al.*, 2018; Blommaert, 2020). However, TEs are not the only repetitive elements that can lead to a genome size expansion. In *Arabidopsis thaliana* genomes, the 45S rDNA has been shown to have the strongest impact on genome size variation, followed by 5S rDNA variation, and contributions by centromeric repeats and TEs (Long *et al.*, 2013). In conjunction, recent work demonstrates how repetitive sequences are drivers of genome size evolution over short evolutionary timescales.

The activity of TEs is controlled by complex selection regimes within species. Actively transposing elements may accelerate genome evolution and underpin expansions. Hence, genomic defenses should evolve to efficiently target recently active TEs. Here, we show that TE activity and counteracting genomic defenses have established a tenuous equilibrium across the species range. We show that population subdivisions are at the origin of highly differentiated TE content within a species matching genome size changes emerging over the span of only decades and centuries. In conclusion, population-level analyses of genome size can recapitulate genome expansions typically observed across much deeper time scales providing fundamentally new insights into genome evolution.

Methods

Fungal isolate collection and sequencing

We analyzed 295 *Z. tritici* isolates covering six populations originating from four geographic locations and four continents (Figure 5-figure supplement 1), including: Middle East 1992 (*n* = 30 isolates, Nahal Oz, Israel), Australia 2001 (*n* = 27, Wagga Wagga), Europe 1999 (*n* = 33, Berg am Irchel, Switzerland), Europe 2016 (*n* = 52, Eschikon, ca. 15km from Berg am Irchel, Switzerland), North America 1990 and 2015 (*n* = 56 and *n* = 97, Willamette Valley, Oregon, United States) (McDonald *et al.*, 1996; Linde *et al.*, 2002; Zhan *et al.*, 2002, 2003, 2005). Illumina short read data from the Middle Eastern, Australian, European 1999 and North American 1990 populations were obtained from the NCBI Sequence Read Archive (SRA) under the BioProject PRJNA327615 (Hartmann *et al.*, 2017). For the Switzerland 2016 and Oregon 2015 populations, asexual spores were harvested from infected wheat leaves from naturally infected fields and grown in YSB liquid media including 50 mgL^{-1} kanamycin and stored in silica gel at −80°C. High-quality genomic DNA was extracted from liquid cultures using the DNeasy Plant Mini Kit from Qiagen (Venlo, Netherlands). The isolates were sequenced on an Illumina HiSeq in paired-end mode and raw reads were deposited at the NCBI SRA under the BioProject PRJNA596434.

TE insertion detection

The quality of Illumina short reads was determined with FastQC version 0.11.5 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) (Figure 1A). To remove spuriously sequenced Illumina adaptors and low quality reads, we trimmed the sequences with Trimmomatic version 0.36, using the following filter parameters: illuminaclip:TruSeq3-PE-2.fa:2:30:10 leading:10 trailing:10 slidingwindow:5:10 minlen:50 (Bolger *et al.*, 2014). We created repeat consensus sequences for TE families (sequences are available on https://github.com/crolllab/datasets; Figure 1-source data 2) in the complete reference genome IPO323 (Goodwin *et al.*, 2011) with RepeatModeler version open-4.0.7 (http://www.repeatmasker.org/RepeatModeler/) based on the RepBase Sequence Database and de novo (Bao *et al.*, 2015). TE classification into superfamilies and families was based on an approach combining detection of conserved protein sequences and tools to detect non-autonomous TEs (Badet *et al.*, 2020). To detect TE insertions, we used the R-based tool ngs_te_mapper version 79ef861f1d52cdd08eb2d51f145223fad0b2363c integrated into the McClintock pipeline version 20cb912497394fabddcdaa175402adacf5130bd1, using bwa version 0.7.4-r385 to map Illumina short reads, samtools version 0.1.19 to convert alignment file formats and R version 3.2.3 (Li & Durbin, 2009; Li *et al.*, 2009; Linheiro & Bergman, 2012; R Core Team, 2017; Nelson *et al.*, 2017).

Down-sampling analysis

We performed a down-sampling analysis to estimate the sensitivity of the TE detection with ngs_te_mapper based on variation in read depth. We selected one isolate per population matching the average coverage of the population. We extracted the per-base pair read depth with the genomecov function of bedtools version 2.27.1 and calculated the genome-wide mean read depth (Quinlan & Hall, 2010). The number of reads in the original fastq file was reduced in steps of 10% to simulate the impact of reduced coverage. We analyzed each of the obtained reduced read subsets with ngs_te_mapper using the same parameters as described above. The correlation between the number of detected insertions and the read depth was visualized using the function nls with model SSlogis in R and visualized with ggplot2 (Wickham, 2016). The number of detected TEs increased with the number of reads until reaching a plateau indicating saturation (Figure 1B). Saturation was reached at a coverage of approximately 15X, hence we retained only isolates with an average read depth above 15X for further analyses. We thus excluded one isolate from the Oregon 2015 population and ten isolates from the Switzerland 2016 population.

Validation procedure for predicted TE insertions

ngs_te_mapper detects the presence but not the absence of a TE at any given locus. We devised additional validation steps to ascertain both the presence as well as the absence of a TE across all loci in all individuals. TEs absent in the reference genome were validated by re-analyzing mapped Illumina reads. Reads spanning both parts of a TE sequence and an adjacent chromosomal sequence should only map to the reference genome sequence and cover the target site duplication of the TE (Figure 1C). We used bowtie2 version 2.3.0 with the parameter --very-sensitive-local to map Illumina short reads of each isolate on the reference genome IPO323 (Langmead & Salzberg, 2012). Mapped Illumina short reads were then sorted and indexed with samtools and the resulting bam file was converted to a bed file with the function bamtobed in bedtools. We extracted all mapped reads with an end point located within 100 bp of the target site duplication (Figure 1C). We tested whether the number of reads with a mapped end around the target site duplication significantly deviated if the mapping ended exactly at the boundary. A mapped read ending exactly at the target site duplication boundary is indicative of a split read mapping to a TE sequence absent in the reference genome. To test for the deviation in the number of read mappings around the target site duplication, we used a Poisson distribution and the *ppois* function in R version 3.5.1 (Figure 1C). We identified a TE as present in an isolate if tests on either side of the target site duplication had a p-value < 0.001 (Figure 5-figure supplement 1; Figure 1-figure supplement 1B and Figure 1-source data 1).

For TEs present in the reference genome, we analyzed evidence for spliced junction reads spanning the region containing the TE. Spliced reads are indicative of a discontinuous sequence and, hence, absence of the TE in a particular isolate (Figure 1D). We used STAR version 2.5.3a to detect spliced junction reads with the following set of parameters: --runThreadN 1 --outFilterMultimapNmax 100 --winAnchorMultimapNmax 200 --outSAMmultNmax 100 --outSAMtype BAM Unsorted --outFilterMismatchNmax 5 --alignIntronMin 150 --alignIntronMax 15000 (Dobin *et al.*, 2012). We then sorted and indexed the resulting bam file with samtools and converted split junction reads with the function bam2hints in bamtools version 2.5.1 (Barnett *et al.*, 2011). We selected loci without overlapping spliced junction reads using the function intersect in bedtools with the parameter -loj -v. We considered a TE as truly absent in an isolate if ngs_te_mapper did not detect a TE and evidence for spliced junction reads were found, indicating that the isolate had no inserted TE in this region. If the absence of a TE could not be confirmed by spliced junction reads, we labelled the genotype as missing.

Finally, we excluded TE loci with more than 20% missing data from further investigations (Figure 1D and Figure 1-figure supplement 1C).

Clustering of TE insertions into loci

We identified insertions across isolates as being the same locus if all detected TEs belonged to the same TE family and insertion sites differed by ≤100 bp (Figure 1-figure supplement 2). We used the R package *GenomicRanges* version 1.28.6 with the functions makeGRangesFromDataFrame and findOverlaps and the R package *devtools* version 1.13.4 (Lawrence *et al.*, 2013; Wickham & Chang, 2016). We used the R package *dplyr* version 0.7.4 to summarize datasets (https://dplyr.tidyverse.org/). Population-specific frequencies of insertions were calculated with the function allele.count in the R package *hierfstat* version 0.4.22 (Goudet, 2005). We conducted a principal component analysis for TE insertion frequencies filtering for a minor allele frequency ≥ 5%. We also performed a principal component analysis for genome-wide single nucleotide polymorphism (SNP) data obtained from Hartmann et al (2017) and Singh et al (2020). As described previously, SNPs were hard-filtered with VariantFiltration and SelectVariants tools integrated in the Genome Analysis Toolkit (GATK) (McKenna *et al.*, 2010). SNPs were removed if any of the following filter conditions applied: QUAL<250; QD<20.0; MQ<30.0; -2 > BaseQRankSum > 2; -2 > MQRankSum > 2; -2 > ReadPosRankSum > 2; FS>0.1. SNPs were excluded with vcftools version 0.1.17 and plink version 1.9 requiring a genotyping rate >90% and a minor allele frequency >5% (https://www.cog-genomics.org/plink2, Chang et al., 2015). Finally, we converted tri-allelic SNPs to bi-allelic SNPs by recoding the least frequent allele as a missing genotype. Principal component analysis was performed using the *gdsfmt* and *SNPRelate* packages in R (Zheng *et al.*, 2012, 2017). For a second principal component analysis with a reduced set of random markers, we randomly selected SNPs with vcftools and the following set of parameters: --maf 0.05 –thin 200000 to obtain an approximately equivalent number of SNPs as TE loci.

Evaluation of singleton insertions

To evaluate the reliability of singleton TE insertion loci, we analyzed singleton loci in isolates for which we had both Illumina datasets and complete reference-quality genomes (Badet *et al.*, 2020). From a set of 19 long-read PacBio reference genomes spanning the global distribution of *Z. tritici,* one isolate each from Australia, Israel, North America (1990) and four isolates

from Europe (1999) were also included in the TE insertion screening. To assess the reliability of singleton TE insertions, we first investigated structural variation analyses among the reference genomes (Badet *et al.*, 2021, Supplementary Data 1 and 2). The structural variation was called both based on split read mapping of PacBio reads and pairwise whole-genome alignments. Using bedtools intersect, we recovered for the 31 singleton TE loci in the 7 analyzed genomes a total of 17 loci showing either an indel, translocation, copy number polymorphism, duplication, inverted duplication, inversion, or inverted translocation at the same location. We visually inspected the PacBio read alignment bam files against the IPO323 reference genome using IGV version 2.4.16 (Robinson *et al.*, 2011), and found a typical coverage increase at the target site duplication, with most read mappings interrupted at the target site duplication as expected for an inserted TE. For the 14 remaining TE loci, we extracted the region of the predicted insertion and padded the sequence on both ends with an additional 500 bp using samtools faidx. We used blast to identify a homologous region in the assembled reference-quality genomes. Matching regions were inspected based on blastn for the presence of a TE sequence matching the TE family originally detected at the locus. With this second approach, we confirmed an additional five singletons to be true insertions. Both methods combined produced supportive evidence for 22 out of 31 singleton insertions (71%). We calculated the read coverage after mapping to the reference genome IPO323 with bedtools genomecov for each PacBio long-read dataset and calculated mean coverage for 500 bp regions around singleton TE insertions.

Population differentiation in TE frequencies

We calculated Nei's fixation index (F_{ST}) between pairs of populations using the R packages *hierfstat* and *adegenet* version 2.1.0 (Jombart, 2008; Jombart & Ahmed, 2011). To understand the chromosomal context of TE insertion loci across isolates, we analyzed draft genome assemblies. We generated *de novo* genome assemblies for all isolates using SPAdes version 3.5.0 with the parameter --careful and a kmer range of "21, 29, 37, 45, 53, 61, 79, 87" (Bankevich *et al.*, 2012). We used blastn to locate genes adjacent to TE insertion loci on genomic scaffolds of each isolate. We then extracted scaffold sequences surrounding 10 kb up- and downstream of the localized gene with the function faidx in samtools and reverse complemented the sequence if needed. Then, we performed multiple sequence alignments for each locus across all isolates with MAFFT version 7.407 with parameter --maxiterate 1000 (Katoh & Standley, 2013). We performed visual inspections to ensure correct alignments across isolates using Jalview version 2.10.5 (Waterhouse *et al.*, 2009). To generate

phylogenetic trees of individual gene or TE loci, we extracted specific sections of the alignment using the function extractalign in EMBOSS version 6.6.0 (Rice *et al.*, 2000) and converted the multiple sequence alignment into PHYLIP format with jmodeltest version 2.1.10 using the -getPhylip parameter. We then estimated maximum likelihood phylogenetic trees with the software PhyML version 3.0, the K80 substitution model and 100 bootstraps on the ATGC South of France bioinformatics platform (Guindon & Gascuel, 2003; Guindon *et al.*, 2010; Darriba *et al.*, 2012). Bifurcations with a supporting value lower than 10% were collapsed in TreeGraph version 2.15.0-887 beta and trees were visualized as circular phylograms in Dendroscope version 2.7.4 (Huson *et al.*, 2007; Stöver & Müller, 2010). For loci showing complex rearrangements, we generated synteny plots using 19 completely sequenced genomes from the same species using the R package *genoplotR* version 0.8.9 (Guy *et al.*, 2010; Badet *et al.*, 2020). We calculated the population-specific allele frequency for TE loci and estimated the exponential decay curve with a self-starting Nls asymptomatic regression model nls(p_loci ~ SSasymp(p_round, Asym, R0, lrc) in R.

We analyzed signatures of selective sweeps based on genome-wide SNPs using the extended haplotype homozygosity (EHH) tests implemented in the R package *REHH* (Sabeti *et al.*, 2007; Gautier & Vitalis, 2012). We analyzed within-population signatures based on the iHS statistic and chose a maximum gap distance of 20 kb. We also analyzed cross-population signatures based on the XP-EHH statistic for the following two population pairs: North America 1990 versus North America 2015, Europe 1999 versus Europe 2016. We defined significant selective sweeps as being among the 99.9th percentile outliers of the iHS and XP-EHH statistics. Significant SNPs at less than 5 kb were clustered into a single selective sweep region adding +/- 2.5 kb. Finally, we analyzed whether TE loci in the population pairs were within 10 kb of a region identified as a selective sweep by XP-EHH using the function intersect from bedtools.

Genomic location of TE insertions

To characterize the genomic environment of TE insertion loci, we split the reference genome into non-overlapping windows of 10 kb using the function splitter from EMBOSS. TEs were located in the reference genome using RepeatMasker providing consensus sequences from RepeatModeler (http://www.repeatmasker.org/). To analyze coding sequence, we retrieved the gene annotation for the reference genome (Grandaubert *et al.*, 2015). We estimated the percentage covered by genes or TEs per window using the function intersect in bedtools. Additionally, we calculated the GC content using the tool get_gc_content (https://github.com/spundhir/RNA-Seq/blob/master/get_gc_content.pl). We extracted the number of TEs present in 1 kb windows around each annotated core gene in the reference genome IPO323, using the function window in bedtools. We calculated the relative distances between each gene and the closest TE with the function bedtools closest. For the TEs inserted into genes, we used the function intersect in bedtools to distinguish intron and exon insertions with the parameters -wo and -v, respectively. TEs that overlap more than one exon were only counted once. For each 100 bp segment in the 1 kb windows as well as for introns and exons, we calculated the mean number of observed TE insertions per base pair. We calculated the mean number of TEs per window and calculated the log2 of the observed number of TE insertions divided by the expected value. We extracted information about recombination hotspots from Croll *et al.* (2015). This dataset is based on two experimental crosses initiated from isolates included in our analyses (1A5x1E4, 3D1x3D7). The recombination rates were assessed based on the reference genome IPO323 and analyzed with the *R/qtl* package in R. We used bedtools intersect to compare both TE density in IPO323 and TE insertion polymorphism with predicted recombination hotspots.

Core genome size estimation

Accessory chromosomes show presence/absence variation within the species and length polymorphism (Goodwin *et al.*, 2011; Croll *et al.*, 2013) and thus impact genome size. We controlled for this effect by first mapping sequencing reads to the reference genome IPO323 using bowtie2 with --very-sensitive-local settings and retained only reads mapping to any of the 13 core chromosomes using seqtk subseq v1.3-r106 (https://github.com/lh3/seqtk/). Furthermore, we found that different sequencing runs showed minor variation in the distribution of the per read GC content. In particular, reads of a GC content lower than 30 % were underrepresented in the Australian (mean reads < 30 % of the total readset: 0.05 %), North American 1990 (0.07 %) and Middle East (0.1 %) populations, and higher in the Europe 1999

(1.3 %), North American 2015 (3.0 %) and Europe 2016 (4.02 %) populations (Figure 1-figure supplement 3). Library preparation protocols and Illumina sequencer generations are known factors influencing the recovery of reads of varying GC content (Benjamini & Speed, 2012).

To control a potential bias stemming from this, we subsampled reads based on GC content to create homogeneous datasets. For this, we first retrieved the mean GC content for each read pair using geecee in EMBOSS and binned reads according to GC content. For the bins with a GC content <30%, we calculated the mean proportion of reads from the genome over all samples. We then used seqtk subseq to subsample reads of <30% to adjust the mean GC content among readsets. We generated *de novo* genome assemblies using the SPAdes assembler version with the parameters --careful and a kmer range of "21, 29, 37, 45, 53, 61, 79, 87". The SPAdes assembler is optimized for the assembly of relatively small eukaryotic genomes. We evaluated the completeness of the assemblies using BUSCO v4.1.1 with the fungi_odb10 gene test set (Simão *et al.*, 2015). We finally ran Quast v5.0.2 to retrieve assembly metrics including scaffolds of at least 1 kb (Mikheenko *et al.*, 2018).

Fungicide resistance assay

To quantify susceptibility towards propiconazole we used a previously published microtiter plate assay dataset with 3 replicates performed for each isolate and concentration. optical density was used to estimate growth rates under different fungicide concentrations (0, 0.00006, 0.00017, 0.0051, 0.0086, 0.015, 0.025, 0.042, 0.072, 0.20, 0.55, 1.5 mgL^{-1}) (Hartmann *et al.*, 2020). We calculated dose-response curves and estimated the half-maximal lethal concentration EC_{50} with a 4-parameter logistics curve in the R package *drc* (Ritz & Streibig, 2005).

Data availability

Sequence data is deposited at the NCBI SRA under the accession numbers PRJNA327615, PRJNA596434 and PRJNA178194. Transposable element consensus sequences are available from https://github.com/crolllab/datasets.

Supplementary Data available on https://zenodo.org/record/6029536.

Acknowledgments

We thank Andrea Sánchez Vallet, Anne C. Roulin, Luzia Stalder, Adam Taranto, Emilie Chanclud and Alice Feurtey for helpful discussions and comments on previous versions of the manuscript. We also thank the three reviewers for very helpful suggestions. We thank C. Sarai Reyes-Avila for advice on statistical analyses. DC is supported by the Swiss National Science (grants 31003A_173265) and the Fondation Pierre Mercier pour la Science.

Literature chapter 1

Ali Shahin S, Khan Mojibur, Mullins Ewen, Doohan Fiona M. 2014. Identification of Fusarium oxysporum Genes Associated with Lignocellulose Bioconversion Competency. Bioenergy Research 7: 110–119.

Badet Thomas, Fouché Simone, Hartmann Fanny E, Zala Marcello, Croll Daniel. 2021. Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. Nature Communications 12.

Badet Thomas, Oggenfuss Ursula, Abraham Leen, McDonald Bruce A, Croll Daniel. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biology 18: 12.

Bankevich Anton, Nurk Sergey, Antipov Dmitry, Gurevich Alexey A, Dvorkin Mikhail, Kulikov Alexander S, Lesin Valery M, Nikolenko Sergey I, Pham Son, Prjibelski Andrey D, Pyshkin Alexey V, Sirotkin Alexander V, Vyahhi Nikolay, Tesler Glenn, Alekseyev Max A, Pevzner Pavel A. 2012. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. Journal of computational biology : a journal of computational molecular cell biology 19: 455–77.

Bao Weidong, Kojima Kenji K, Kohany Oleksiy. 2015. Repbase Update, a database of repetitive elements in eukaryotic genomes. Mobile DNA 6: 4–9.

Barnett Derek W, Garrison Erik K, Quinlan Aaron R, Stŕmberg Michael P, Marth Gabor T. 2011. Bamtools: A C++ API and toolkit for analyzing and managing BAM files. Bioinformatics 27: 1691–1692.

Baucom Regina S, Estill James C, Leebens-Mack Jim, Bennetzen Jeffrey L. 2008. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Research 19: 243–254.

Benjamini Yuval, Speed Terence P. 2012. Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Research 40: 1–14.

Blommaert Julie. 2020. Genome size evolution: towards new model systems for old questions. Proceedings. Biological sciences 287: 20201441.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30: 2114–2120.

Bonifacino Juan S, Glick Benjamin S. 2004. The Mechanisms of Vesicle Budding and Fusion. Cell 116: 153–166.

Bourque Guillaume, Burns Kathleen H, Gehring Mary, Gorbunova Vera, Seluanov Andrei, Hammell Molly, Imbeault Michaël, Izsvák Zsuzsanna, Levin Henry L, Macfarlan Todd S, Mager Dixie L, Feschotte Cédric. 2018. Ten things you should know about transposable elements. Genome Biology 19: 199.

Charlesworth Brian, Charlesworth Deborah. 1983. The population dynamics of transposable elements. Genetical Research 42: 1–27.

Chuong EB, Elde NC, Feschotte Cédric. 2017. Regulatory activities of transposable elements: from conflicts to benefits. Nature Reviews Genetics 18: 71–86.

Cridland Julie M, Macdonald Stuart J, Long Anthony D, Thornton Kevin R. 2013. Abundance and distribution of transposable elements in two drosophila QTL mapping resources. Molecular Biology and Evolution 30: 2311–2327.

Croll D, Lendenmann MH, Stewart E, McDonald BA. 2015. The Impact of Recombination Hotspots on Genome Evolution of a Fungal Plant Pathogen. Genetics 201: 1213-U787.

Croll Daniel, Zala Marcello, McDonald Bruce A. 2013. Breakage-fusion-bridge Cycles and Large Insertions Contribute to the Rapid Evolution of Accessory Chromosomes in a Fungal Pathogen (Joseph Heitman, Ed.). PLOS Genetics 9: e1003567.

Darriba Diego, Taboada Guillermo L, Doallo Ramon, Posada David. 2012. jModelTest 2: more models, new heuristics and parallel computing. Nature Methods 9: 772.

Dobin Alexander, Davis Carrie A, Schlesinger Felix, Drenkow Jorg, Zaleski Chris, Jha Sonali, Gingeras Thomas R, Batut Philippe, Chaisson Mark. 2012. STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29: 15–21.

Eichler EE, Sankoff D. 2003. Structural dynamics of eukaryotic chromosome evolution. Science 301: 793–797.

Estep LK, Torriani SFF, Zala M, Anderson NP, Flowers MD, Mcdonald BA, Mundt CC, Brunner PC. 2015. Emergence and early evolution of fungicide resistance in North American populations of Zymoseptoria tritici. Plant Pathology 64: 961–971.

Feschotte Cédric. 2008. Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics 9: 397–405.

Feurtey Alice, Lorrain Cécile, Croll Daniel, Eschenbrenner Christoph, Freitag Michael, Habig Michael, Haueisen Janine, Möller Mareike, Schotanus Klaas, Stukenbrock Eva H. 2020. Genome compartmentalization predates species divergence in the plant pathogen genus Zymoseptoria. BMC genomics 21: 588.

Fouché Simone, Badet Thomas, Oggenfuss Ursula, Plissonneau Clémence, Francisco Carolina Sardinha, Croll Daniel. 2019. Stress-driven transposable element de-repression dynamics in a fungal pathogen. Molecular Biology and Evolution.

Frantzeskakis Lamprinos, Kracher Barbara, Kusch Stefan, Yoshikawa-Maekawa Makoto, Bauer Saskia, Pedersen Carsten, Spanu Pietro D, Maekawa Takaki, Schulze-Lefert Paul, Panstruga Ralph. 2018. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics 19: 1–23.

Fu Yu, Kawabe Akira, Etcheverry Mathilde, Ito Tasuku, Toyoda Atsushi, Fujiyama Asao, Colot Vincent, Tarutani Yoshiaki, Kakutani Tetsuji. 2013. Mobilization of a plant transposon by expression of the transposon-encoded anti-silencing factor. EMBO Journal 32: 2407–2417.

Gautier Mathieu, Vitalis Renaud. 2012. Rehh An R package to detect footprints of selection in genome-wide SNP data from haplotype structure. Bioinformatics 28: 1176–1177.

Gilly A, Etcheverry M, Madoui MA, Guy J, Quadrana L, Alberti A, Martin A, Heitkam T, Engelen S, Labadie K, Le Pen J, Wincker P, Colot V, Aury JM. 2014. TE-Tracker: systematic identification of transposition events through whole-genome resequencing. Bmc Bioinformatics 15.

González-Sayer Sandra, Oggenfuss Ursula, García Ibonne, Aristizabal Fabio. 2021. High-quality genome assembly of Pseudocercospora ulei the main threat to natural rubber trees. Genetics and Molecular Biology: 0–1.

Goodwin Stephen B, Ben M'Barek Sarrah, Dhillon Braham, Wittenberg Alexander HJ, Crane Charles F, Hane James K, Foster Andrew J, Van der Lee Theo AJ, Grimwood Jane, Aerts Andrea, Antoniw John, Bailey Andy, Bluhm Burt, Bowler Judith, Bristow Jim, van der Burgt Ate, Canto-Canché Blondy, Churchill Alice CL, Conde-Ferràez Laura, et al. 2011. Finished Genome of the Fungal Wheat Pathogen Mycosphaerella graminicola Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis (Harmit S Malik, Ed.). PLOS Genetics 7: e1002070.

Goudet J. 2005. Hierstat, a package for R to compute and test heirarchical F-statistics. Molecular Ecology Notes 5: 184–186.

Grandaubert J, Bhattacharyya A, Stukenbrock EH. 2015. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3-Genes Genomes Genetics 5: 1323–1333.

Guindon Stephane, Dufayard Jean-Francois, Lefort Vincent, Anisimova Maria, Hordijk Wim, Gascuel Olivier. 2010. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. Systematic Biology 59: 307–321.

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology 52: 696–704.

Guy Lionel, Kultima Jens Roat, Andersson Siv GE. 2010. GenoPlotR: comparative gene and genome visualization in R. Bioinformatics 26: 2334–2335.

Hartmann FE, Croll D. 2017. Distinct Trajectories of Massive Recent Gene Gains and Losses in Populations of a Microbial Eukaryotic Pathogen. Molecular Biology and Evolution.

Hartmann FE, McDonald MC, Croll D. 2018. Genome-wide evidence for divergent selection between populations of a major agricultural pathogen. Molecular Ecology 27: 2725–2741.

Hartmann Fanny E, Sánchez-Vallet Andrea, McDonald Bruce A, Croll Daniel. 2017. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. The ISME Journal 11: 1189–1204.

Hartmann Fanny E, Vonlanthen Tiziana, Singh Nikhil Kumar, McDonald Megan C, Milgate Andrew, Croll Daniel. 2020. The complex genomic basis of rapid convergent adaptation to pesticides across continents in a fungal plant pathogen. Molecular Ecology.

Hollister Jesse D, Gaut Brandon S. 2009. Epigenetic silencing of transposable elements: A trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Research 19: 1419–1428.

Huson Daniel H, Richter Daniel C, Rausch Christian, Dezulian Tobias, Franz Markus, Rupp Regula. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8: 1–6.

Jiao Wen-Biao, Schneeberger Korbinian. 2019. Chromosome-level assemblies of multiple Arabidopsis thaliana accessions reveal hotspots of genomic rearrangements. bioRxiv: 738880.

Jombart Thibaut. 2008. Adegenet: A R package for the multivariate analysis of genetic markers. Bioinformatics 24: 1403–1405.

Jombart Thibaut, Ahmed Ismaïl. 2011. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. Bioinformatics 27: 3070–3071.

Katoh Kazutaka, Standley Daron M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Kidwell Margaret G. 2002. Transposable elements and the evolution of genome size in eukaryotes. Genetica 115: 49–63.

Krishnan Parvathy, Meile Lukas, Plissonneau Clémence, Ma Xin, Hartmann Fanny E, Croll Daniel, McDonald Bruce A, Sánchez-Vallet Andrea. 2018. Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. BMC Biology 16: 1–18.

Lai Xianjun, Schnable James C, Liao Zhengqiao, Xu Jie, Zhang Gengyun, Li Chuan, Hu Erliang, Rong Tingzhao, Xu Yunbi, Lu Yanli. 2017. Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize. BMC Genomics 18: 1–13.

Langmead Ben, Salzberg Steven L. 2012. Fast gapped-read alignment with Bowtie 2. Nature Methods 9: 357–359.

Lawrence Michael, Huber Wolfgang, Pagès Hervé, Aboyoun Patrick, Carlson Marc, Gentleman Robert, Morgan Martin T, Carey Vincent J. 2013. Software for Computing and Annotating Genomic Ranges (Andreas Prlic, Ed.). PLOS Computational Biology 9: e1003118.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li Heng, Handsaker Bob, Wysoker Alec, Fennell Tim, Ruan Jue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Liang Xiaofei, Wang Bo, Dong Qiuyue, Li Lingnan, Rollins Jeffrey A, Zhang Rong, Sun Guangyu. 2018. Pathogenic adaptations of Colletotrichum fungi revealed by genome wide gene family evolutionary analyses. PLoS ONE 13: 1–25.

Lim Johng K. 1988. Intrachromosomal rearrangements mediated by hobo transposons in Drosophila melanogaster. PNAS 85: 9153–9157.

Linde CC, Zhan J, McDonald BA. 2002. Population Structure of Mycosphaerella graminicola : From Lesions to Continents. Phytopathology 92: 946–955.

Linheiro Raquel S, Bergman Casey M. 2012. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster (Jason E Stajich, Ed.). PLOS ONE 7: e30008.

Long Quan, Rabanal Fernando A, Meng Dazhe, Huber Christian D, Farlow Ashley, Platzer Alexander, Zhang Qingrun, Vilhjálmsson Bjarni J, Korte Arthur, Nizhynska Viktoria, Voronin Viktor, Korte Pamela, Sedman Laura, Mandáková Terezie, Lysak Martin A, Seren Ümit, Hellmann Ines, Nordborg Magnus. 2013. Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nature Genetics 45: 884–890.

Lu Lu, Chen Jinfeng, Robb Sofia MC, Okumoto Yutaka, Stajich Jason E, Wessler Susan R. 2017. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. Proceedings of the National Academy of Sciences: 201716459.

Luo Shiqi, Zhang Hong, Duan Yuange, Yao Xinmin, Clark Andrew G, Lu Jian. 2020. The evolutionary arms race between transposable elements and piRNAs in Drosophila melanogaster. BMC Evolutionary Biology 20: 14.

Lynch Michael. 2007. The Origins of Genome Architecture. Sunderland MA: Sinauer Associates.

McDonald Bruce A, Mundt Christopher C, Chen Ruey-shyang. 1996. The role of selection on the genetic structure of pathogen populations : Evidence from field experiments with Mycosphaerella graminicola on wheat. Euphytica 92: 73–80.

McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M, DePristo MA. 2010. The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. Genome Research 20: 1297–1303.

Meile Lukas, Croll Daniel, Brunner Patrick C, Plissonneau Clémence, Hartmann Fanny E, McDonald Bruce A, Sánchez-Vallet Andrea. 2018. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to septoria tritici blotch. New Phytologist 219: 1048–1061.

Miga Karen H, Koren Sergey, Rhie Arang, Vollger Mitchell R, Gershman Ariel, Bzikadze Andrey, Brooks Shelise, Howe Edmund, Porubsky David, Logsdon Glennis A, Schneider Valerie A, Potapova Tamara, Wood Jonathan, Chow William, Armstrong Joel, Fredrickson Jeanne, Pak Evgenia, Tigyi Kristof, Kremitzki Milinn, et al. 2020. Telomere-to-telomere assembly of a complete human X chromosome. Nature 585: 79–84.

Mikheenko Alla, Prjibelski Andrey, Saveliev Vladislav, Antipov Dmitry, Gurevich Alexey. 2018. Versatile genome assembly evaluation with QUAST-LG. Bioinformatics 34: i142–i150.

Miyao Akio, Tanaka Katsuyuki, Murata Kazumasa, Sawaki Hiromichi, Takeda Shin, Abe Kiyomi, Shinozuka Yoriko, Onosato Katsura, Hirochika Hirohiko. 2003. Target site specificity of the Tos17 retrotransposon shows a preference for insertion within genes and against insertion in retrotransposon-rich regions of the genome. Plant Cell 15: 1771–1780.

Nagarajan Niranjan, Pop Mihai. 2013. Sequence assembly demystified. Nature Reviews Genetics 14: 157–167.

Nakamura Kensuke, Oshima Taku, Morimoto Takuya, Ikeda Shun, Yoshikawa Hirofumi, Shiwa Yuh, Ishikawa Shu, Linak Margaret C, Hirai Aki, Takahashi Hiroki, Altaf-Ul-Amin Md, Ogasawara Naotake, Kanaya Shigehiko. 2011. Sequence-specific error profile of Illumina sequencers. Nucleic Acids Research 39.

Nelson Michael G, Linheiro Raquel S, Bergman Casey M. 2017. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. G3: Genes|Genomes|Genetics 7: 2763–2778.

Oliver Keith R, McComb Jen A, Greene Wayne K. 2013. Transposable elements: Powerful contributors to angiosperm evolution and diversity. Genome Biology and Evolution 5: 1886–1901.

Omrane Selim, Audéon Colette, Ignace Amandine, Duplaix Clémentine, Aouini Lamia, Kema Gert, Walker Anne-Sophie, Fillinger Sabine. 2017. Plasticity of the MFS1 promoter leads to multi drug resistance in the wheat pathogen Zymoseptoria tritici. mSphere: 1–42.

Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, Fillinger S. 2015. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. Environmental Microbiology 17: 2805–2823.

Peter Martina, Kohler Annegret, Ohm Robin A, Kuo Alan, Krützmann Jennifer, Morin Emmanuelle, Arend Matthias, Barry Kerrie W, Binder Manfred, Choi Cindy, Clum Alicia, Copeland Alex, Grisel Nadine, Haridas Sajeet, Kipfer Tabea, LaButti Kurt, Lindquist Erika, Lipzen Anna, Maire Renaud, et al. 2016. Ectomycorrhizal ecology is imprinted in the genome of the dominant symbiotic fungus Cenococcum geophilum. Nature Communications 7: 1–15.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. Molecular Biology and Evolution 20: 880–892.

Piegu Benoit, Guyot Romain, Picault Nathalie, Roulin Anne, Sanyal Abhijit, Kim Hyeran, Collura Kristi, Brar Darshan S, Jackson Scott, Wing Rod A, Panaud Olivier. 2006. Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in Oryza australiensis, a wild relative of rice. Genome Research 21: 1201.

Plissonneau Clémence, Stürchler Alessandra, Croll Daniel. 2016. The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat. mBio 7: 1–13.

Quadrana L, Silveira AB, Mayhew GF, LeBlanc C, Martienssen RA, Jeddeloh JA, Colot V. 2016. The Arabidopsis thaliana mobilome and its impact at the species level. Elife 5.

Quinlan Aaron R, Hall Ira M. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

R Core Team. 2017. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology 10: 417–430.

Rice Peter, Longden Lan, Bleasby Alan. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277.

Ritz Christian, Streibig Jens C. 2005. Bioassay analysis using R. Journal of Statistical Software 12: 1–22.

Robinson James T, Thorvaldsdóttir Helga, Winckler Wendy, Guttman Mitchell, Lander Eric S, Getz Gad, Mesirov Jill P. 2011. Integrative Genome Viewer. Nature Biotechnology 29: 24–6.

Rogers P David, Barker Katherine S. 2003. Genome-wide expression profile analysis reveals coordinately regulated genes associated with stepwise acquisition of azole resistance in Candida albicans clinical isolates. Antimicrobial Agents and Chemotherapy 47: 1220–1227.

Rogers Jeffrey, Gibbs Richard A. 2014. Content and Dynamics. Nature Reviews Genetics 15: 347–359.

Rouxel Thierry, Grandaubert Jonathan, Hane James K, Hoede Claire, van de Wouw Angela P, Couloux Arnaud, Dominguez Victoria, Anthouard Véronique, Bally Pascal, Bourras Salim, Cozijnsen Anton J, Ciuffetti Lynda M, Degrave Alexandre, Dilmaghani Azita, Duret Laurent, Fudal Isabelle, Goodwin Stephen B, Gout Lilian, Glaser Nicolas, et al. 2011. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nature communications 2: 202.

Sabeti Pardis C, Varilly Patrick, Fry Ben, Lohmueller Jason, Hostetter Elizabeth, Cotsapas Chris, Xie Xiaohui, Byrne Elizabeth H, McCarroll Steven A, Gaudet Rachelle, Schaffner Stephen F, Lander Eric S. 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. 1998. The paleontology of intergene retrotransposons of maize. Nature Genetics 20: 43–45.

Shen Richard M, Batzer Mark A, Deininger Prescott L. 1991. Evolution of the master Alu gene(s). Journal of Molecular Evolution 33: 311–320.

Simão Felipe A, Waterhouse Robert M, Ioannidis Panagiotis, Kriventseva Evgenia V., Zdobnov Evgeny M. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31: 3210–3212.

Singh Nikhil Kumar, Chanclud Emilie, Croll Daniel. 2020. Population-level deep sequencing reveals the interplay of clonal and sexual reproduction in the fungal wheat pathogen Zymoseptoria tritici.

Sirisattha Sophon, Momose Yuko, Kitagawa Emiko, Iwahashi Hitoshi. 2004. Toxicity of anionic detergents determined by Saccharomyces cerevisiae microarray analysis. Water Research 38: 61–70.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272–285.

Soustre Isabelle, Letourneux Yves, Karst Francis. 1996. Characterization of the Saccharomyces cerevisiae RTA1 gene involved in 7-aminocholesterol resistance. Current Genetics 30: 121–125.

Stöver Ben C, Müller Kai F. 2010. TreeGraph 2: Combining and visualizing evidence from different phylogenetic analyses. BMC Bioinformatics 11: 1–9.

Stritt Christoph, Gordon Sean P, Wicker Thomas, Vogel John P, Roulin Anne C. 2017. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass Brachypodium distachyon. Genome Biology and Evolution 10: 1–38.

Stuart Tim, Eichten Steven R, Cahn Jonathan, Karpievitch Yuliya V, Borevitz Justin O, Lister Ryan. 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife 5: 1–27.

Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. 2007. Origin and domestication of the fungal wheat pathogen Mycosphaerella graminicola via sympatric speciation. Molecular Biology and Evolution 24: 398–411.

Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li RQ, Zala M, McDonald BA, Wang J, Schierup MH. 2011. The making of a new pathogen: Insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. Genome Research 21: 2157–2166.

Torriani Stefano FF, Melichar James PE, Mills Colin, Pain Naomi, Sierotzki Helge, Courbot Mikaël. 2015. Zymoseptoria tritici: A major threat to wheat production, integrated approaches to control. Fungal Genetics and Biology 79: 8–12.

Walser Jean-Claude, Chen Bing, Feder Martin E. 2006. Heat-Shock Promoters: Targets for Evolution by P Transposable Elements in Drosophila. PLOS Genetics 2: e165.

Waterhouse Andrew M, Procter James B, Martin David MA, Clamp Michèle, Barton Geoffrey J. 2009. Jalview Version 2-A multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189–1191.

Wickham Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Wickham Hadley, Chang Winston. 2016. devtools: Tools to Make Developing R Packages Easier.

Wong Wai Yee, Simakov Oleg, Bridge Diane M, Cartwright Paulyn, Bellantuono Anthony J, Kuhn Anne, Holstein Thomas W, David Charles N, Steele Robert E, Martínez Daniel E. 2019. Expansion of a single transposable element family is associated with genome-size increase and radiation in the genus Hydra. Proceedings of the National Academy of Sciences 116: 22915–22917.

Yew Su Mei, Chan Chai Ling, Kuan Chee Sian, Toh Yue Fen, Ngeow Yun Fong, Na Shiang Ling, Lee Kok Wei, Hoh Chee Choong, Yee Wai Yan, Ng Kee Peng. 2016. The genome of newly classified Ochroconis mirabilis: Insights into fungal adaptation to different living conditions. BMC Genomics 17: 1–17.

Zhan J, Kema GHJ, Waalwijk C, McDonald BA. 2002. Distribution of mating type alleles in the wheat pathogen Mycosphaerella graminicola over spatial scales from lesions to continents. Fungal Genetics and Biology 36: 128–136.

Zhan J, Linde CC, Jurgens T, Merz U, Steinebrunner F, McDonald BA. 2005. Variation for neutral markers is correlated with variation for quantitative traits in the plant pathogenic fungus Mycosphaerella graminicola. Mol Ecol 14: 2683–2693.

Zhan J, Pettway RE, McDonald BA. 2003. The global genetic structure of the wheat pathogen Mycosphaerella graminicola is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. Fungal Genetics and Biology 38: 286–297.

Zheng Xiuwen, Gogarten Stephanie M, Lawrence Michael, Stilp Adrienne, Conomos Matthew P, Weir Bruce S, Laurie Cathy, Levine David. 2017. SeqArray-a storage-efficient high-performance data format for WGS variant calls. Bioinformatics 33: 2251–2257.

Zheng Xiuwen, Levine David, Shen Jess, Gogarten Stephanie M, Laurie Cathy, Weir Bruce S. 2012. A high-performance computing toolset for relatedness and principal component analysis of SNP data. Bioinformatics 28: 3326–3328.

Supplementary Files

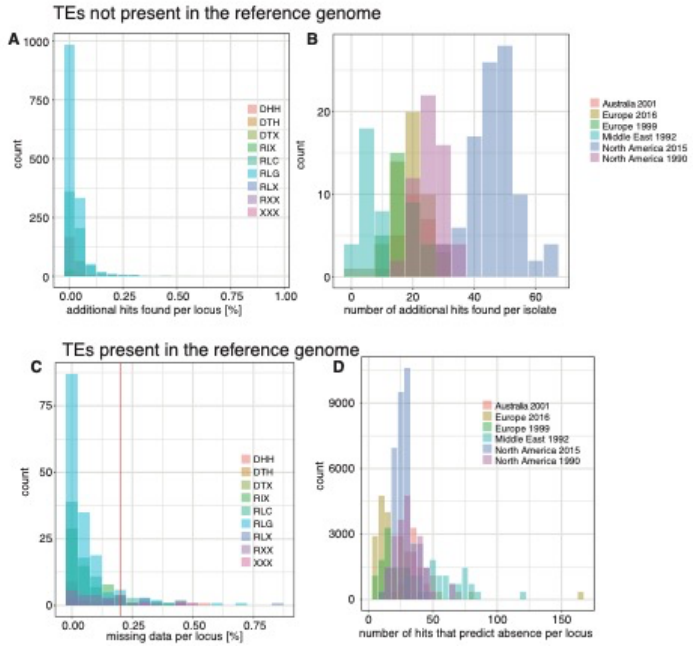


Figure 1 supplement 1: Validation of transposable element (TE) insertion predictions. (A) TEs not present in the reference genome: distribution of additional TE hits found per locus after the outlier test. Color indicates superfamilies. (B) TEs not present in the reference genome: distribution of additional TE hits found per population after the outlier test. Colors indicate populations. (C) TEs present in the reference genome: distribution of missing data per locus after the validation with spliced junction reads. Missing data indicates that the TE was not predicted with ngs_te_mapper and that there was no indication of spliced reads. The red line (=20 %) indicates the threshold for missing data. TE loci with an amount of missing data > 20 % were completely excluded from further analyses. Color indicates superfamily. (D) TEs present in the reference genome: detection of strong outlier isolates with a high number of split reads. Color indicates the population.

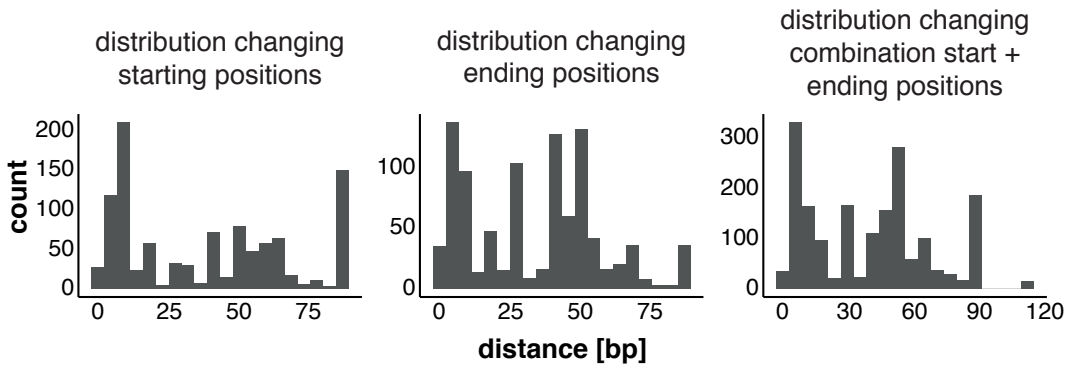


Figure 1 supplement 2: Establishment of transposable element (TE) loci with differing start and end positions in the isolates. Distribution of length of distance for start position, end position and both start and end combined after the correction.

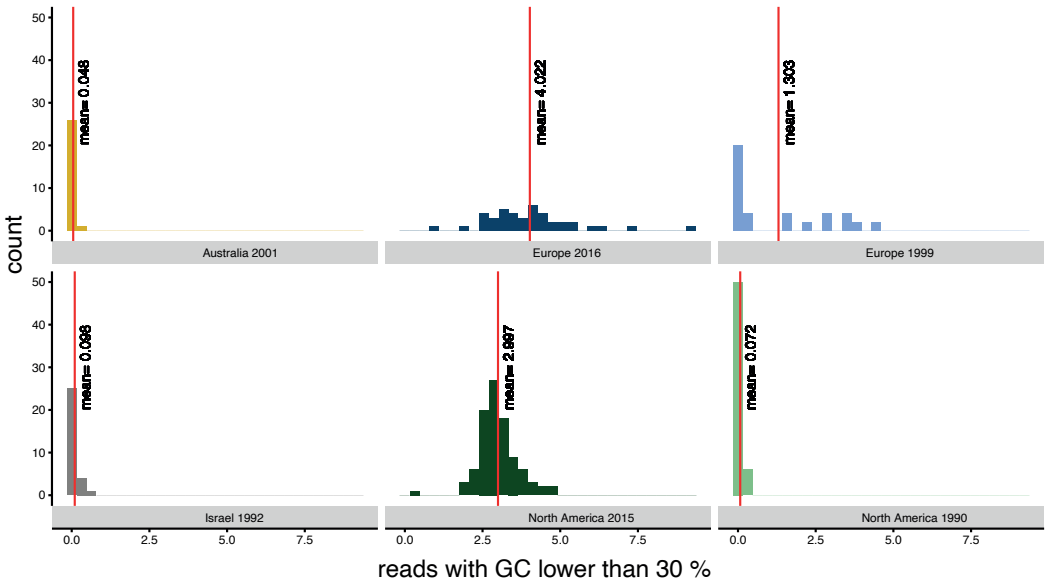


Figure 1 supplement 3: Bias for reads with a GC content lower than 30 % per population. Red lines indicate the mean.

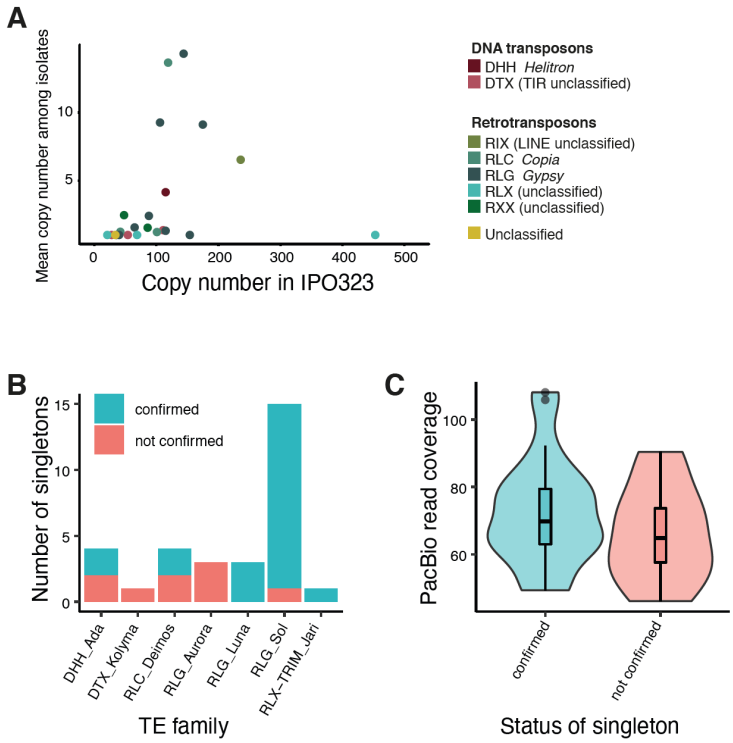


Figure 2 supplement 1. Validation of transposable element (TE) insertion predictions. (A) TEs not present in the reference genome: distribution of additional TE hits found per locus after the outlier test. Color indicates superfamilies. (B) TEs not present in the reference genome: distribution of additional TE hits found per population after the outlier test. Colors indicate populations. (C) TEs present in the reference genome: distribution of missing data per locus after the validation with spliced junction reads. Missing data indicates that the TE was not predicted with ngs_te_mapper and that there was no indication of spliced reads. The red line (=20 %) indicates the threshold for missing data. TE loci with an amount of missing data > 20 % were completely excluded from further analyses. Color indicates superfamily. (D) TEs present in the reference genome: detection of strong outlier isolates with a high number of split reads. Color indicates the population.

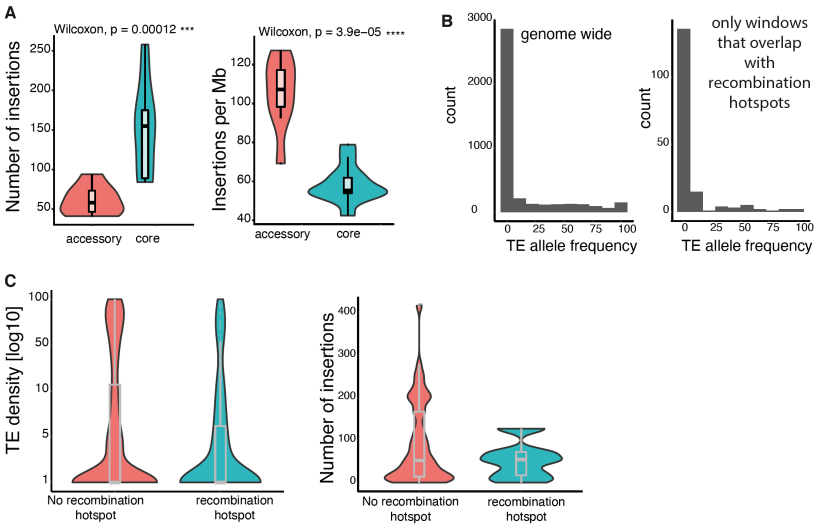


Figure 2 supplement 2. Establishment of transposable element (TE) loci with differing start and end positions in the isolates. Distribution of length of distance for start position, end position and both start and end combined after the correction.

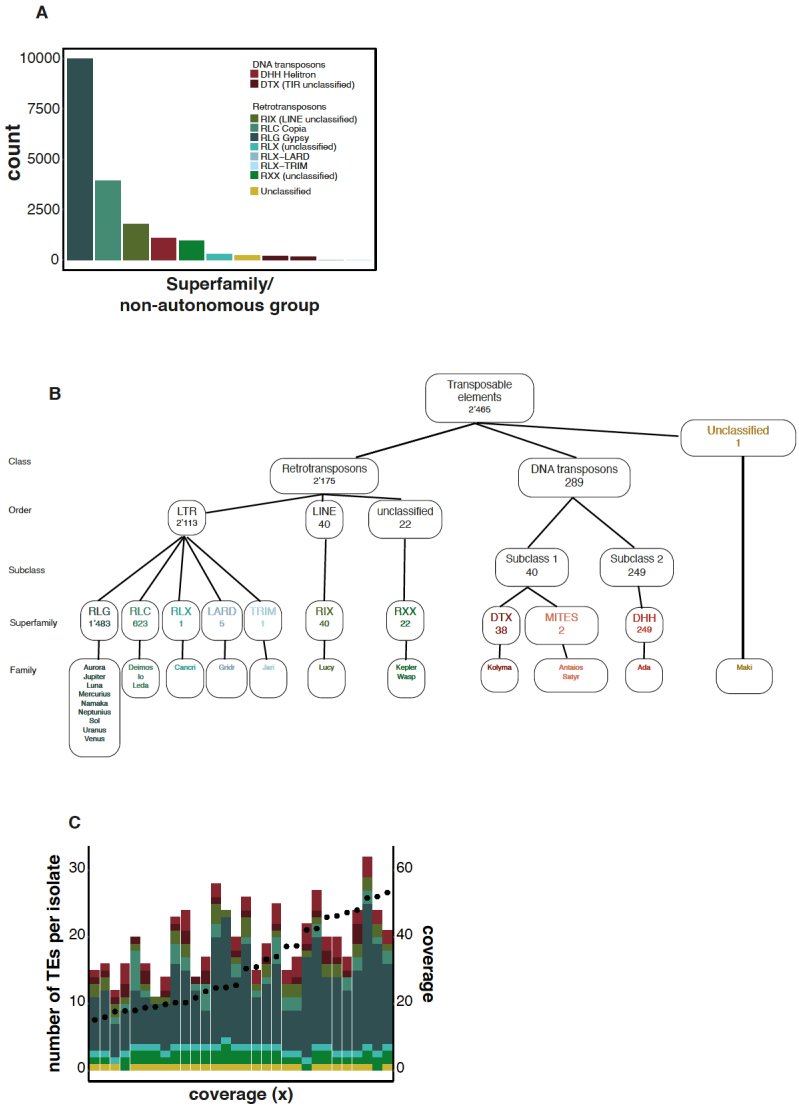


Figure 2 supplement 3. Bias for reads with a GC content lower than 30 % per population. Red lines indicate the mean.

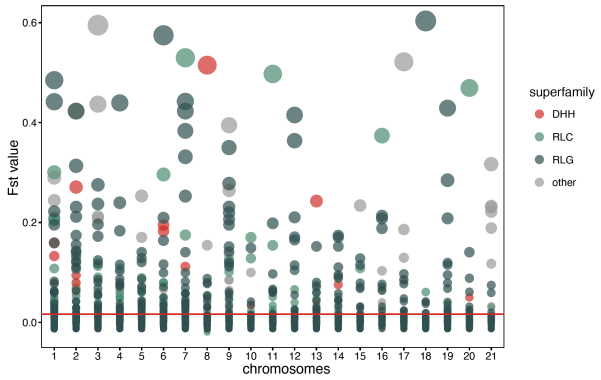


Figure 3 supplement 1. Global pairwise FST distributions shown separately for the 21 chromosomes. The red horizontal line indicates the mean FST = 0.0163. Colors are according to the three main superfamilies (RLG, RLC, DHH).

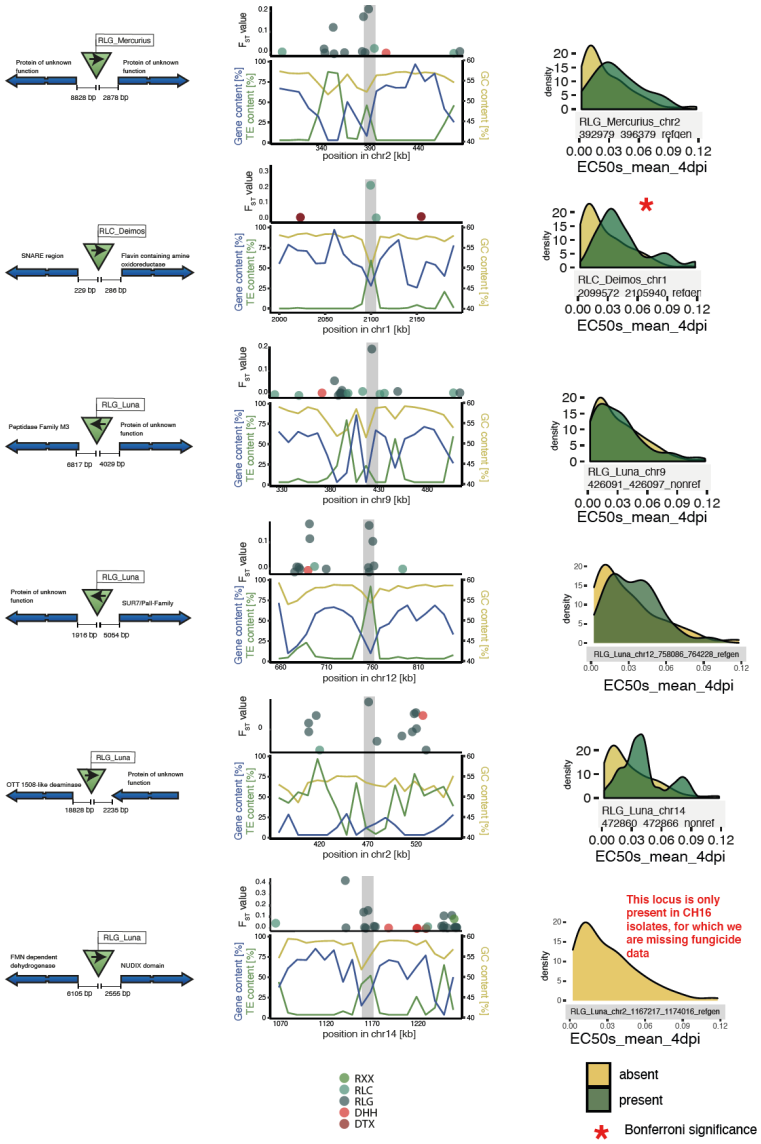


Figure 4 supplement 1. Additional top loci. Six additional candidate adaptive transposable element (TE) insertions. Each row corresponds to a candidate, with the first five being candidates detected in the North American populations and the last one in the European populations. For each candidate, the direction of the TE and the direction, function and distance of the closest two genes are indicated. The middle column indicates the location of the TE in the genomic niche, with TE content, gene content and GC content for the surrounding windows. The third column indicates resistance levels towards azole antifungals for isolates with and without the TE insertion.

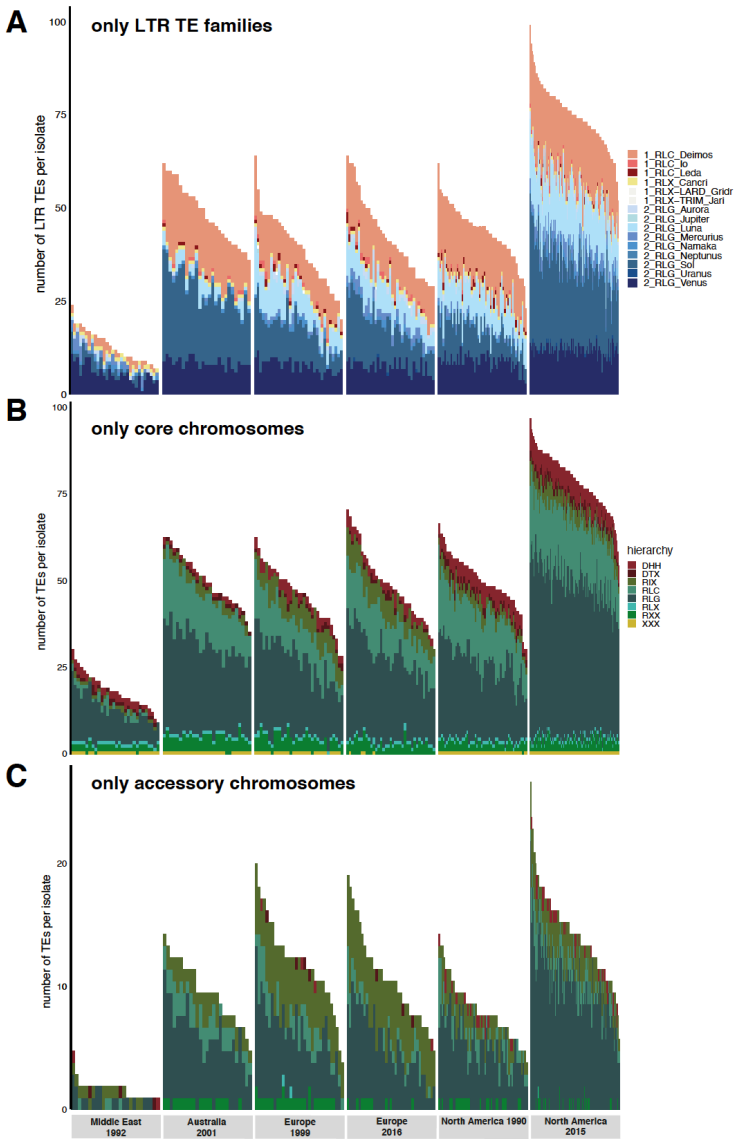


Figure 6 supplement 1. Population changes additional. Variation in transposable element (TE) content per isolate across populations. (A) Total TE copies per superfamily (colored) and per isolate only including LTR (long terminal repeat) TEs Copia and Gypsy. Color indicates the family. (B) Total TE copies per superfamily (colored) and per isolate only on the core chromosomes. (C) Total TE copies per superfamily (colored) and per isolate only on the accessory chromosomes.

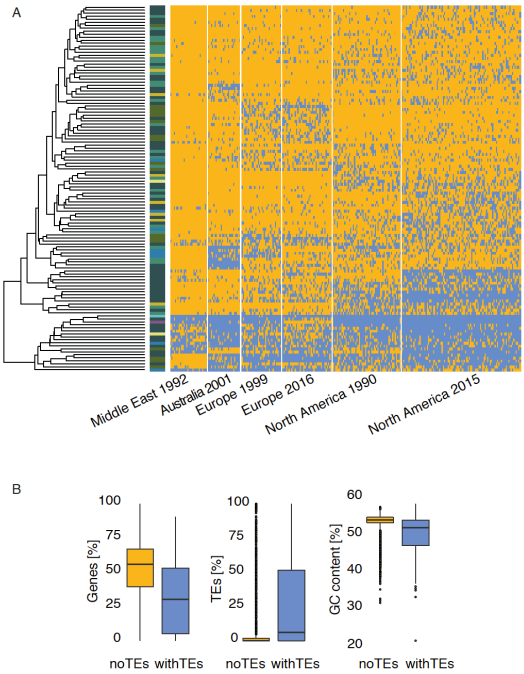


Figure 6 supplement 2. Heatmap loci. (A) Presence (blue) and absence (yellow) matrix for all transposable element (TE) loci in all isolates per population. Colors on the left side indicate the superfamily. (B) Comparison of different genomic regions with and without TE insertions in IPO323.

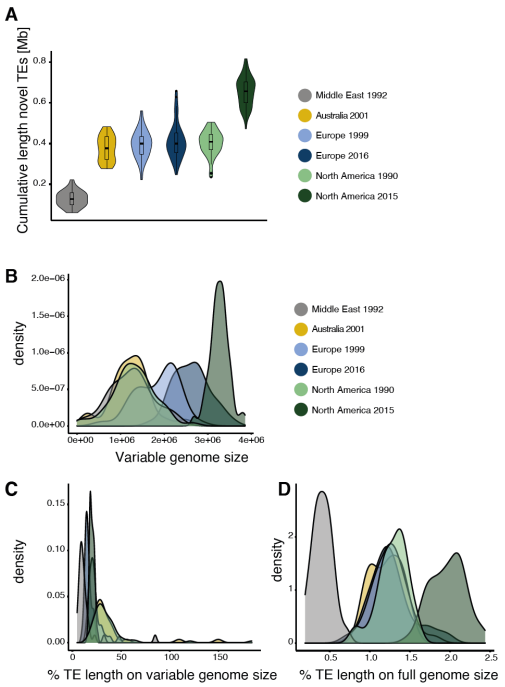


Figure supplement 1. Genome size expansion. (A) Estimated length of TE insertions per isolate and population. (B) Genome size variation per population. (C) Percentage of TEs content variation compared to the variation in genome size. (D) TE contributions to genome size variation compared to full genome size.

Chapter 2: Expansion routes of transposable elements across the genome

Ursula Oggenfuss[1], Daniel Croll[1]

[1] Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland

Author contributions: UO, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft; DC, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - original draft, Writing - review and editing

Unpublished manuscript

"There was nowhere to go but everywhere, so just keep on rolling under the stars."

Jack Kerouac (On the road)

Abstract

The activity of transposable elements (TEs) contributes to genome evolution. TEs can destabilize genome integrity but also confer adaptive changes in phenotypic traits. De-repression of silenced TEs can initiate bursts, creating a large number of novel copies in the genome. Strong purifying selection tends to quickly remove new deleterious TE insertions. A major gap in our knowledge is how transpositional activity, genomic defenses and selection jointly control TE expansion routes along the genome. To retrace the expansion of individual TE families, we analyzed a set of 19 telomere-to-telomere genomes of the fungal wheat pathogen *Zymoseptoria tritici*. Reconstructing the evolutionary history and ancestral states of individual TE families using phylogenetic methods, we show that TEs have undergone distinct activation and repression periods, leading to highly uneven copy numbers between genomes. Most TEs are clustered in gene poor compartments, indicating strong purifying selection against TE insertions nearby coding sequences. TE families with high copy numbers show low sequence divergence and strong signatures of defense mechanisms (*i.e.,* RIP). In contrast, small non-autonomous TEs (*i.e.,* MITEs) are less impacted by defense mechanisms and are often in close proximity to genes. We then reconstructed the expansion of individual TE families including copies from all genomes. TE families showed several distinct rapid expansion events with a large number of nearly identical copies being created. We find that a *Copia* element had initiated a burst from copies being inserted substantially closer to genes compared to older insertions. Overall, TE copies having most likely initiated a transposition burst are in more GC-rich regions and less affected by genomic defenses. Our work shows that specific genomic environments have a substantial impact on TE proliferation.

Introduction

Transposable elements create novel copies by duplication or relocation in the genome. Without counteraction, TEs can proliferate in the genome, leading to expanded genome sizes, increased ectopic recombination and the potential of deleterious insertions into coding and regulatory regions (Petrov *et al.*, 2003). Defense mechanisms have evolved to reversibly inactivate (*i.e.*, silence) or irreversibly mutate TEs (Daboussi & Capy, 2003; Lisch & Bennetzen, 2011). Defense against TEs include histone modifications, cytosine and chromatin methylation, small RNA based silencing or KRAB zinc finger based transcriptional silencing (Lisch, 2009; Jacobs

et al., 2014; Yang *et al.*, 2017; Schmitz *et al.*, 2019). Some TEs have the ability to regulate their own expression trough small RNA (Rebollo *et al.*, 2012). In certain ascomycete fungi, TEs are also targeted by repeat-induced point mutations (RIP) during sexual recombination (Galagan & Selker, 2004; Gladyshev & Kleckner, 2017; van Wyk *et al.*, 2021). RIP introduces a number of CpG -> TpA mutations in all copies of a duplicated sequences. RIP generally decreases the GC content and likely introduces preliminary stop codons. In the ascomycete *Neurospora crassa*, a few generations of sexual recombination are sufficient to render TE copies unfunctional with RIP mutations (Wang *et al.*, 2020b). TE expansion is additionally counterbalanced by deletion via ectopic recombination, purifying selection and genetic drift (Charlesworth & Charlesworth, 1983; Devos *et al.*, 2002; Petrov *et al.*, 2003). Defense mechanisms against TEs can lose effectiveness under stress conditions or are lost over evolutionary time scales (González *et al.*, 2008; Horváth *et al.*, 2017; Lorrain *et al.*, 2020). A loss of control over TEs can lead to de-repression and rapid amplifications (*i.e.*, bursts) of new TE copies characterized by low sequence diversity (Frantzeskakis *et al.*, 2018). Where individual TE bursts are initiated remains largely unclear. For most TEs that were recently de-repressed, or newly introduced into a species via horizontal transfer, a strong increase in copies occurs with genomic defenses being deployed to counterbalance the spread (Le Rouzic & Capy, 2005). In the absence of bursts, most TE families reach a plateau in copy numbers, which indicates a balance between insertion and deletion (Charlesworth & Charlesworth, 1983; Petrov *et al.*, 2003; Le Rouzic & Capy, 2005).

TE activity reshapes the genomic landscape with new insertions depending on the insertion rate and subsequent selection on the insertion loci (Sigman & Slotkin, 2016). Generally, strong purifying selection acts on new insertions in plant, animal and fungal genomes (Cridland *et al.*, 2013; Stuart *et al.*, 2016; Lai *et al.*, 2017; Stritt *et al.*, 2017; Oggenfuss *et al.*, 2021). The evolving genomic landscape in turn provides niches to tolerate new TE insertions (Kremer *et al.*, 2020; Stitzer *et al.*, 2021). Some TE superfamilies have preferred insertion sites, either into a specific motif or a specific region in the genome (Bridier-Nahmias *et al.*, 2015; Sultana *et al.*, 2017). Many fungal plant pathogens accumulate TEs in gene-sparse compartments with relaxed purifying selection or even in accessory chromosomes, often co-located with pathogen associated genes (Rouxel *et al.*, 2011; Raffaele & Kamoun, 2012; Van Dam *et al.*, 2017; de Freitas Pereira *et al.*, 2018; Torres *et al.*, 2020). Few TE insertions are under positive selection, especially after external stress conditions led to an increase in TE activity (González *et al.*, 2008; van't Hof *et al.*, 2016; Horváth *et al.*, 2017; Zhang *et al.*, 2019). As a consequence of

insertion site bias, selection and defense mechanisms, most TEs are unevenly distributed in the genome (Bourque *et al.*, 2018). Despite significant progress, it remains unclear to what extent TEs have adapted to specific niches of the genome or whether variation in insertion frequencies reflects purifying selection.

Active TEs face ongoing counterselection in the host genome and it has been proposed that some TEs have adapted by retaining low copy numbers (Stritt *et al.*, 2017; Blumenstiel, 2019). Few TEs show patterns of becoming co-opted, *e.g.,* as telomeric sequences or as part of regulatory networks (Dhillon *et al.*, 2014; Chuong *et al.*, 2016; Cosby *et al.*, 2019). Low divergence in TE sequences of the family indicates that transpositions occurred recently (Lerat *et al.*, 2003). Hence, phylogenetic analyses of TE copies can reconstruct the evolutionary history of the TE family. Analogous to viral birth-death models, bursts of transpositions should leave distinct marks of short internal branches in phylogenetic trees (Volz *et al.*, 2013; Blumenstiel, 2019). Copies with long terminal branches have likely been silenced TEs that remained in populations as remnants, accumulated mutations and ultimately degenerated and are no longer functional. In contrast, transposition bursts are characterized by a most recent common ancestor likely reflecting the copy initiating the expansion. Reconstructing the sequence of events leading to transposition bursts is often challenged by the difficulty in recovering all copies of a transposable element due to the incomplete nature of genome assemblies or the fact that insertions are not fixed in populations. Recent bursts of TEs were documented in large collections of rice genome resequencing datasets showing that highly active elements can be successfully identified (Lu *et al.*, 2017). Recovering full-length copies of transposition bursts remains challenging though using short-read datasets.

Many filamentous fungal plant pathogens have compact genomes amenable to TE analyses and a high degree of compartmentalization into TE dense and gene dense regions (Frantzeskakis *et al.*, 2019; Torres *et al.*, 2020). Genes present in TE rich compartments are often dispensable for survival but may be associated with virulence on the host (*i.e.*, effectors) (Raffaele & Kamoun, 2012; Croll & McDonald, 2012; Faino *et al.*, 2016; Torres *et al.*, 2020). *Zymoseptoria tritici* is an important fungal plant pathogen on wheat that co-evolved with its host (Stukenbrock *et al.*, 2007). TEs cover between 16.5 and 24 % of the genome of *Z. tritici*, and TE bursts are associated with genome size expansions, where not only new insertions would lead to a genome size increase, but also TE-derived structural variation and large-scale chromosomal rearrangements (Fouché *et al.*, 2020; Badet *et al.*, 2020; Oggenfuss *et al.*, 2021).

Insertion of different TE families in the promoter region of a major facilitator superfamily transporter led to multidrug resistance in *Z. tritici* (Omrane *et al.*, 2015, 2017; Mäe *et al.*, 2020). A DNA transposon was shown to reduce asexual spore production and is associated with chromosomal breakage points (Fouché, 2020; Wang *et al.*, 2021). TEs are associated with melanization and avirulence (Krishnan *et al.*, 2018; Meile *et al.*, 2018). Some populations seem to have lost dim2, an essential gene of the RIP machinery (Möller *et al.*, 2020; Lorrain *et al.*, 2021b). TE defenses based on cytosine methylation was also lost after a duplication event of the MgDNMT gene and subsequent mutations by RIP that rendered all copies non-functional (Dhillon *et al.*, 2010, 2014). Histone modification and small RNAs might be responsible for TE silencing in *Zymoseptoria tritici* (Schotanus *et al.*, 2015; Kettles *et al.*, 2019; Habig *et al.*, 2021). We recently defined 304 TE families based on analyses of 19 complete genomes. Most TE families are silenced, but some are de-repressed and active during stress conditions, leading to bursts of TE proliferation (Fouché *et al.*, 2020; Badet *et al.*, 2020; Oggenfuss *et al.*, 2021).

Given the challenges to comprehensively retrace insertion events, our understanding of how TEs expand in genomes remains limited. Here, we use phylogenetic reconstructions to establish how TEs have proliferated within genomes. We combine the total evidence for TE copies in 19 telomere-to-telomere genome assemblies of the same fungal pathogen species. We identify characteristics of genomic niches where TEs have recently been activated and caused bursts of proliferation.

Methods

Sequences and TE detection

We used a previously published set of 19 reference-quality genomes of *Z. tritici* assembled using PacBio sequencing (Badet *et al.*, 2020; European Nucleotide Archive BioProject PRJEB33986). The genomes cover the global genetic diversity of species and represent 14 countries on six continents (Figure 1A, Supplementary Table S1). We used a recently published improved TE annotation for the species with elements retrieved from all assembled genomes. TE annotation steps included using RepeatMasker, LTR-Finder, MITE-Tracker, SINE-Finder, Sine-Scan and extensive manual curation with WICKERsoft (Smit *et al.*; Xu & Wang, 2007; Breen *et al.*, 2010; Wenke *et al.*, 2011; Ma *et al.*, 2015; Gao *et al.*, 2016; Mao & Wang, 2017;

Crescente *et al.*, 2018; Badet *et al.*, 2020). The primary TE annotation was followed by stringent filtering steps to detect nested insertions and to join TE fragments. Simple repeats, low complexity regions and elements smaller than 100 bp were removed. TEs belonging to the same family overlapping by more than 100 bp were merged. TEs belonging to different families overlapping by more than 100 bp were considered as nested insertions. TEs belonging to the same family separated by less than 200 bp were considered as fragmented TEs and merged to one element (Badet *et al.*, 2020). We additionally annotated *Z. tritici* specific TEs using the same pipeline in high quality genomes of the sister species *Z. ardabiliae*, *Z. brevis*, *Z. pseudotritici* and *Z. passerinii* (Feurtey *et al.*, 2020).

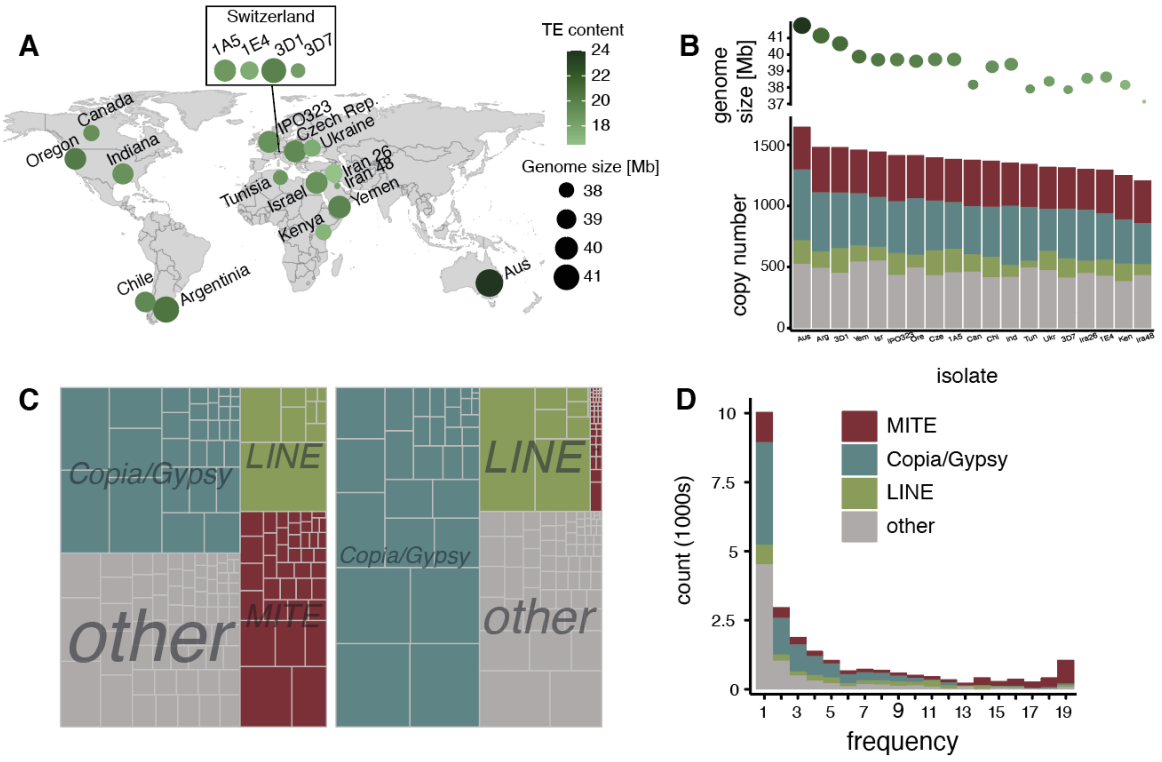


Figure 1: Transposable element (TE) distribution in 19 telomere-to-telomere genomes: (A) Origin of isolates used for TE genome analyses. Circle size indicates the genome size, the green shade indicates the TE content. (B) Genome size and TE copy number per isolate. The colors indicate MITEs (miniature inverted repeat transposable elements, small non-autonomous DNA transposons corresponding to several TE superfamilies), *Copia/Gypsy* (two superfamilies belonging to LTR (long terminal repeat) and LINE (a class of different superfamilies of non-LTR long non-interspersed elements). (C) Copy numbers of TEs (left) and total length (right) in all 19 genomes. Smaller boxes correspond to TE families. (D) Allele frequencies of TEs at orthologous insertion loci among genomes. TEs were defined as orthologous if they were located between the same set of orthologous genes.

TE multiple sequence alignment

We created multiple sequence alignments for all copies belonging to the same TE family from the 19 *Z. tritici* and four sister species genomes. We extracted all sequences of TE families with copy number ≥ 20 with the function *faidx* in samtools version 1.9 (Li *et al.*, 2009). In case

of fragmented elements, we extracted all fragments as individual copies. We reverse-complemented sequences where necessary prior to sequence alignment. To extract coding regions, we performed blastx searches against the PTREP18 database and against the non-redundant protein database from NCBI (09/2020) with diamond blast version 0.9.32.133 and selected the hit with the highest bit score with at least 200 bp length (Thomas Wicker; http://botserv2.uzh.ch/kelldata/trep-db/index.html) (Altschul *et al.*, 1990; Buchfink *et al.*, 2014). For small non-autonomous TE families lacking a coding region, we selected the whole sequence. We created multiple sequence alignments for each family with MAFFT version 7.453 and the following parameters: --thread 1 --reorder --localpair --maxiterate 1000 --nomemsave --leavegappyregion (Katoh & Standley, 2013). For four TE families with high copy numbers and large coding regions (RII_Cassini, RLG_Luna, RLG_Sol, RLC_Deimos), we slightly decreased accuracy of MAFFT, using the parameters --6merpair instead of --localpair.

TE family divergence

TE families are expected to be active during different time spans and evolve at different rates. To estimate the age of the TE families, we ran RIPCAL with --windowsize 1000 --model consensus to create an additional consensus sequence that includes all copies of a TE family (Hane & Oliver, 2008, 2010). In R version 4.0.2 we created DNAbin objects with the R package ape version 5.3 and calculated nucleotide diversity of the multiple sequence alignments for each TE family with *nuc.div* in the package pegas version 0.13 (Paradis, 2010; Paradis & Schliep, 2019; R Core Team, 2020). To compare between TE families, we divided the nucleotide diversity by the length of the corresponding TE coding region. We did a relative age estimation of TE bursts per family using RepeatMasker. To compare recent activity or bursts, we created a repeat landscape using *build Summary*, *calcDivergenceFromAlign* using Kimura divergence and *createRepeatLandscape* in RepeatMasker and visualized the results with ggplot (Kimura, 1980).

Genomic environment of TE insertions

We described the genomic characteristics of niches containing TE insertions. For 5 kb windows centered on TE insertion loci, we calculated the TE and gene content based on TE and gene annotations, respectively, using the *intersect* command in bedtools version 2.28.0. We calculated GC content with the *geecee* tool in EMBOSS version 6.6.0 (Rice *et al.*, 2000; Quinlan & Hall, 2010; Grandaubert *et al.*, 2015). We also calculated the distance to the closest

gene and TE with the *closest* command in bedtools. We used Occultercut version 1.1 with default parameters to detect isochores with low (≤ 49 %) or moderate (> 49 %) GC content (Testa *et al.*, 2016). We used TheRIPper to identify large RIP affected regions in all analyzed genomes, and calculated the overlap of TE insertions and RIP affected regions with bedtools *intersect* (van Wyk *et al.*, 2019). For the reference genome IPO323, we used available ChIP-seq information (http://ascobase.cgrb.oregonstate.edu/cgi-bin/gb2/gbrowse/ncrassa_public/) to define the chromatin structure in niches around TE insertions (Schotanus *et al.*, 2015).

Characteristics of TE insertions

Many TE sequences in the genome are fragmented due to nested insertions or partial deletions. To improve the quality of multiple sequence alignments, we selected only TE coding regions for phylogenetic analyses. For this, we trimmed the multiple sequence alignments with *extractalign* from EMBOSS, based on the position of the coding sequences, removed empty sequences with trimAl -gt 0 version 1.4.rev15 (http://trimal.cgenomics.org) and removed fragments that contained more than 50 % gap positions in the coding region (Capella-Gutiérrez *et al.*, 2009). We calculated the GC content of each TE coding region with *geecee* in EMBOSS. To quantify RIP-like mutation signatures, we extracted dinucleotide frequencies for each TE family alignment with *count* in the package seqinr (Charif & Lobry, 2007). To define locus specific TE dynamics, we looked at the closest up- and downstream fixed orthologous genes from the pangenome annotation with *closest* in bedtools (Badet *et al.*, 2020). We defined TE insertions belonging to the same TE family and being located between the same fixed orthologous genes as orthologous TE groups. Visualizations were made with ggplot (Wickham, 2016).

Maximum likelihood trees

We estimated maximum likelihood trees for all TE families with indications for recent activity and bursts in the species. We extracted conserved blocks of the coding region with Gblocks version 0.91b, using the following parameters: -t=d -b3=10 -b4=5 -b5=a -b0=5 (Castresana, 2000). For each TE family, we included two sequences retrieved from the same TE in sister species genomes to root trees. We estimated maximum likelihood trees with RAxML version 8.2 (Stamatakis, 2014). For this, we generated 20 ML trees with each a different starting tree and extracted the starting tree with best likelihood with the following parameters: raxmlHPC-PTHREADS-SSE3 -T 4 -m GTRGAMMA -p 12345 -# 10 --print-identical-sequences. We

performed a bootstrap analysis to obtain branch support values with the following parameters: raxmlHPC-PTHREADS-SSE3 -T 4 -m GTRGAMMA -p 12345 -b 12345 -# 50 --print-identical-sequences. Finally, we draw bipartitions on the best ML tree with the following parameters: raxmlHPC-PTHREADS-SSE3 -T 4 -m GTRGAMMA -p 12345 -f b --print-identical-sequences.

Ancestral state reconstruction

We performed ancestral state reconstructions for each TE family including characteristics of the TE sequences or the niche of the TE insertion. We imported the trees into R using *read.tree* from the package treeio version 1.10.0 (Wang *et al.*, 2020a). We rooted the trees with *root* in package ape, using sister species sequences as an outgroup. We converted trees objects to tibble with *as_tibble* from the package tibble version 3.0.1 in tidyverse version 1.3.0 and added metadata using dplyr version 0.8.5 in tidyverse (Wickham *et al.*, 2019, 2020; Müller & Wickham, 2020). We converted the tibble objects back to tree formats with *as.treedata* in treeio (Wickham *et al.*, 2019). Using *fastAnc* and *contMap* from package phytools version 0.7-47, we performed ancestral state reconstruction for characteristics of the following continuous traits: gene density, TE density and GC content of 2.5 kb windows surrounding the TEs, closest gene, GC and RIP-like mutations per bp (Revell, 2012). To estimate ancestral states for binary characteristics (GC rich *vs.* poor, accessory *vs.* core chromosomes), we used *make.simmap* from the package phytools with an equal rates model and 100 simulations. We visualized the trees with *ggtree* version 2.0.1 (Yu *et al.*, 2017). To retrieve clades representing recent bursts, we created polytomy trees from binary trees, using the command *CollapseNode* from TreeTools version 1.4.4 in R at branch lengths smaller than 1.1e-05 (Smith, 2019). For each burst clade, we defined the parental branch and an outgroup of the clade with *offspring* in the *treeio* package as the ancestral branch. We compared characteristics of ancestral branches of bursts with the distribution of the characteristics in all elements outside of bursts. We performed associating mapping for shared characteristics along the phylogenetic tree using treeWAS (Collins & Didelot, 2018).

Results

TE diversity in the pangenome

We analyzed TE copies in 19 completely assembled genomes to comprehensively map each TE family found in *Z. tritici* (Figure 1A). Both genome size and TE content vary considerably, and are positively correlated. The Australian isolate has the highest TE content (24 %) and largest genome size (41.76 Mb) and an Iranian isolate has the lowest TE content (18.1 %) and smallest genome size (37.13 Mb) (Figure 1B; published Supplementary Table S1 in Badet et al, 2020). We focused on TE families with at least 20 copies that are making up a total of 24,520 copies across all analyzed genomes. Half of the copies belong to DNA transposons (*n* = 104 families) and half to retrotransposons (*n* = 59 families; Supplementary Table S2) without meaningful differences among genomes (Figure 1C; Supplementary Figure S1). We analyzed allele frequencies at TE insertion sites. Most TE insertions are singletons or at low frequency with few TE insertions being fixed among genomes (*n* = 122; Figure 1D). Fixed TE insertions are predominantly MITEs.

TE insertion niches are of low gene content

We analyzed niches of TE insertions by retrieving key features of 5 kb windows centered around insertions (Figure 2A). For the reference genome IPO323, we found no accumulation of TE insertions in regions with euchromatin marks, and only a subset of TEs overlapping with obligate heterochromatin marks (Figure 2A). Across all 19 genomes, we found that most TEs are located on core chromosomes, but TEs are more densely packed on accessory chromosomes (Figure 2B). The majority of TE copies are located in compartments with a GC content below 50 %, with the exception of MITEs (Figure 2B). Only a small subset of TE copies is located in regions overlapping a gene or a subtelomeric region (Figure 2B). We found no relationship between TE insertions and large RIP affected regions. Most *Copia/Gypsy* and LINE are located within large RIP affected regions and most MITEs outside of large RIP affected regions (Figure 2B).

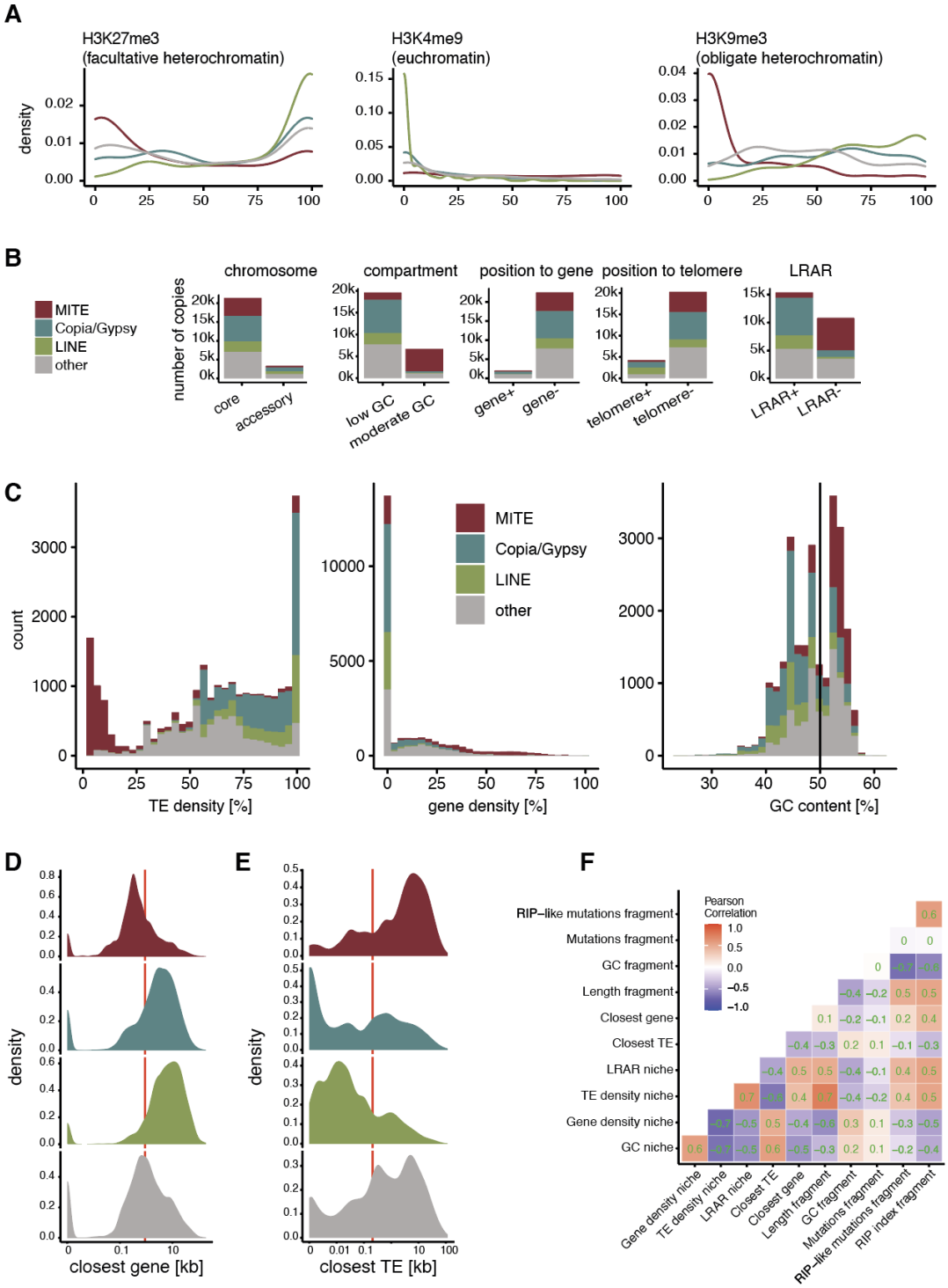


Figure 2: Characteristics of TE insertion niches in the genome: (A) Proportional overlap of H3K27me2, H3K4me9 and H3K9me3 histone methylation marks with TE insertion niches in the reference genome IPO323. Colors indicate the group of TE. (B) TE insertion sites between core and accessory chromosomes, TE insertions into niches with a moderate (≥ 50%) or low (< 50%) GC content, TE insertions into regions annotated as genes, TE insertions into subtelomeric region and TE insertions into large RIP affected regions. (C) Overlap of TE content, gene content and GC content with TE insertion niches. (D) Distribution of the distance to the closest gene in MITEs, *Copia/Gypsy*, LINE and other TEs. The red line indicates the mean distance. (E) Distances to the next TE MITEs, *Copia/Gypsy*, LINE and other TEs. The red line indicates the mean distance. (F) Spearman correlation matrix of 11 characteristics of TE insertion niches and TE copy characteristics. Dark red indicates strong positive correlation, dark blue indicates strong negative correlation of two characteristics.

We found more than one third of TEs inserted into niches with more than 80 % TE content. In contrast, MITEs are preferentially in TE poor regions (Figure 2C). GC content in TE insertion niches varies between 25 and 60 %. We found more than one third of TE insertions being 1-10 kb away from the next gene, with MITEs being on average closer (Figure 2D). TE insertions were often close to LINEs and *Copia/Gypsy* copies (902 and 2431 bp, respectively). MITEs generally were at a large distance from the next TE (8,037 bp; Figure 2E). Overall, niche TE density is negatively correlated with gene density and GC content (Figure 2F). Longer TE copies tend to be located in already TE rich regions.

Recent activity of high-copy TE families

Recently active TE families typically carry a high number of weakly differentiated TE copies in the genome. We first filtered for a subset of TE families with more than 100 copies in all 19 genomes combined. The 61 retained TE families include predominantly MITEs (n = 12) as well as *Gypsy* and *Copia* copies (n = 11 and 5, respectively) (Figure 3A). We find that high-copy TE families tend to have also more variable copy numbers among 19 genomes, indicating ongoing activity of the individual families (Figure 3B). The GC content of high copy TE families is generally near the genome-wide GC, with the exception of MITEs with an overall higher GC content around 52%, and an RLC_Deimos family where most copies have a GC content below 40 % (Figure 3C). Full length copies range from 218 to 13,907 bp, with the shorter copies belonging to the non-autonomous MITEs lacking coding regions (Figure 3D). To estimate the relative age of TE copy number increases, we calculated the nucleotide diversity for each TE family. TE families with lower copy number tend to have a higher nucleotide diversity. TE families with high copy numbers have very low nucleotide diversity consistent with recent proliferation in the genome. MITEs tend to have higher nucleotide diversity at similar copy numbers compared to other TE families (Figure 3E, H). MITEs are also less affected by genomic defenses related mutations (Figure 3F, H). Terminal branch lengths of individual TE copies are a further indication of the age of recent transposition. Copies of MITEs tend to have overall short terminal branch lengths compared to other TEs (Figure 3G). The short length of MITEs might constrain the potential to accumulate mutations compared to longer TEs.

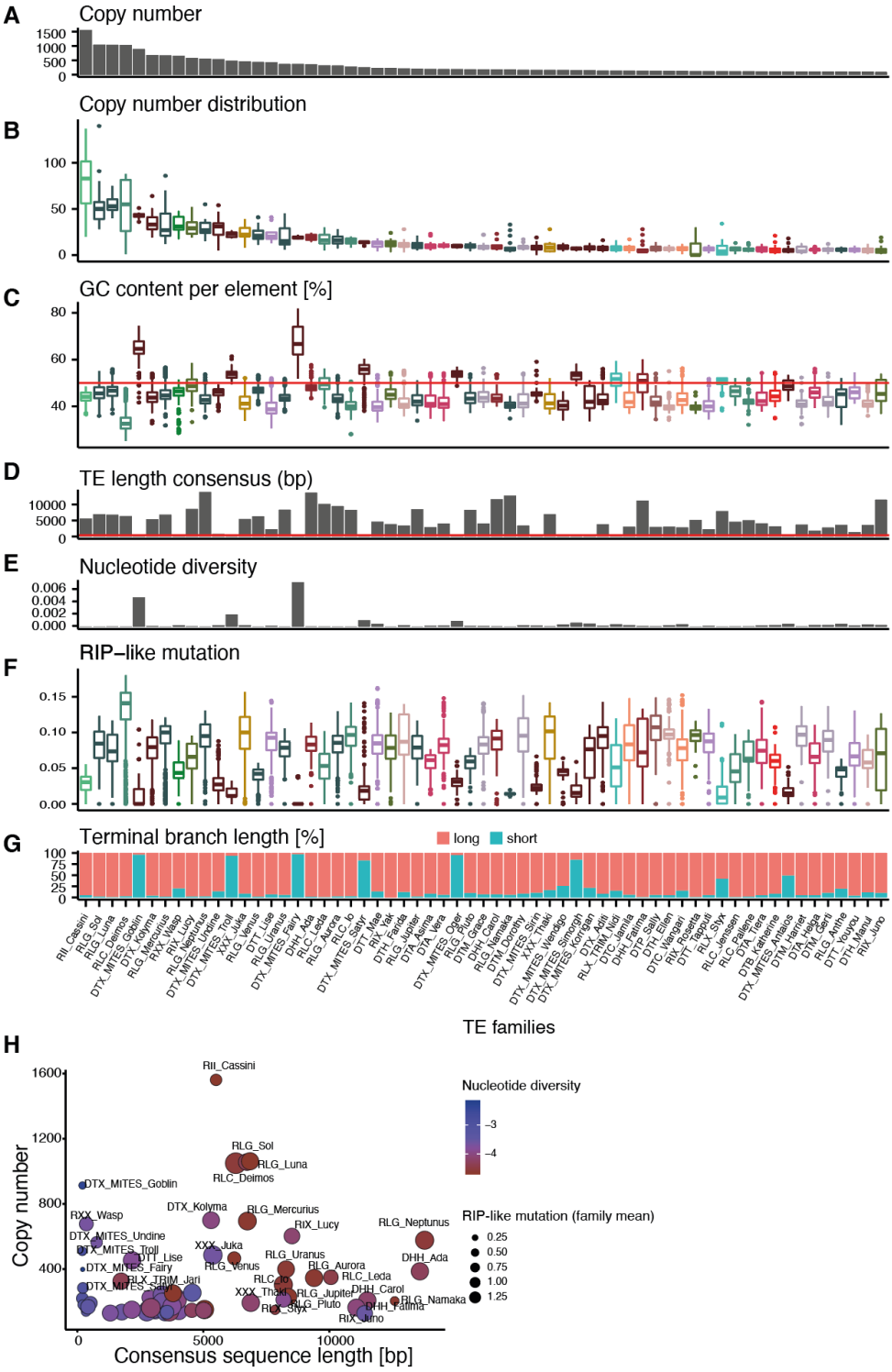


Figure 3: Characteristics of high-copy TE families: This figure is ordered from highest total copy numbers to lowest copy numbers (right) in all 19 analyzed genomes combined. (A) Total copy numbers. (B) Copy numbers per TE family. (C) GC content distribution per TE family. (D) Length of the consensus sequence corresponding to the full-length consensus sequence excluding nested TEs or partial deletions. (E) Nucleotide diversity of the TE family. (F) Number of RIP-like mutation (CpA<->TpA/TpG<->TpA) per TE copy, corrected for the length of the TE. (G) Long (> 0.00001; red) and short (≤0.00001; blue) terminal branch lengths of individual copies identifying two classes of divergence times. (H) Correlation between copy numbers and consensus sequence lengths for TE families. Circle size corresponds to the mean number of RIP-like mutations and the color indicates the nucleotide diversity.

Consistent with this, many MITEs show long internal branch lengths between distinct clades characterizing independent bursts (Supplementary Figure S2). Overall, TE families with high copy numbers and long consensus sequences show a lower nucleotide diversity yet a higher impact of RIP-like mutations (Figure 3H). Our findings show that recent activity typically characterized by high copy numbers and low nucleotide diversity is in a complex interplay with genomic defenses and the length of TEs.

Expansion routes assessed through phylogenetic relationships

We investigated the timing of TE bursts by calculating genetic distances among copies (Figure 4A). Most TE families show their highest activity in a similar, recent range. We found two TE families with ongoing activity (Styx and Thrym). For the high copy TE families, the RII_Cassini was most recently active. RLG_Luna, RLG_Sol, RIX_Lucy and RLC_Deimos had undergone earlier bursts with both RLC_Deimos and RLG_Luna showing indications for an even earlier burst of proliferation. To reconstruct TE expansion routes in the genome, we created phylogenetic trees, rooted using copies found in genomes of the *Zymoseptoria* sister species. For each inserted TE, we analyzed whether the characteristics of the genomic niche and TE sequences themselves have evolved from the ancestor sequence it has most likely derived from. For this, we used ancestral state reconstruction to identify features of all the most recent branching points leading to individual TE copies. We found increases from ancestor sequences to individual TEs in the GC and gene content of the genomic niche as well as the GC content of the copy (Figure 4B). While most TEs inserted on a different chromosome compared to their ancestor, new insertions typically remained on core chromosomes (65.4 %) or switched from accessory to core chromosomes (21.5 %). We found that more than half of the insertions remained in isochores of low GC content (58.4 %) or jumped from moderate to low GC content (20.5 %). Additionally, more than half of all TE insertions remained in large RIP affected regions or jumped into such regions (20.7 %).

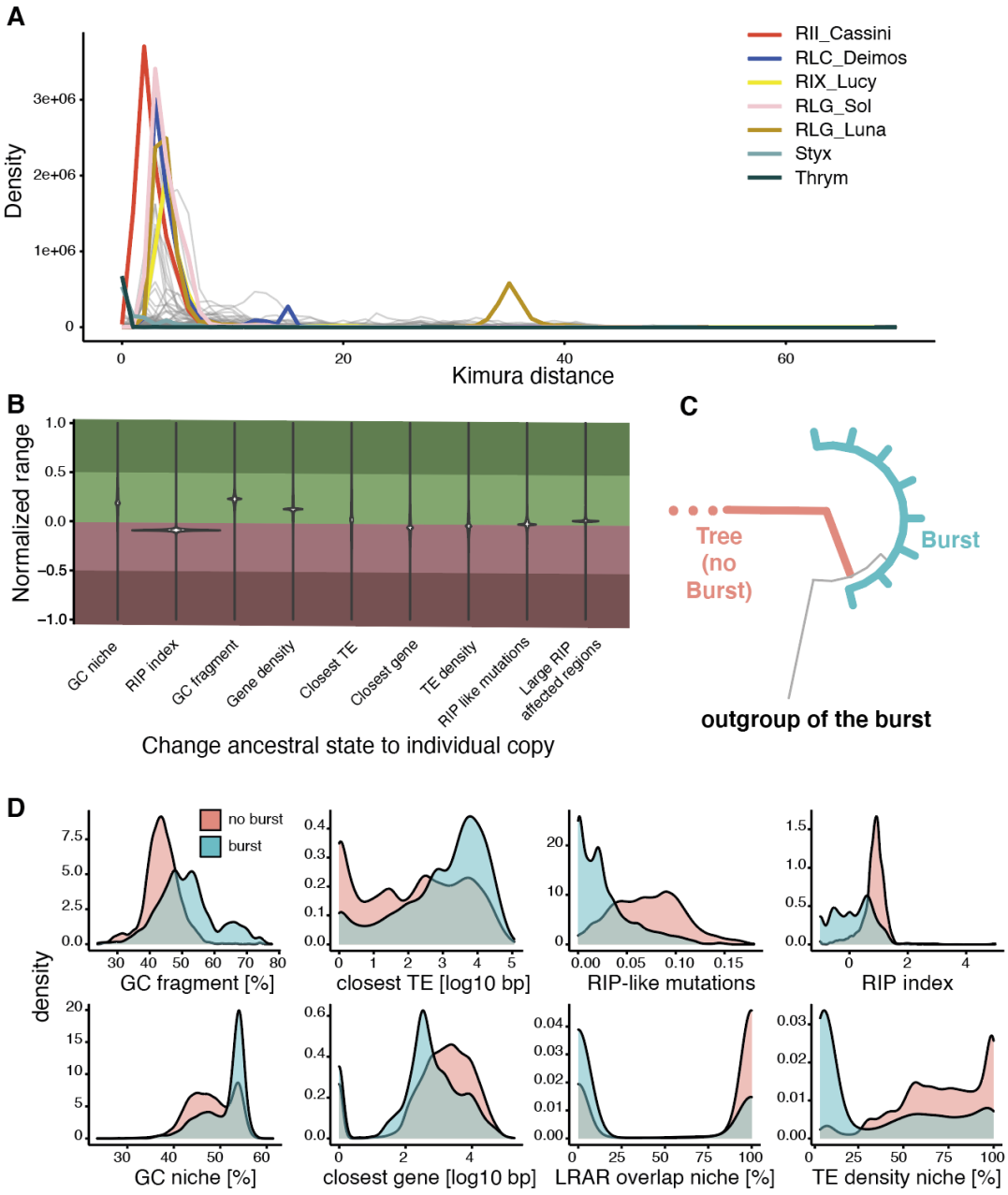


Figure 4: Origin and features of TE transposition bursts: (A) Repeat landscape of the TE families with the highest copy numbers. Colors indicate the highest copy numbers (RII_Cassini, RLC_Deimos, RLG_Sol, RLG_Luna), TEs with multiple bursts (RLC_Deimos, RIX_Lucy and RLG_Luna) or very recent burst (Styx, Thrym). (B) Normalized range of characteristics of TE copies and their genomic niche compared between ancestral states and derived copies. Green indicates an increase and red a decrease in individual features. (C) Scheme of the definition of burst and burst outgroups based on phylogenetic trees of TE families. Green indicates the copies of a burst with low terminal branch lengths and the red outgroup indicates the closest related sister branch of a burst. (D) Distribution of niche and TE copy characteristics of copies belonging to a burst clade (red) compared to all other copies (blue).

We identified individual bursts within TE families as groups of almost identical sequences by retrieving clades of highly similar sequences distinct from the rest of the tree (Figure 4C). We defined the outgroup of individual bursts as the most likely parental copy that preceded the burst. Overall, half of TE families experienced at least one recent burst and 10% (n = 32) of all TE families show several bursts. Copies of individual bursts were often found only in a subset

of the analyzed genomes of the species consistent with the bursts being very recent and local. Bursts observed in MITE families often include a large proportion of all copies suggesting recent expansion events. To identify general properties of TE bursts, we compared copies included in a burst with all other copies of a TE (Figure 4D). Burst copies generally have a higher GC content, less RIP-like mutations, are closer to genes and more distant to other TEs compared to non-burst copies (Figure 4D). We found also that burst copies tend to be located in genomic niches with lower TE density, overlap less with large RIP affected regions and have a higher GC content.

Niche characteristics of massively expanded TE families

To identify niche insertion preferences in the evolutionary history of massively expanded TE families, we focused on five TE families with high copy numbers and evidence for recent bursts of activity (LINE/*I* RII_Cassini, LTR/*Gypsy* Luna and Sol, the LTR/*Copia* Deimos, and the DTX_MITE_Goblin). During the recent burst of RLC_Deimos, new copies were mostly inserted outside of large RIP affected regions and closer to genes. These copies are also of higher GC content and show only few RIP-like mutations (Figure 5). Similarly, the recent burst of RII_Cassini has created copies with a higher GC content and lower numbers of RIP-like mutations (Figure 5). In contrast, RLG_Sol and RLG_Luna show no evidence of recent bursts of activity (Figure 5). DTX_MITEs_Goblin showed a general high copy number, with similar copy numbers per isolates indicating that these are old insertions shared among the genomes.

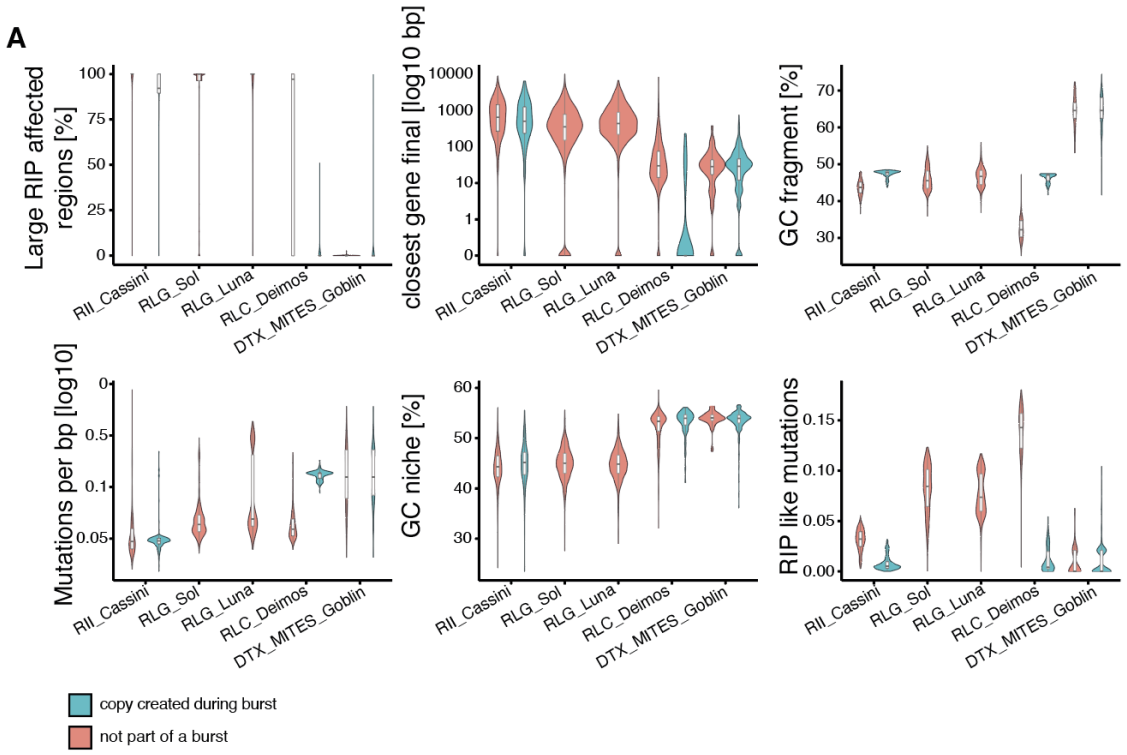


Figure 5: High copy number TE families and characteristics of burst copies: Comparison of copies in bursts (red) and all other copies (blue) for the five TE families with the highest copy numbers. The TE families RLG_Luna and RLG_Sol do not have any copies in burst clades.

RII_Cassini shows evidence for six individual bursts, two of which are likely very recent as all copies of an individual burst were found in a single genome (isolates from Australia and Canada, respectively). Most copies show a low to moderate number of RIP-like mutations, and moderate GC content, with the copies generated during bursts having a higher GC content and lower number of RIP-like mutations compared all other copies (Figure 6). Most copies were inserted into niches with strong signatures of RIP, moderate GC content, and far from coding sequences. The RLC_Deimos has undergone a single large burst with copies carrying nearly no RIP-like mutations (Figure 7) in contrast to a higher number of RIP-like mutations in all other copies.

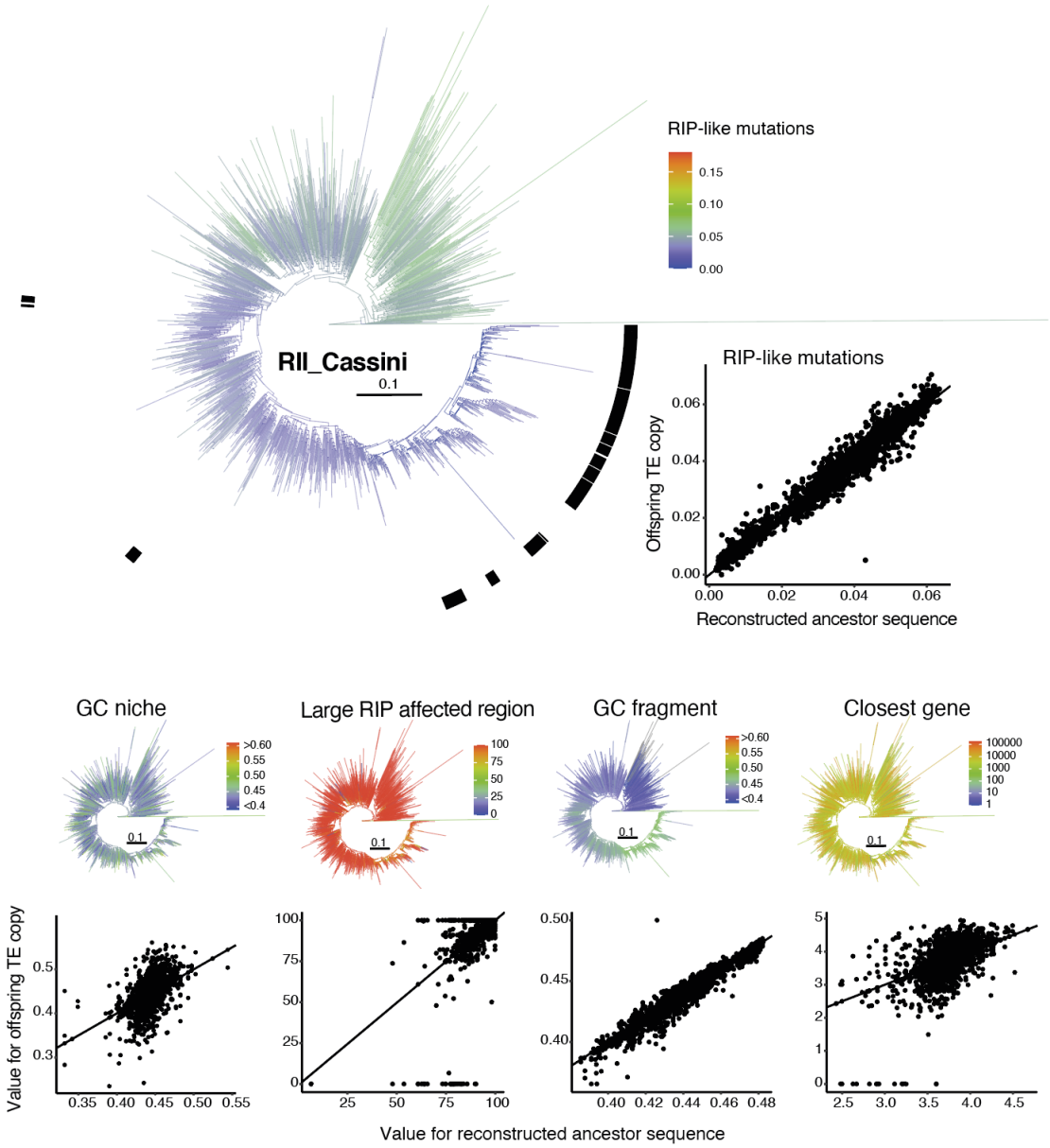


Figure 6: Phylogenetic tree of the TE family RII_Cassini: (A) Phylogenetic tree. Colors indicate the number of RIP-like mutations. The black bar marks the different burst clades. The dot plot shows the changes RIP like mutations in from the ancestor to offspring for all internal and terminal branches from the ancestral state reconstruction. (B-E) Phylogenetic trees and ancestor-offspring changes for (B) the GC content of the niche, (C) the overlap of the niche with large RIP affected regions (LRAR), (D) the GC content of the copy and (E) the distance to the closest gene.

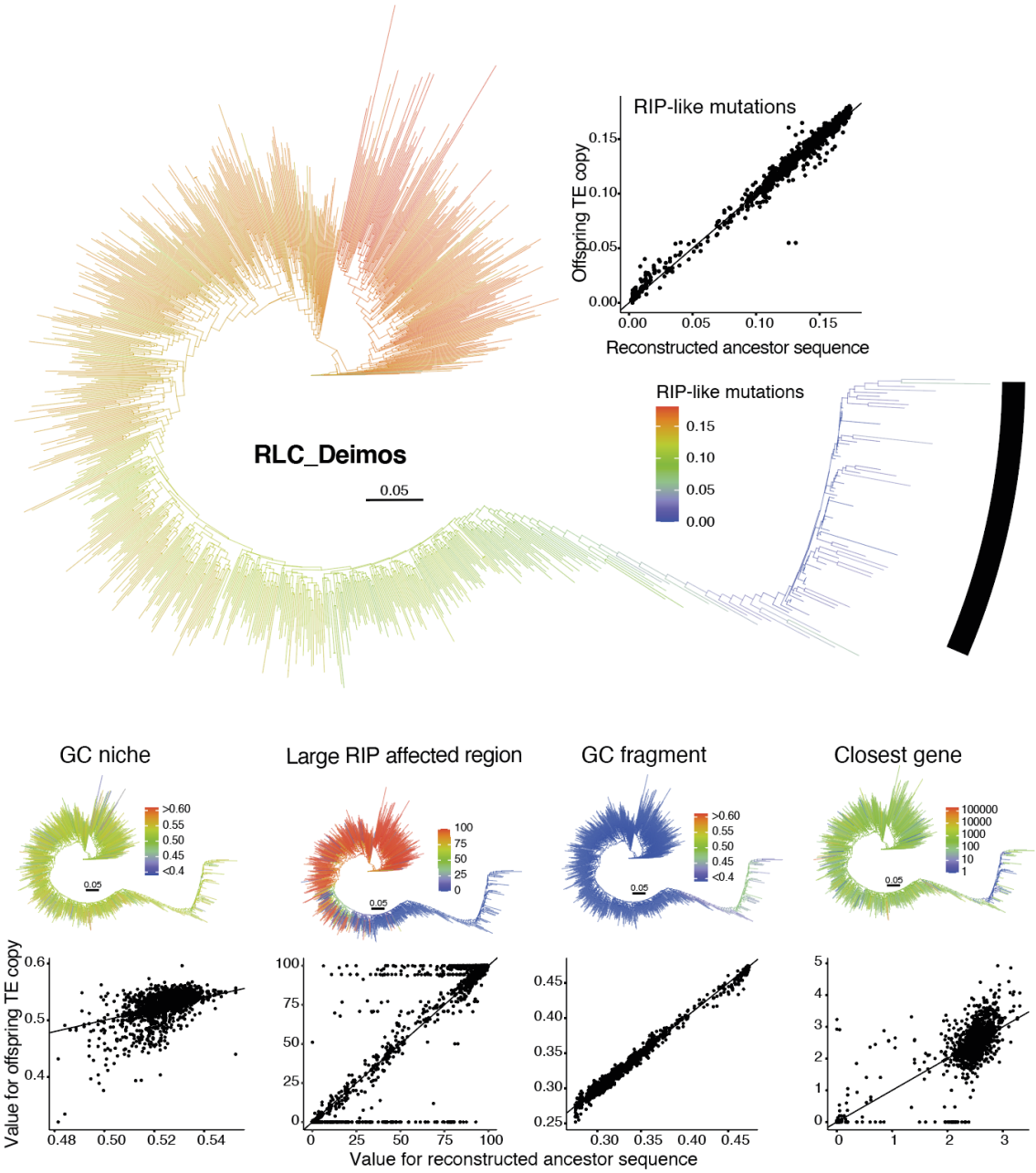


Figure 7: Phylogenetic tree of the TE family RLC_Deimos: (A) Phylogenetic tree with colors indicating the number of RIP-like mutations. The black bar marks the different burst clades. The dot plot shows the changes in RIP-like mutations from the ancestor to offspring for all internal and terminal branches from the ancestral state reconstruction. (B-E) Phylogenetic trees and ancestor-offspring changes for (B) the GC content of the niche, (C) the overlap of the niche with large RIP affected regions, (D) the GC content of the copy and (E) the distance to the closest gene.

Burst copies show an increase in GC content and are inserted into regions devoid of RIP-like mutations compared to all other copies. Interestingly, one of the likely ancestral copies leading to the burst of RLC_Deimos copies is inserted into a gene found in a single genome only and encodes an alpha/beta hydrolase. The expansion of DTX_MITEs_Goblin is characterized by a high number of bursts. Copies from individual bursts are typically shared among all isolates. Despite high nucleotide diversity, RIP-like mutations are very low. Taken together, the

DTX_MITEs_Goblin family might consist of older insertions shared between genomes, that are not affected by RIP. Across major TEs showing recent burst activity, the genomic niche characteristics between TE families are largely depending on the presence of recent bursts.

Discussion

TE activity is an important disruptor of genome integrity and a source of potentially beneficial mutations. The deleterious impact of TEs is strongly influenced by the length of inserted sequences and site preferences, in particular during bursts. Our analyses of 19 complete genomes of the same species show that TEs were highly active in the recent evolutionary history of the species. A substantial number of TEs have produced one or multiple bursts of proliferation with distinct insertion site characteristics from copies unrelated to bursts. Genomic defense mechanisms were only partially effective against proliferating TEs. Our analyses suggest that a key trigger point for initiating new bursts is the successful insertion close to coding sequences.

TE families undergo phases of inactivation, bursts and diversification

We identified the emergence and expansion of numerous clades within TE families consistent with rapid copy number expansions characteristic of a burst. TE families often include large numbers of inactive copies accumulating mutations that are unlikely to cause new insertions. Consistent with the dynamic nature of TE families, many TEs of *Z. tritici* are not detectable in the genomes of sister species. Searching for more distant homologs to *Z. tritici* TEs might reveal common ancestors though as observed in other species (Wicker *et al.*, 2007; Bleykasten-Grosshans *et al.*, 2021). *Z. tritici* MITEs families are most often absent in the sister species. This is consistent with a rapid divergence of these non-autonomous elements, as they are not carrying coding sequences. A small subset of all TE families seems to be driving most recent insertions. Such new insertions are expected to have persisted despite potential purifying selection against new insertions, mutations introduced by genomics defenses (*i.e.,* RIP) or silencing. We identified such rapid proliferations of new copies (*i.e.,* bursts) to be highly uneven among TE families. While most families seem largely inactive, we identified a small subset of families having undergone repeated bursts. Some TE families including RLC_Deimos were activated in several waves starting new branches or ”subfamilies” from

distinct TE copies in the genome. Consistent with the independent bursts, the family has also an overall elevated nucleotide diversity. What triggers the activation of TEs remains largely unexplored. *Z. tritici* has undergone substantial colonization bottlenecks during its emergence as a wheat pathogen and global migration (Zhan *et al.*, 2003; Stukenbrock *et al.*, 2007). The species has also experienced significant environmental stress factors including colder climates and, more recently, fungicide applications (Croll & McDonald, 2017). Stress can de-repress silenced TEs and therefore underpin the activation of TEs (Ito et al., 2011; Cavrak et al., 2014). *Z. tritici* specifically de-represses TEs upon contact with the plant host caused by plant defenses being activated (Fouché *et al.*, 2020). How TE bursts in the evolutionary history of the species are associated with changes in the lifestyle or demography could provide important insights into the biology governing TE activity.

The impact of defense mechanisms on TE proliferation

As ongoing activity should trigger defense mechanisms, we expected recently active TE families to be highly affected by RIP. However, our results suggest a low impact of RIP on more recent TE copies. Weak RIP signatures were predominantly found in MITEs and such short elements are most likely to escape detection by the RIP machinery as seen in another species (Pereira *et al.*, 2021). Longer elements show stronger impacts of RIP using both GC content and RIP-like mutations as proxies. We found evidence for RIP mostly in older copies, while copies generated during a burst show nearly no RIP signatures. Escaping effects of RIP may be a prerequisite for the initiation of a burst, hence the strong association of age and RIP mutations driven by this necessity. RIP has mostly been studied in the ascomycete *N. crassa*, where RIP introduces a large number of mutations after just one generation of sexual recombination (Wang *et al.*, 2020b). The life cycle of *Z. tritici* is thought to consist of several cycles of asexual reproduction during the growing season, and only one round of sexual reproduction at the end of the season (Chen & McDonald, 1996). Hence, TEs may proliferate despite RIP for short periods but the ubiquitous sexual reproduction should provide ample opportunity for RIP to act on copies of any recent burst. Evidence for RIP mutations are widespread in the genome of *Z. tritici* and the machinery for RIP appears conserved (van Wyk *et al.*, 2021). However, species-wide analyses have suggested that newly established populations have lost a functional RIP machinery (Lorrain *et al.*, 2020). As RIP is a key factor in reducing GC content of fungal genomes, loss of the RIP machinery would therefore maintain repetitive sequences at a GC content or even lead to an increase through for example GC biased gene conversion (Duret & Galtier, 2009). To what extent the loss of RIP and counter-acting

factors have shaped the distribution of GC content among TE copies remains unknown. A loss of RIP efficiency would be consistent with the loss of RIP-like mutations in nearly all recently expanded TE families. Our data on recent TE proliferation strongly supports the idea that genomic defenses have been recently weakened in the species.

The insertion niche has a strong impact on the fate of a TE copy

Young TE copies show distinct niche associations in the genome compared to older copies. Older copies tend to be located in regions with high TE content, new copies from recent bursts are closer to genes. If the association of different TE copies with genomic niches is driven by preferred insertion sites this pattern is most likely confounded by the action of purifying selection. The strength of purifying selection against TEs depends on the fitness penalties incurred by the insertion. As TEs can disrupt encoded proteins or change expression profiles of neighboring genes, purifying selection is most likely strongest for TEs inserting into gene rich regions. In contrast, TEs inserting into already TE-rich regions by disrupting of existing TE sequences (*i.e.*, nested insertions) have likely only a minor impact on fitness. In some fungal plant pathogens including *Z. tritici*, such nested insertions led to the compartmentalization of niches with high TE density and niches mostly composed of genes (Plissonneau *et al.*, 2016). Repeated insertions of TEs into such TE-rich regions likely exacerbates the compartmentalization of the genome. Selection might also have an impact on the effectiveness of defenses against TEs. Silencing or hypermutation close to genes can disrupt the functionality of genes, hence the efficiency of genomic defenses may be weakened by selection to reduce pleiotropic effects. Hence, it is conceivable that otherwise silenced TEs can create new copies and lead to a burst of inserted sufficiently close to genes.

Beyond having possibly reduced genomic defenses, the propensity of bursts being initiated by TEs inserting unusually close to genes could also be related to beneficial impacts of the TE insertion itself. We found that copies at the start of bursts tend to have higher allele frequency within the species (*i.e.,* present in most analyzed genomes). Our reference genome dataset is too small to analyze selection acting on parental copies initiating bursts. It is conceivable though that beneficial effects of an individual TE insertion are linked to triggers of TE bursts. Such a multi-level model of selection was recently suggested to represent a Devil's Bargain in plant pathogens (Fouché *et al.*, 2021). Reconstructing fitness consequences of individual TE insertions along the expansion history of TEs will allow to test hypotheses about proximate drivers of TE expansions over short evolutionary time periods.

Acknowledgements

DC is supported by the Swiss National Science (grants 31003A_173265) and the Fondation Pierre Mercier pour la Science.

Data availability

Supplementary Data available on https://zenodo.org/record/6029566.

Literature chapter 2

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. Journal of Molecular Biology 215: 403–410.

Badet T, Oggenfuss U, Abraham L, McDonald BA, Croll D. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biology 18: 12.

Bleykasten-Grosshans C, Fabrizio R, Friedrich A, Schacherer J. 2021. Species-Wide Transposable Element Repertoires Retrace the Evolutionary History of the Saccharomyces cerevisiae Host . Molecular Biology and Evolution 38: 4334–4345.

Blumenstiel JP. 2019. Birth, School, Work, Death, and Resurrection: The Life Stages and Dynamics of Transposable Element Proliferation. Genes 10.

Bourque G, Burns KH, Gehring M, Gorbunova V, Seluanov A, Hammell M, Imbeault M, Izsvák Z, Levin HL, Macfarlan TS, et al. 2018. Ten things you should know about transposable elements. Genome Biology 19: 199.

Breen J, Li D, Dunn DS, Békés F, Kong X, Zhang J, Jia J, Wicker T, Mago R, Ma W, et al. 2010. Wheat beta-expansin (EXPB11) genes: Identification of the expressed gene on chromosome 3BS carrying a pollen allergen domain. BMC Plant Biology 10.

Bridier-Nahmias A, Tchalikian-Cosson A, Baller JA, Menouni R, Fayol H, Flores A, Saib A, Werner M, Voytas DF, Lesage P. 2015. An RNA polymerase III subunit determines sites of retrotransposon integration. Science 348: 585–588.

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12: 59–60.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Castresana J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular Biology and Evolution 17: 540–552.

Cavrak V V., Lettner N, Jamge S, Kosarewicz A, Bayer LM, Mittelsten Scheid O. 2014. How a Retrotransposon Exploits the Plant's Heat Stress Response for Its Activation. PLoS Genetics 10.

Charif D, Lobry JR. 2007. SeqinR 1.0-2: A Contributed Package to the R Project for Statistical Computing Devoted to Biological Sequences Retrieval and Analysis. : 207–232.

Charlesworth B, Charlesworth D. 1983. The population dynamics of transposable elements. Genetical Research 42: 1–27.

Chen RS, McDonald BA. 1996. Sexual reproduction plays a major role in the genetic structure of populations of the fungus Mycosphaerella graminicola. Genetics 142: 1119–1127.

Chuong EB, Elde NC, Feschotte C. 2016. Regulatory evolution of innate immunity through co-option of endogenous retroviruses. Science 351: 1083–1087.

Collins C, Didelot X. 2018. A phylogenetic method to perform genome-wide association studies in microbes that accounts for population structure and recombination (AC McHardy, Ed.). PLOS Computational Biology 14: e1005958.

Cosby RL, Chang NC, Feschotte C. 2019. Host–transposon interactions: Conflict, cooperation, and cooption. Genes and Development 33: 1098–1116.

Crescente JM, Zavallo D, Helguera M, Vanzetti LS. 2018. MITE Tracker: An accurate approach to identify miniature inverted-repeat transposable elements in large genomes. BMC Bioinformatics 19: 1–10.

Cridland JM, Macdonald SJ, Long AD, Thornton KR. 2013. Abundance and distribution of transposable elements in two drosophila QTL mapping resources. Molecular Biology and Evolution 30: 2311–2327.

Croll D, McDonald BA. 2012. The Accessory Genome as a Cradle for Adaptive Evolution in Pathogens (J Heitman, Ed.). PLOS Pathogens 8: e1002608.

Croll D, McDonald BA. 2017. The genetic basis of local adaptation for pathogenic fungi in agricultural ecosystems. Molecular Ecology 26: 2027–2040.

Daboussi MJ, Capy P. 2003. Transposable Elements in Filamentous Fungi. Annual Review of Microbiology 57: 275–299.

Van Dam P, Fokkens L, Ayukawa Y, Van Der Gragt M, Ter Horst A, Brankovics B, Houterman PM, Arie T, Rep M. 2017. A mobile pathogenicity chromosome in Fusarium oxysporum for infection of multiple cucurbit species. Scientific Reports 7: 1–15.

Devos KM, Brown JKM, Bennetzen JL. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Research 12: 1075–1079.

Dhillon B, Cavaletto JR, Wood K V, Goodwin SB. 2010. Accidental Amplification and Inactivation of a Methyltransferase Gene Eliminates Cytosine Methylation in Mycosphaerella graminicola. Genetics 186: 67-U139.

Dhillon B, Gill N, Hamelin RC, Goodwin SB. 2014. The landscape of transposable elements in the finished genome of the fungal wheat pathogen Mycosphaerella graminicola. Bmc Genomics 15.

Duret L, Galtier N. 2009. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. Annual Review of Genomics and Human Genetics 10: 285–311.

Faino L, Seidl MF, Shi-Kunne X, Pauper M, van den GC, Wittenberg AH, PHJ Thomma B, Bart PHJ Thomma D. 2016. Transposons passively and actively contribute to evolution of the two-speed genome 1 of a fungal pathogen The Netherlands Running title: Genome evolution by transposable elements. : 1091–1100.

Feurtey A, Lorrain C, Croll D, Eschenbrenner C, Freitag M, Habig M, Haueisen J, Möller M, Schotanus K, Stukenbrock EH. 2020. Genome compartmentalization predates species divergence in the plant pathogen genus Zymoseptoria. BMC genomics 21: 588.

Fouché S. 2020. Drivers of genome evolution in a fungal pathogen of wheat.

Fouché S, Badet T, Oggenfuss U, Plissonneau C, Francisco CS, Croll D. 2020. Stress-Driven Transposable Element De-repression Dynamics and Virulence Evolution in a Fungal Pathogen. Molecular biology and evolution 37: 221–239.

Fouché S, Oggenfuss U, Chanclud E, Croll D. 2021. A devil's bargain with transposable elements in plant pathogens. Trends in Genetics: 1–9.

Frantzeskakis L, Kracher B, Kusch S, Yoshikawa-Maekawa M, Bauer S, Pedersen C, Spanu PD, Maekawa T, Schulze-Lefert P, Panstruga R. 2018. Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. BMC Genomics 19: 1–23.

Frantzeskakis L, Kusch S, Panstruga R. 2019. The need for speed: compartmentalized genome evolution in filamentous phytopathogens. Molecular Plant Pathology 20: 3–7.

de Freitas Pereira M, Veneault-Fourrey C, Vion P, Guinet F, Morin E, Barry KW, Lipzen A, Singan V, Pfister S, Na H, et al. 2018. Secretome Analysis from the Ectomycorrhizal Ascomycete Cenococcum geophilum. Frontiers in Microbiology 9: 1–17.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends in Genetics 20: 417–423.

Gao D, Li Y, Kim K Do, Abernathy B, Jackson SA. 2016. Landscape and evolutionary dynamics of terminal repeat retrotransposons in miniature in plant genomes. Genome biology 17: 7.

Gladyshev E, Kleckner N. 2017. Recombination-independent recognition of DNA homology for repeat-induced point mutation. Current Genetics 63: 389–400.

González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. 2008. High Rate of Recent Transposable Element-Induced Adaptation in Drosophila melanogaster. Plos Biology 6: 2109–2129.

Grandaubert J, Bhattacharyya A, Stukenbrock EH. 2015. RNA-seq-Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus Zymoseptoria Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3-Genes Genomes Genetics 5: 1323–1333.

Habig M, Schotanus K, Hufnagel K, Happel P, Stukenbrock EH. 2021. Ago1 affects the virulence of the fungal plant pathogen zymoseptoria tritici. Genes 12.

Hane JK, Oliver RP. 2008. RIPCAL: A tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics 9: 1–12.

Hane JK, Oliver RP. 2010. In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. BMC Genomics 11.

Horváth V, Merenciano M, González J. 2017. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends in Genetics 33: 832–841.

Ito H, Gaubert H, Bucher E, Mirouze M, Vaillant I, Paszkowski J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. Nature 472: 115–120.

Jacobs FMJ, Greenberg D, Nguyen N, Haeussler M, Ewing AD, Katzman S, Paten B, Salama SR, Haussler D. 2014. An evolutionary arms race between KRAB zinc-finger genes ZNF91/93 and SVA/L1 retrotransposons. Nature 516: 242–245.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Kettles GJ, Hofinger BJ, Hu P, Bayon C, Rudd JJ, Balmer D, Courbot M, Hammond-Kosack KE, Scalliet G, Kanyuka K. 2019. SRNA profiling combined with gene function analysis reveals a lack of evidence for cross-kingdom RNAi in the wheat – Zymoseptoria tritici pathosystem. Frontiers in Plant Science 10.

Kimura M. 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. Journal of Molecular Evolution 16: 111–120.

Kremer SC, Linquist S, Saylor B, Elliott TA, Gregory TR, Cottenie K. 2020. Transposable element persistence via potential genome-level ecosystem engineering. BMC Genomics 21: 1–15.

Krishnan P, Meile L, Plissonneau C, Ma X, Hartmann FE, Croll D, McDonald BA, Sánchez-Vallet A. 2018. Transposable element insertions shape gene regulation and melanin production in a fungal pathogen of wheat. BMC Biology 16: 1–18.

Lai X, Schnable JC, Liao Z, Xu J, Zhang G, Li C, Hu E, Rong T, Xu Y, Lu Y. 2017. Genome-wide characterization of non-reference transposable element insertion polymorphisms reveals genetic diversity in tropical and temperate maize. BMC Genomics 18: 1–13.

Lerat E, Rizzon C, Biémont C. 2003. Sequence divergence within transposable element families in the Drosophila melanogaster genome. Genome Research 13: 1889–1896.

Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Lisch D. 2009. Epigenetic regulation of transposable elements in plants. Annual Review of Plant Biology 60: 43–66.

Lisch D, Bennetzen JL. 2011. Transposable element origins of epigenetic gene regulation. Current Opinion in Plant Biology 14: 156–161.

Lorrain C, Feurtey A, Möller M, Haueisen J, Stukenbrock EH. 2020. Dynamics of transposable elements in recently diverged fungal pathogens: lineage-specific transposable element content and efficiency of genome defences. bioRxivc.

Lorrain C, Oggenfuss U, Croll D, Duplessis S, Stukenbrock E. 2021. Transposable Elements in Fungi: Coevolution With the Host Genome Shapes, Genome Architecture, Plasticity and Adaptation. In: Encyclopedia of Mycology. Elsevier, 1–2.

Lu L, Chen J, Robb SMC, Okumoto Y, Stajich JE, Wessler SR. 2017. Tracking the genome-wide outcomes of a transposable element burst over decades of amplification. Proceedings of the National Academy of Sciences: 201716459.

Ma B, Li T, Xiang Z, He N. 2015. MnTEdb, a collective resource for mulberry transposable elements. Database 2015: 1–10.

Mäe A, Fillinger S, Sooväli P, Heick TM. 2020. Fungicide Sensitivity Shifting of Zymoseptoria tritici in the Finnish-Baltic Region and a Novel Insertion in the MFS1 Promoter. Frontiers in Plant Science 11: 1–10.

Mao H, Wang H. 2017. SINE-scan: An efficient tool to discover short interspersed nuclear elements (SINEs) in large-scale genomic datasets. Bioinformatics 33: 743–745.

Meile L, Croll D, Brunner PC, Plissonneau C, Hartmann FE, McDonald BA, Sánchez-Vallet A. 2018. A fungal avirulence factor encoded in a highly plastic genomic region triggers partial resistance to septoria tritici blotch. New Phytologist 219: 1048–1061.

Möller M, Habig M, Lorrain C, Feurtey A, Haueisen J, Fagundes W, Alizadeh A, Freitag M, Stukenbrock E. 2020. Recent loss of the Dim2 DNA methyltransferase decreases mutation rate in repeats and changes evolutionary trajectory in a fungal pathogen. : 1–27.

Müller K, Wickham H. 2020. tibble: Simple Data Frames. R package version 3.0.1.

Oggenfuss U, Badet T, Wicker T, Hartmann FE, Singh NK, Abraham L, Karisto P, Vonlanthen T, Mundt C, McDonald BA, et al. 2021. A population-level invasion by transposable elements triggers genome expansion in a fungal pathogen. eLife 10: 1–25.

Omrane S, Audéon C, Ignace A, Duplaix C, Aouini L, Kema G, Walker A-S, Fillinger S. 2017. Plasticity of the MFS1 promoter leads to multi drug resistance in the wheat pathogen Zymoseptoria tritici. mSphere: 1–42.

Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, Fillinger S. 2015. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. Environmental Microbiology 17: 2805–2823.

Paradis E. 2010. Pegas: An R package for population genetics with an integrated-modular approach. Bioinformatics 26: 419–420.

Paradis E, Schliep K. 2019. Ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. Bioinformatics 35: 526–528.

Pereira D, Oggenfuss U, McDonald BA, Croll D. 2021. Population genomics of transposable element activation in the highly repressive genome of an agricultural pathogen. Microbial Genomics 7.

Petrov DA, Aminetzach YT, Davis JC, Bensasson D, Hirsh AE. 2003. Size matters: Non-LTR retrotransposable elements and ectopic recombination in Drosophila. Molecular Biology and Evolution 20: 880–892.

Plissonneau C, Stürchler A, Croll D. 2016. The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat. mBio 7: 1–13.

Quinlan AR, Hall IM. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

R Core Team. 2020. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology 10: 417–430.

Rebollo R, Romanish MT, Mager DL. 2012. Transposable Elements: An Abundant and Natural Source of Regulatory Sequences for Host Genes. In: Bassler BL, ed. Annual Review of Genetics, Vol 46. 21–42.

Revell LJ. 2012. phytools: An R package for phylogenetic comparative biology (and other things). Methods in Ecology and Evolution 3: 217–223.

Rice P, Longden L, Bleasby A. 2000. EMBOSS: The European Molecular Biology Open Software Suite. Trends in Genetics 16: 276–277.

Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, et al. 2011. Effector diversification within compartments of the Leptosphaeria maculans genome affected by Repeat-Induced Point mutations. Nature communications 2: 202.

Le Rouzic A, Capy P. 2005. The first steps of transposable elements invasion: Parasitic strategy vs. genetic drift. Genetics 169: 1033–1043.

Schmitz RJ, Lewis ZA, Goll MG. 2019. DNA Methylation: Shared and Divergent Features across Eukaryotes. Trends in Genetics 35: 818–827.

Schotanus K, Soyer JL, Connolly LR, Grandaubert J, Happel P, Smith KM, Freitag M, Stukenbrock EH. 2015. Histone modifications rather than the novel regional centromeres of Zymoseptoria tritici distinguish core and accessory chromosomes. Epigenetics & Chromatin 8.

Sigman MJ, Slotkin RK. 2016. The First Rule of Plant Transposable Element Silencing: Location, Location, Location. The Plant Cell 28: 304–313.

Smit A, Hubley R, Green P. RepeatMasker Open-4.0.

Smith MR. 2019. TreeTools: create, modify and analyse phylogenetic trees.

Stamatakis A. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Stitzer MC, Anderson SN, Springer NM, Ross-Ibarra J. 2021. The genomic ecosystem of transposable elements in maize. PLOS Genetics 17: e1009768.

Stritt C, Gordon SP, Wicker T, Vogel JP, Roulin AC. 2017. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass Brachypodium distachyon. Genome Biology and Evolution 10: 1–38.

Stuart T, Eichten SR, Cahn J, Karpievitch Y V, Borevitz JO, Lister R. 2016. Population scale mapping of transposable element diversity reveals links to gene regulation and epigenomic variation. eLife 5: 1–27.

Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA. 2007. Origin and domestication of the fungal wheat pathogen Mycosphaerella graminicola via sympatric speciation. Molecular Biology and Evolution 24: 398–411.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 18: 292–308.

Testa AC, Oliver RP, Hane JK. 2016. OcculterCut: A comprehensive survey of at-rich regions in fungal genomes. Genome Biology and Evolution 8: 2044–2064.

Torres DE, Oggenfuss U, Croll D, Seidl MF. 2020. Genome evolution in fungal plant pathogens: looking beyond the two-speed genome model. Fungal Biology Reviews 34: 136–143.

van't Hof AE, Campagne P, Rigden DJ, Yung CJ, Lingley J, Quail MA, Hall N, Darby AC, Saccheri IJ. 2016. The industrial melanism mutation in British peppered moths is a transposable element. Nature 534: 102–105.

Volz EM, Koelle K, Bedford T. 2013. Viral Phylodynamics. PLoS Computational Biology 9.

Wang LG, Lam TTY, Xu S, Dai Z, Zhou L, Feng T, Guo P, Dunn CW, Jones BR, Bradley T, et al. 2020a. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Molecular Biology and Evolution 37: 599–603.

Wang C, Milgate AW, Solomon PS, McDonald MC. 2021. The identification of a transposon affecting the asexual reproduction of the wheat pathogen Zymoseptoria tritici. Molecular Plant Pathology 22: 800–816.

Wang L, Sun Y, Sun X, Yu L, Xue L, He Z, Huang J, Tian D, Tian D, Hurst LD, et al. 2020b. Repeat-induced point mutation in Neurospora crassa causes the highest known mutation rate and mutational burden of any cellular life. Genome Biology 21: 1–23.

Wenke T, Dobel T, Sorensen TR, Junghans H, Weisshaar B, Schmidt T. 2011. Targeted Identification of Short Interspersed Nuclear Element Families Shows Their Widespread Existence and Extreme Heterogeneity in Plant Genomes. the Plant Cell Online 23: 3117–3128.

Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982.

Wickham H. 2016. ggplot2: Elegant Graphics for Data Analysis. New York: Springer-Verlag.

Wickham H, Averick M, Bryan J, Chang W, McGowan L, François R, Grolemund G, Hayes A, Henry L, Hester J, et al. 2019. Welcome to the Tidyverse. Journal of Open Source Software 4: 1686.

Wickham H, François R, Henry L, Müller K. 2020. dplyr: A Grammar of Data Manipulation. R package version 1.0.0.

van Wyk S, Harrison CH, Wingfield BD, De Vos L, van der Merwe NA, Steenkamp ET. 2019. The RIPper, a web-based tool for genome-wide quantification of Repeat-Induced Point (RIP) mutations. PeerJ 7: e7447.

van Wyk S, Wingfield BD, De Vos L, van der Merwe NA, Steenkamp ET. 2021. Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota. Frontiers in Microbiology 11.

Xu Z, Wang H. 2007. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Research 35: 265–268.

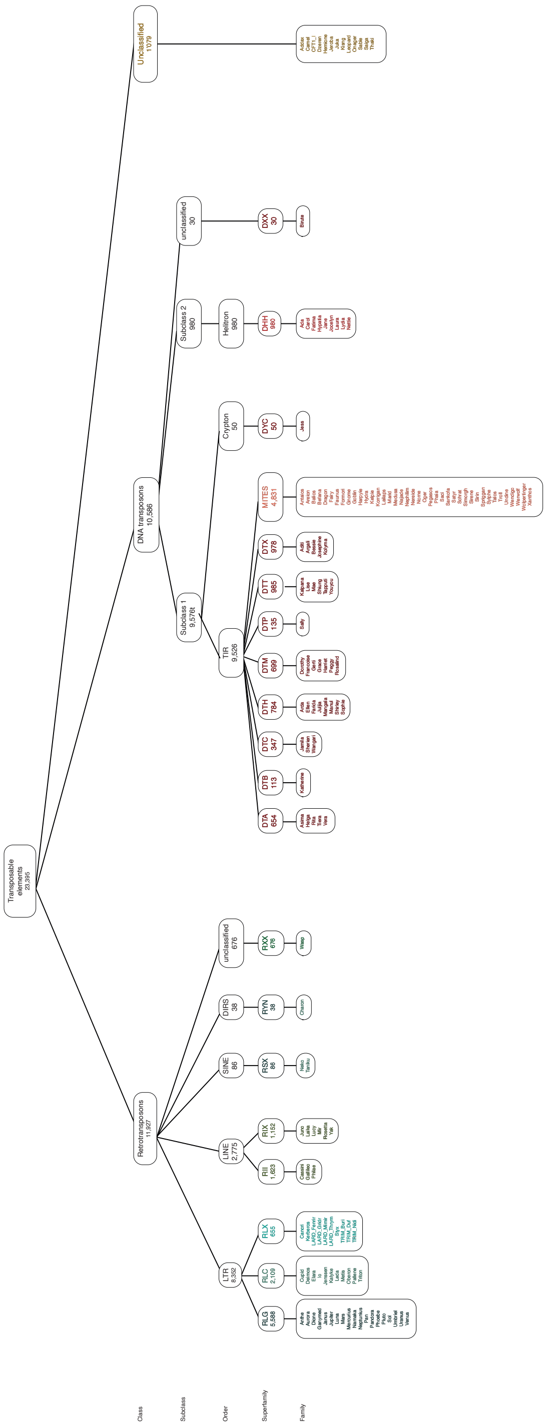
Yang P, Wang Y, Macfarlan TS. 2017. The Role of KRAB-ZFPs in Transposable Element Repression and Mammalian Evolution. Trends in Genetics 33: 871–881.

Yu G, Smith DK, Zhu H, Guan Y, Lam TTY. 2017. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods in Ecology and Evolution 8: 28–36.

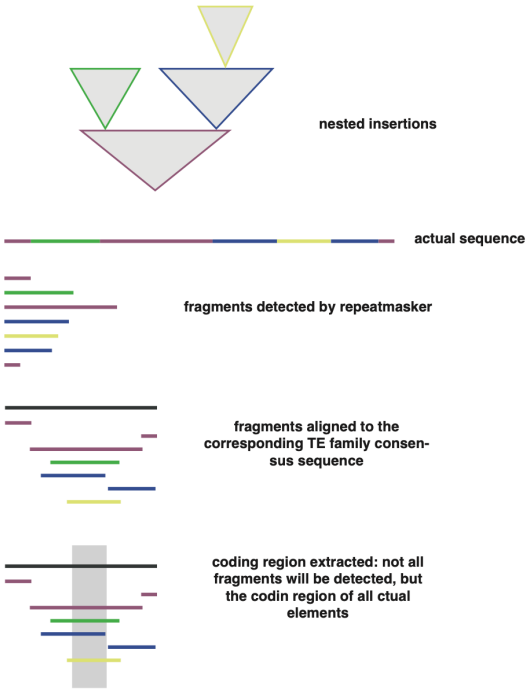
Zhan J, Pettway RE, McDonald BA. 2003. The global genetic structure of the wheat pathogen Mycosphaerella graminicola is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. Fungal Genetics and Biology 38: 286–297.

Zhang L, Hu J, Han X, Li J, Gao Y, Richards CM, Zhang C, Tian Y, Liu G, Gul H, et al. 2019. A high-quality apple genome assembly reveals the association of a retrotransposon and red fruit colour. Nature Communications 10: 1494.

Supplementary Files



Supplementary Figure S1: Hierarchy TE superfamilies: Classes, subclasses, orders, superfamilies as well as the tree-letter code according to Wicker et al (2007). The *Z. tritici* specific family names are according to Badet et al (2020).



Supplementary Figure S2: Approach to obtain multiple sequence alignments among TE copies. Due to the high number of nested insertions and partially deleted fragments, we aligned only coding regions.

Supplementary Table S1: Collection of 19 reference-quality genomes of *Zymoseptoria tritici*. Data from Badet et al (2020)

Supplementary Table S2: Metadata for all loci. Includes TE family, information about isolates of origin, position in the genome, niche and copy characteristics.

Chapter 3: Co-option of transposable elements across the fungal kingdom

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

[1]Laboratory of Evolutionary Genetics, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland.

[2]Laboratory of Cell and Molecular Biology, Institute of Biology, University of Neuchâtel, 2000 Neuchâtel, Switzerland

Author contributions: UO, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Writing - original draft; TB, Conceptualization, Data curation, Formal analysis, Methodology, Supervision, Writing; DC, Conceptualization, Funding acquisition, Methodology, Project administration, Resources, Supervision, Writing - original draft, Writing - review and editing

Unpublished manuscript

"In this world, there are things you can only do alone, and things you can only do with somebody else. It's important to combine the two in just the right amount."

Haruki Murakami (After Dark)

Abstract

How novel protein functions are gained is a central question in molecular biology. Key paths to novelty are gene duplications or recombination. Transposable elements (TEs) are increasingly recognized as major source of novel domain-encoding sequences driven by the ubiquitous TE transposition activity. Yet, the impact of TE coding sequences on the evolution of the proteome remains largely unknown. Here, we analyzed 1,327 genomes spanning the phylogenetic breadth of the fungal kingdom. We scanned proteomes for evidence of co-occurrence of TE-derived domains together with other unrelated conventional protein functional domains. We detected more than 13,000 predicted proteins with a TE-derived domain, of which 825 were present in more than 5 species, indicating that host-TE fusions have persisted over longer evolutionary time scales. We used the phylogenetic context to identify the age and retention of individual TE-derived domains. Most of the TE-derived protein domains originate from *Helitrons*, a TE superfamily that can capture surrounding genomic regions including genes during proliferation. Additionally, we found TE co-options at a higher rate in the Saccharomycotina, a subphylum of ascomycetes with generally small genomes depleted of TEs. Finally, we suggest that a predicted protein including two TE domains plays a role in TE and gene silencing in a subset of the ascomycete phyla. The gene shows multiple full and partial losses within the ascomycetes. Overall, our work establishes a kingdom-wide view of how TE activity influences the evolution of protein functions.

Introduction

Proteomes evolve through the loss of genes, changes in gene functions, horizontal acquisition or *de novo* gains (Van Oss & Carvunis, 2019). Novel functions and *de novo* genes often emerge after the duplication of already existing genes, followed by relaxed purifying selection and neofunctionalization of one of the copies (Andersson *et al.*, 2015). *De novo* genes can emerge through the addition of new open reading frames, the truncation or merging of open reading frames in existing genes or additional exonisation via novel splice sites in introns (Van Oss & Carvunis, 2019). Some *de novo* genes emerged without any homology to genes in the same species or closely related species, indicating that genes can also emerge from previously non-coding sequences or via horizontal gene transfer (Nishida, 2006; Keeling & Palmer, 2008; McLysaght & Guerzoni, 2015). Novel gene function also evolve by exon shuffling trough the

activity of transposable elements (TEs) (Kolkman & Stemmer, 2001; Ejima & Yang, 2003; Morgante *et al.*, 2005; Grabundzija *et al.*, 2016). Additionally, TEs can be an important source of new protein domains encoded by *de novo* genes (Bourque *et al.*, 2018). TE insertion into genes typically has deleterious effects on the encoded protein functions, therefore most TE insertions into genes are under strong purifying selection (Baucom *et al.*, 2008; Stritt *et al.*, 2017). In duplicated genes, pseudogenes or non-essential genes, TE insertions are less likely to be deleterious and can lead to neofunctionalization of the gene. Furthermore, some TEs can capture and amplify segments containing coding sequences, as described for *Helitrons* in maize or *Pack-Mules* in rice (Jiang *et al.*, 2004; Gupta *et al.*, 2005). Either TE insertion into existing genes or gene capture by TEs can lead to fusion proteins with domains derived both from TEs and genes. If retained over time, such host-TE fusions are likely to lose functions related to TE proliferation, a process also identified as TE domestication or exaptation (Hoen & Bureau, 2015; Bourque *et al.*, 2018). Host-TE fusions that provide essential new functions are expected to be retained, although the evolutionary timeframes of such domestication events remain poorly understood.

The fusion of host genes with TE derived domains are generated at a considerable rate (Hoen & Bureau, 2015). The initial function of TE coding sequences is typically restricted to few functions related to the mobilization and duplication of the elements (Wicker *et al.*, 2007; Wells & Feschotte, 2020). Yet, how TE sequences provide additional functions for existing coding sequences is largely unclear with few exceptions. A well-studied example of host-TE fusion is the V(D)J recombination that leads to immunoglobin diversification and provides highly conserved adaptive immunity in vertebrates (Agrawal *et al.*, 1998; Koonin *et al.*, 2020). The recombination activating genes *RAG1* and *RAG2* code for proteins ensuring DNA cleavage at recombination signal sequences, which are derived from the terminal inverted repeats of the TE (Martin *et al.*, 2020). *RAG1* and *RAG2* retained mobility and can re-shuffle recombination signal sequences creating the basis for rapid sequences changes in the face of new antigens (Agrawal *et al.*, 1998). Even though the V(D)J recombination is not conserved across vertebrates, the fusion is thought to have occurred ~500 million years ago (Agrawal *et al.*, 1998; Kapitonov & Jurka, 2005). *RAG1* is a host-TE fusion gene, containing the transposase of the *Transib*-like DNA transposon and a RING finger ubiquitin ligase at the N-terminal that probably acts in dimerization and as a ligase for ubiquitination (Yurchenko *et al.*, 2003). *KRABINER* is a host-TE fusion in vespertilionid bats including a *Mariner* DNA transposon inserted into the intron of *ZNF112*, followed by alternative splicing and a single base pair

deletion that fused the TE and host genes into a single open reading frame (Cosby *et al.*, 2021). *KRABINER* controls the regulation of a large network of other genes (Cosby *et al.*, 2021). Host-TE fusion events were recently reviewed in Hoen & Bureau (2015) and Jangam et al (2017), highlighting that TE-derived sequences often remain largely intact including terminal inverted repeats and open reading frames matching known TE functions.

Fungal genomes show high variability in TE content and composition even between closely related species suggesting significant recent TE activity (Raffaele & Kamoun, 2012; Hess *et al.*, 2014; González-Sayer *et al.*, 2021). TEs have played an important role in the evolution of host-associated lifestyles or local adaptation to external stress including tolerance of pesticides (Omrane *et al.*, 2015, 2017; Hartmann *et al.*, 2017; Gladyshev, 2017). Fungi associated with animals and pathogenic lifestyles in general tend to have higher numbers of TE insertions into genes, which could either be recent insertions in non-essential genes or host-TE fusions (Muszewska *et al.*, 2019). Old TE insertions are more likely to affect genes with enzymatic rather than protein-protein interaction functions (Muszewska *et al.*, 2019). In the fission yeast *Schizosaccharomyces pombe*, *Abp1*, *Cbh1*, *Cbh2* are centromeric *pogo* derived host-TE fusions that led to retrotransposon silencing (Smit & Riggs, 1996; Cam *et al.*, 2008; Mateo & González, 2014). A *Bel-Pao* derived *gag* sequence was recently shown to have fused with *PEX14* gene, acquiring an intron and being co-opted in fungi (Wang & Han, 2021). The recent TE activity in many lineages of the fungal kingdom and the exceptional genomic resources available for such compact genomes provide a vast potential to retrace the emergence of host-TE fusions over deep evolutionary timeframes.

Here, we used a systematic approach to detect host-TE fusions in the genomes of 1,327 fungal species. We used gene orthology and phylogenomic analyses to detect the emergence and retention of TE-derived domains in fungal proteomes. We found that TE-derived helicases are the dominant TE partner in host-TE fusions. The subphylum Saccharomycotina including the model yeasts *S. pombe* and *Saccharomyces cerevisiae* shows elevated content of host-TE fusions despite their typically small and repeat-poor genomes. Host-TE fusions are enriched for binding functions to heterocyclic compounds, organic cyclic compounds ATP, adenyl ribonuclease and adenyl nucleotide. We found also widespread host-TE fusions in ascomycetes involved in gene silencing originating from *Helitron* and *Maverick* domains. Phylogenetic analyses suggest independent origins of identical host-TE fusions, uneven rates of gene retention and secondary losses.

Methods

Retrieval of genomes and gene annotations

We retrieved high-quality genomes and gene annotations of 1,327 fungal and oomycetes species from two different sources. A total of 1,011 genomes belong to the phylum Ascomycota, 206 Basidiomycota, 30 Microsporidia, 28 Mucoromycota, 16 Chytridiomycota, 16 Zoopagomycota, 2 Blastocladiomycota, 2 Cryptomycota and 16 Oomycota as an outgroup (see Supplementary Table S1 for full references and additional data). The budding yeast genomes were retrieved from Shen and colleagues (2018). We retrieved additional genomes and gene annotation from fungal and oomycetes genomes were retrieved from NCBI (see Supplementary Table S1 for full references and additional data).

Species phylogeny reconstruction

For the species tree reconstruction, we followed an approach published by Li et al (2020) to reconstruct the fungal tree of life. Briefly, we first identified a set of single-copy orthologous genes in each of the 1,327 genomes using BUSCO v 4.1.4 searching the fungi or oomycote orthology database version 10 for fungal and oomycete species, respectively (Manni *et al.*, 2021). The method identified a maximum set of 756 BUSCO genes in the genome of the fungus *Colletotrichum plurivorum*. The identified BUSCO genes were then translated into protein sequences respecting the relevant genetic code (code 12 for Saccharomycotina species except for *Pachysolen tannophilu*s (Pactanno for which code 26 was used and code 1 for all other species) (Mühlhausen *et al.*, 2016). Of the 756 BUSCO genes identified, a random sample of 100 of the resulting BUSCO protein sequences was then concatenated using the geneStitcher.py script (https://github.com/ballesterus/Utensils) and aligned using mafft v 7.475 with the parameters *--maxiterate 1000 --auto* (Katoh & Standley, 2013). The resulting alignment was then trimmed using trimAl v 1.4.rev15 with the *-gappyout* option (Capella-Gutiérrez *et al.*, 2009). We estimated the best fitting evolutionary models for the concatenated 100 protein sequences using partitionfinder v 2 with the quick option *-q* and default RAxML v 8.2.12 (Lanfear *et al.*, 2012; Stamatakis, 2014). The resulting partitioned model was then applied for phylogenetic inference using iqtree2 v 2.1.2 after 1,000 replicates for ultrafast bootstrap and 2 independent runs with *-B 1000 --runs 2* (Minh *et al.*, 2020). We rooted the tree

with the oomycete *Phytophthora parasitica* and visualized the tree using the R packages ggtree, ggtreeExtra and treeio (Yu *et al.*, 2017; Wang *et al.*, 2020a; Xu *et al.*, 2021)

Annotation of functional domains in the proteomes

To identify putative functional domains across the species proteomes, we downloaded the annotated domains hidden Markov models from the PFAM release 31 (Mistry *et al.*, 2021). We used the hmmsearch function from the HMMER package v 3.3.2 to scan all the species proteomes for functional domains with the *--noali* option to speed up the process (Eddy, 2011). We then filtered the matching domains for a minimal bitscore of 50 and a maximal e-value of 1e-17 using the HmmPy.py script.

Inference of trophic modes

We categorized species using the CATAStrophy v 0.1.0 according to their trophic mode (Hane *et al.*, 2020). Using the predicted proteins from all genomes, we searched for genes encoding carbohydrate-degrading enzymes (CAZymes) with dbCAN v 8 (Zhang *et al.*, 2018). As for the PFAM annotation, we performed hmmscans on each proteome using the dbCAN hidden Markov models as query. We then applied the CATAStrophy algorithm to predict the most likely trophic mode based on the set of encoded CAZymes.

Gene orthology analysis

We inferred gene orthology among all species based on protein sequence identity. We used orthofinder v 2.4.1, which implements diamond blast v 0.9.24 for homology searches across the pool of predicted proteins (Buchfink *et al.*, 2014; Emms & Kelly, 2018). From the initial set of 13,863,658 individual proteins encoded by all genomes combined, orthofinder grouped 7,860,083 proteins into 299,713 orthogroups.

Detection of candidate host-TE fusions

We retrieved previously reported PFAM domains associated with different fungal TE superfamilies (Muszewska *et al.*, 2019: see there Supplementary Table S2) and filtered for genes encoding such TE-associated PFAM domains. In a second filtering step, we removed proteins annotated exclusively with TE-associated PFAM domains. We excluded PFAM with similarity to any of the fungal TE PFAM based on SCOOP and HHSearch (Hoede *et al.*, 2014) (Supplementary Table S2). We removed all oomycete genes. We filtered out genes if the identified TE and non-TE PFAM domains had an overlap of more than 5% in the amino acid

sequence. Such overlaps may indicate that the two annotations identify the same protein domain. Overlaps were identified using bedtools v 2.30.0 with the *intersect* function (Quinlan & Hall, 2010). We retained candidate orthogroups including host-TE fusion proteins if genes encoding independent TE and non-TE PFAM domains were represented in at least five species and belong to the same orthogroup.

Gene ontology term enrichment analyses

We analyzed the enrichment of specific gene ontology (GO) terms among host-TE fusion genes compared to the background of all genes. To reduce the computational load, we defined the background as a 1% random subset of the entire set of genes (subset: n = 358,350). GO terms were assigned to genes using a GO-PFAM term translation based on Mitchell et al (2015). We created a GOAllFrame object with the AnnotationDbi package v 1.54.1 and constructed a GeneSetCollection with GSEABase v 1.54.0 (Morgan *et al.*, 2021; Pagès *et al.*, 2021). We calculated enrichment *p*-values using the *hyperGTest* in the Category package v 2.58.0 (Gentleman, 2021). For each MF (molecular function), BP (biological process) and CC (cellular component) GO term enrichment, a *p*-value cut-off of 1e-10 and a minimum term size of 20 was applied.

Additional filtering for copy-number variation in host-TE fusion genes

To detect potential activity of the TEs represented by the identified PFAM domains in individual genomes, we analyzed potential copy-number variation of the host-TE fusion genes and their respective PFAM terms. To reduce conservatively detecting host-TE fusion genes generated by recent TE insertion events, we required host-TE fusion genes to be present in at least 5 species belonging to the same order, and present in at least 20 species. Furthermore, we analyzed candidate host-TE fusion genes for their PFAM domain order along the amino acid sequence and removed orthogroups without a conserved domain order. After filtering, we extracted the predicted function of the non-TE candidate function based on the information provided in the PFAM database (Mistry *et al.*, 2021). To remove host-TE fusion gene candidates potentially erroneously identified due to the physical proximity of genes in fungal gene clusters, we performed gene cluster analyses using antiSMASH v 3.0 (Blin *et al.*, 2017). Sequence alignments were generated with MAFFT v 7.487 with the parameters --reorder --localpair --maxiterate 1000 --leavegappyregion (Li & Durbin, 2009; Li *et al.*, 2009).

Results

Analysis of fungal genomes and phylogenetic reconstruction

We analyzed genomes of 1,327 fungal species belonging primarily to phyla of ascomycetes and basidiomycetes (Figure 1A). Based on a set of 100 single-copy genes we constructed a maximum likelihood phylogenetic tree (Figure 1B; Supplementary file F1). The tree resolves the fungal phylogeny consistent with recent analyses of similar breadth (Shen *et al.*, 2020). ascomycetes are segregated into two large groups including the Saccharomycotina that mostly contain yeast species and the Sordariomycotina, Eurotiomycotina, Dothideomycotina and 9 other classes, that mostly contain filamentous species. The analyzed genomes are generally of high completeness based on BUSCO analyses for almost all species (Figure 1C). Genomes sizes are highly variable based on estimates from genome assembly sizes. The observed range is 2.2 – 1,230 Mb with the smallest genomes being found as expected among Cryptomycota and Microsporidia (Figure 1C). Genome-wide GC content is on average 46.4 % with an observed range between 16.3-67.8 % (Figure 1C). Oomycete genomes and Blastocladiomycota tend to have higher GC contents, while the genomes of Saccharomycotina, Chytridiomycota, Cryptomycota, Microsporidia, Mucoromycota and Zoopagomycota have generally lower GC contents.

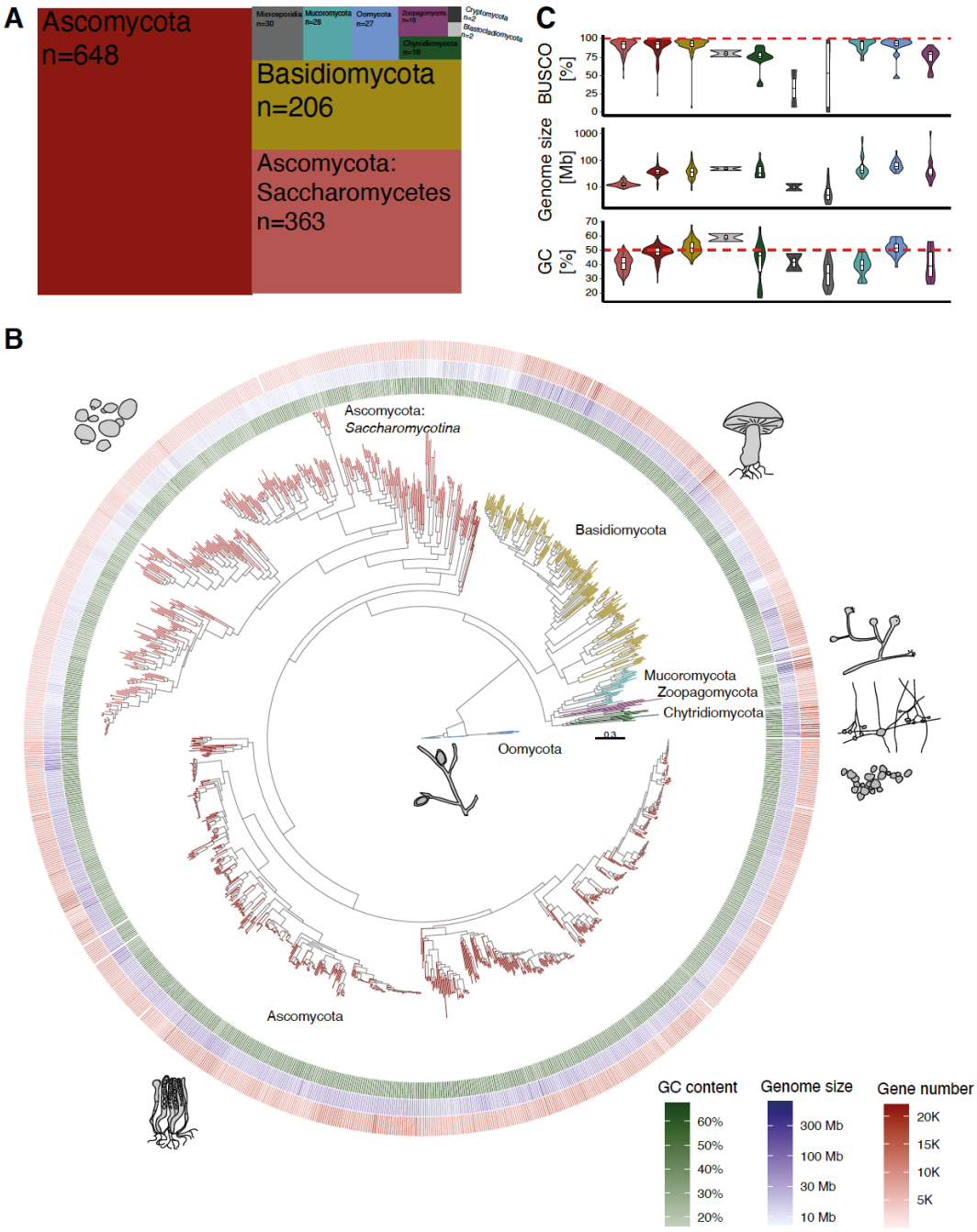


Figure 1: Phylogenomic analysis and genome properties across the fungal kingdom and oomycetes: (A) Number of species per phylum analyzed. The subphylum of Saccharomycotina is shown separately from the other ascomycetes. (B) Phylogenomic tree of the fungal kingdom based on 100 concatenated orthologous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

Uneven rates of host-TE fusions across the fungal kingdom

We next analyzed coding sequences for conserved domains present in the PFAM database. To define candidate host-TE fusion we require that at least one conserved domain matches a domain thought to be exclusively associated with TEs (see methods) and at least one more non-TE domain. The stringent filtering for candidates detected in 20 species or more allowed us to only obtain the most conserved host-TE fusions over deep evolutionary times, and to remove

remnants of deleterious or neutral TE insertions into genes. From a total of 39,655 unique proteins associated with a TE-associated domain across all genomes, we found that 13,342 also contain a non-TE domain (Figure 2A; Supplementary Table S3). A total of 1,305 species (98.3 %) carries at least one gene matching our criterion for host-TE fusion genes. In four species we did not detect any host-TE fusion (*Emmonsia* sp., *Lachaencea* sp., *Penicillium roqueforti* and *Pyricularia* sp.). We found on average 297.5 host-TE fusions (range 0-3,311) per genome (Figure 2B). Overall, 0.6-17.0 % of all annotated genes of a species are host-TE fusions. Saccharomycotina has on average higher proportions of host-TE genes per genome (5.2 % compared to 2.8 % across all genomes; Figure 2B). Species of the Saccharomycotina have overall genomes with fewer genes compared to other ascomycetes (Figure 2B). Additional outlier species with high proportions of host-TE fusion genes include the basidiomycetes *Armillaria ostoyae* and *Microbotryum silene-dioica*, as well as the ascomycete *Fusarium poae* and the mucoromycete *Rhizopus delemar*. We found no correlation between the number of detected host-TE fusions, BUSCO completeness scores, GC content or genome size suggesting the variation in host-TE among lineages is not meaningfully explained by variation in genome assembly quality (Supplementary Figure S1).

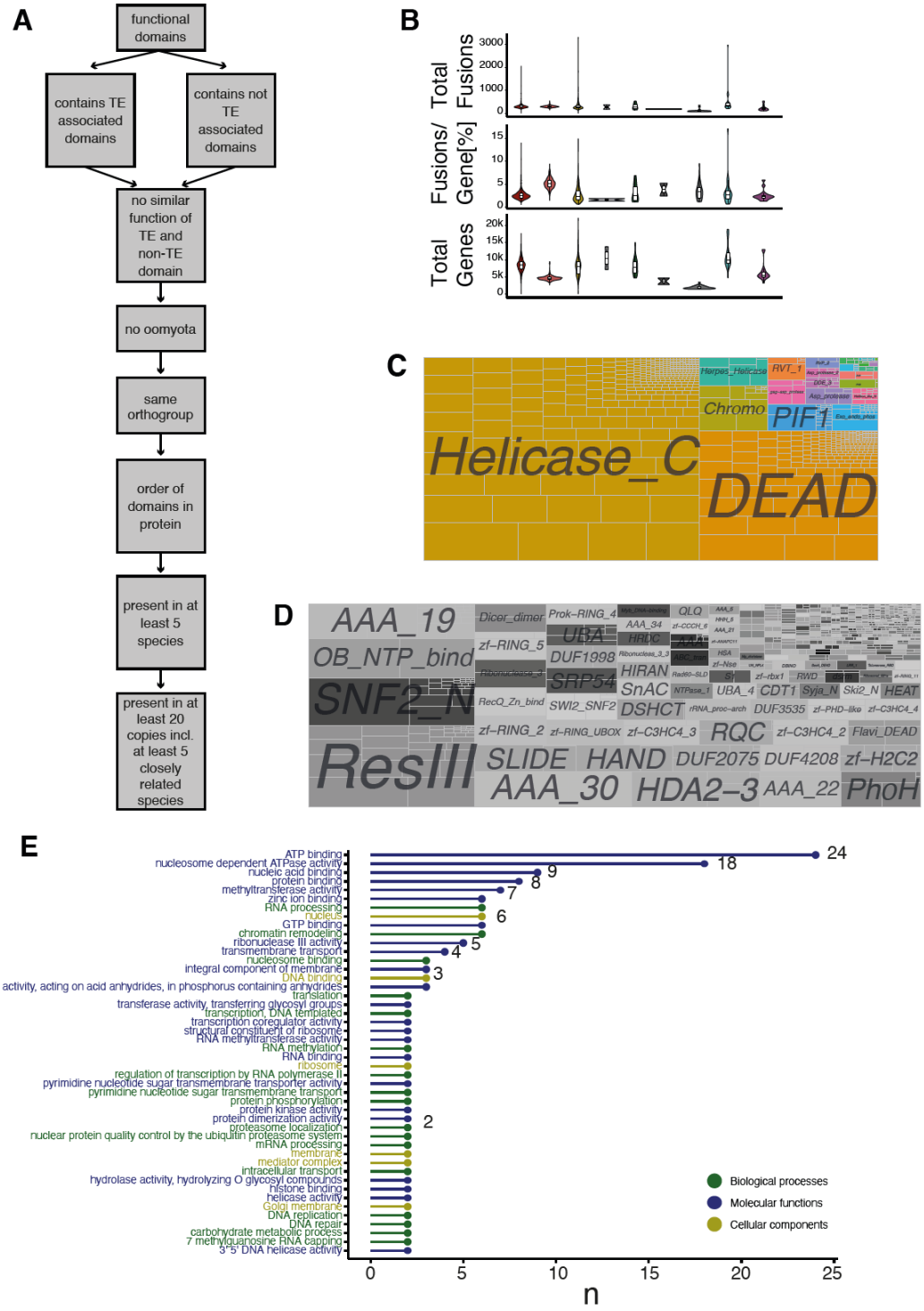


Figure 2: Host-TE fusion events identified across the fungal kingdom: (A) Overview of host-TE fusion detection steps. (B) Number of fusions detected per species, number of fusions detected per gene and number of annotated genes per genome. (B) Function of TE derived domains in all detected host-TE fusions. Squares indicate the number of individual fusions. (D) Function of non-TE derived domains of host-TE fusion genes. (E) Counts of non-TE derived domain gene ontology terms. Gene ontology with a single occurrence can be found in Supplementary Table S7.

Functions of evolutionarily retained host-TE fusion genes

We restricted our analysis to 824 (115,497 occurrences) host-TE fusion orthogroups where an ortholog is present in at least five species, thereby retaining the evolutionarily conserved host-

TE fusions. From the set of 824 individual host-TE fusions (Supplementary Table S4) we identified 22 distinct TE-associated PFAM. The TE-related PFAM covers mostly domains of *Helicase_C* and *DEAD* helicases in either the *AcademH* or *KolobokH* superfamilies (Figure 2C). The diversity in non-TE PFAM domains consistently found across all orthologs of a host-TE fusion protein is substantially higher with 383 individual non-TE PFAMs. In particular, the domains included ResIII, SNF2_N, OB_NTP_bind and AAA_19 functions (Figure 2D). The 383 non-TE PFAM are associated with 66 gene ontology (GO) terms with ATP binding, nucleosome-dependent ATP activity, nucleic acid binding, protein binding, methyltransferase activity, GTP binding, nucleus, zinc ionic binding, RNA processing and hydrolase activity, acting on acid anhydrides, in phosphorus containing anhydrides being the prevalent functions (Figure 2E). Over 40 % of all non-TE domains could not be associated with a GO term.

We then focused on a more restricted set including the most evolutionarily conserved host-TE fusions by requiring an ortholog to be present in at least 20 species (and at least five species belonging to the same order). The resulting subset of host-TE fusions contains 241 genes encoding 125 distinct non-TE PFAM terms. Domains included *SNF2_N*, *UBA*, *zf-H2C2*, *Rad60-SLD* and *UBA_4* present in five or more fusion host-TE fusions (Supplementary Figure S2). Domains with functions related to nucleotide binding are enriched in this set of 241 candidates (Figure 3; Supplementary Table S5). Fusion proteins include a homolog of centromere protein *CENP-B* also described as *Abp1*, *Cbh1* and *Cbh2* centromere protein N-domain in *S. pombe* that were previously described as a host-TE fusion (Smit & Riggs, 1996; Nakagawa *et al.*, 2002; Cam *et al.*, 2008; Mateo & González, 2014). Furthermore, the subset contains predominantly TE-associated functions related to helicases (*Helicase_C*, *DEAD* helicase) (Supplementary Figure S2). We also identified a fusion between a *DEAD* helicase and a Dicer dimerization domain (PF00270_Can55 and PF00271_Can80). The Dicer protein is involved in RNA interference and protection against TE activity or viral infection and has been previously identified as containing a helicase domain (Hammond, 2005).

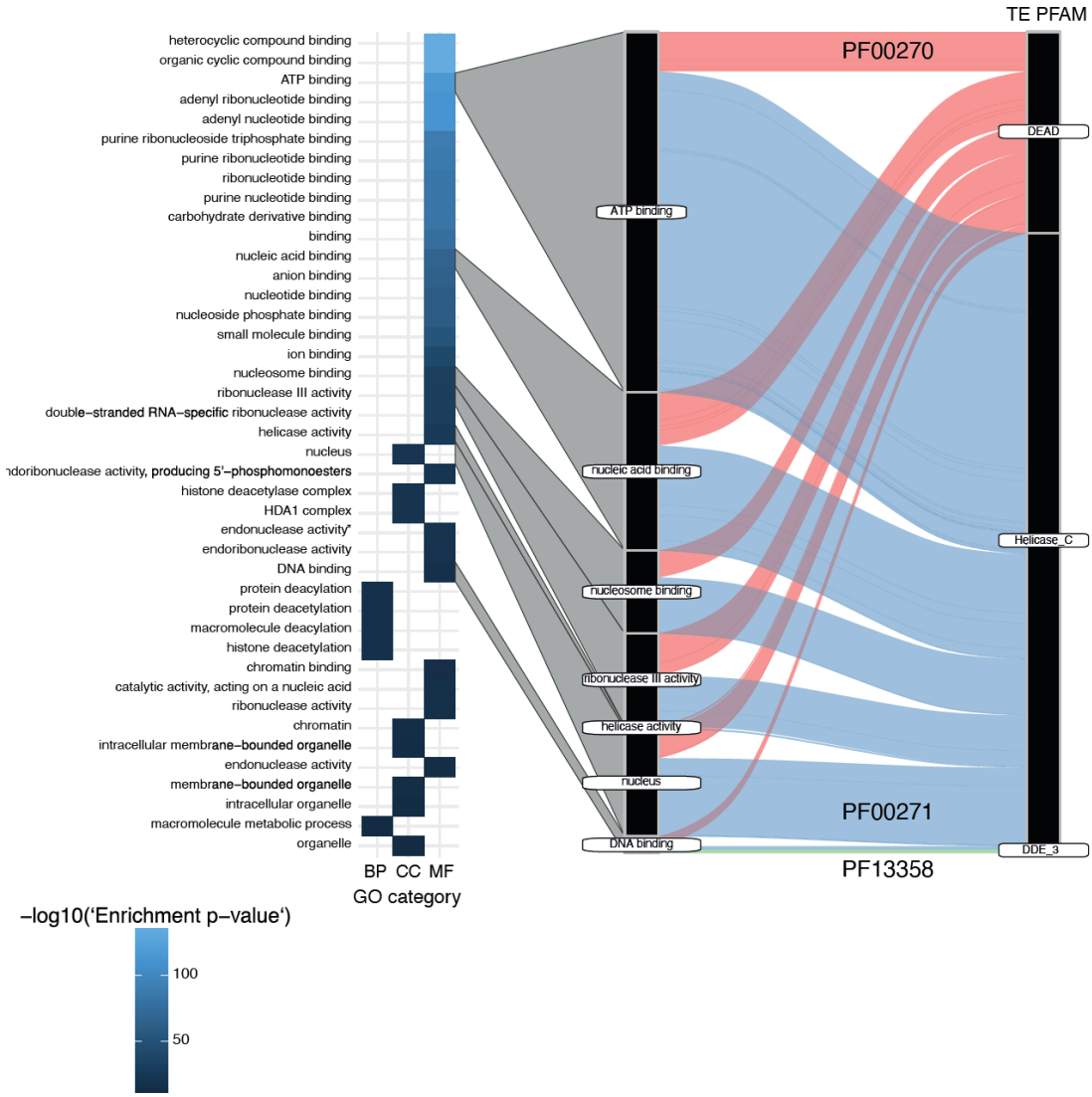


Figure 3: Gene enrichment analysis: Left side: Gene enrichment analysis of the non-TE derived domains, ordered by p-value. BP = biological process, CC = cellular component, MF = molecular function. Right side: corresponding TE-derived domains that are in fusion with enriched GO terms. DEAD and Helicase_C are helicases derived from AcademH or KolobokH TEs. DDE_3 are transposases derived from TIR DNA TEs. * active with either ribo- or deoxyribonucleic acids and producing 5'-p hosphomonoesters.

We highlight an effector complex for heterochromatic transcriptional silencing (SHREC) with a function in heterochromatin silencing (PF00385_Can4 and PF00271_Can16), described in *S. pombe* (Sugiyama *et al.*, 2007). SHREC is a host-TE fusion that includes a *Helicase_C* derived from *AcademH* or *KolobokH*, an additional TE-derived *Chromo* domain from *Maverick* TEs and a conserved non-TE domain *zf-CCCH_6*. In addition to the conserved non-TE domain *zf-CCCH_6* and the two TE domains *Helicase_C* and *Chromo*, almost all copies of SCREC contain *SNF2_N* and *zf-PHD-like* domains. Around half of the fusion protein variants contain *ResII* or *PHD* domains in addition to 88 more rarely associated domains (Figure 4A; Supplementary Figure S3). SCREC is present primarily in ascomycetes with the highest representation in the Eurotiomycetes (*n*=169), Dothideomycetes (*n*=115) and Leotiomycetes

(*n*=34). Lower numbers are found in Lecanoromycetes (*n*=4), Orbiliomycetes (*n*=5), Pezizomycetes (*n*=9), Taphrinomycetes (*n*=1) and Xylonomycetes (*n*=1). SHREC is largely absent in the large class of Saccharomycotina (*n*=1) and only present in *Schizosaccharomyces cryptophilus*, *S. japonicus* and *S. pombe* of the Schizosaccharomycotina. Weak representation is also found in basidiomycetes of the class Agaricomycetes (*n*=4) and Dacrymycetes (*n*=1). In two ascomycete species (*Aspergillus carbonarius* and *Phialophora americana*), SHREC has a paralog, with one duplication that affected the gene with both TE domains and one duplication that affected the *Helicase_C* domain gene. A multiple sequence alignment of the duplicated genomic regions confirms the conservation of the individual domains (Figure 4B).

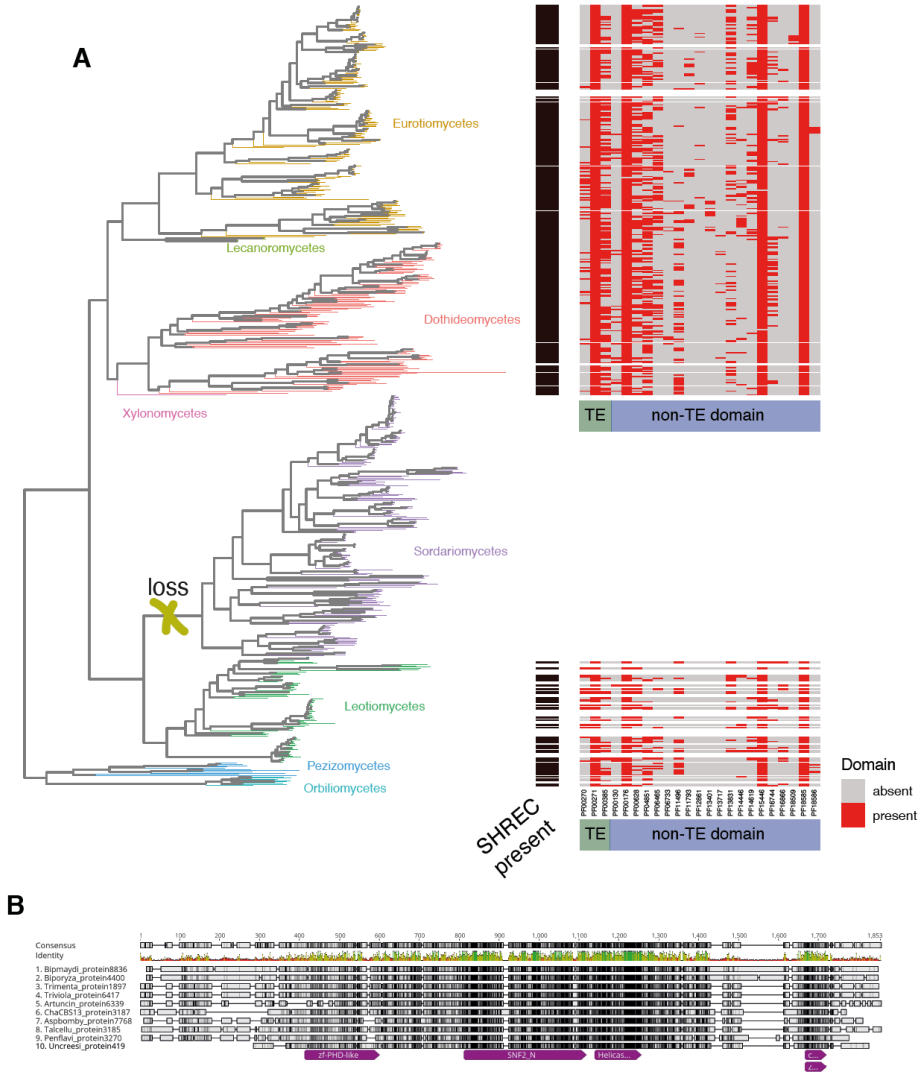


Figure 4: Host-TE fusion candidate SHREC: (A) Presence of the candidate in ascomycetes with Helicase_C and Chromo TE-derived domains. Black = SHREC is present. Red = the domain is present. (B) Multiple sequence alignment of a subset of SHREC ascomycetes with the relative position of the coding regions. The full alignment containing all 348 copies can be found in Supplementary Figure S4. The five coding regions are marked in purple below the alignment.

Discussion

Transposable elements play important roles in triggering chromosomal rearrangements, disrupting genes, and changing gene expression over short evolutionary timeframes. In addition, the retention and domestication of TE-derived domains underpins significant functional diversity in proteomes. Our analyses of host-TE fusion genes across 1,327 fungal genomes revealed an important role of helicase domains. In vertebrates and plants, terminal inverted repeat transposase domains are predominantly associated with host proteins (Hoen & Bureau, 2015; Cosby *et al.*, 2021). Host-TE fusion genes are unevenly distributed among major fungal phylogenies but tend to associate with processes involved in genome integrity and defense against foreign sequences including TEs.

Uneven rates of host-TE fusions across the fungal kingdom

We find that the compact genomes of Saccharomycotina contain a higher proportion of host-TE fusion genes compared to other fungi but have overall similar absolute numbers of host-TE fusions per genome. Interestingly, species of the Saccharomycotina do not carry defense mechanisms against TEs shared by other ascomycetes (*i.e.* repeat-induced point mutations or RIP) (van Wyk *et al.*, 2021). RIP is a mechanism that induces point mutations in all copies of duplicated regions of a certain length, including transposable elements and genes (Selker & Garrett, 1988; Selker, 2002; Galagan & Selker, 2004). RIP can cause early stop codons or other deleterious mutations in coding regions leading to loss-of-function of duplicated gene sequences (Dhillon *et al.*, 2010). Active RIP in a lineage might largely prevent the evolution of essential gene functions through gene duplication (Hane *et al.*, 2015) Consequently, activity of RIP may underpin low rates of duplicated genes among ascomycetes (Skamnioti *et al.*, 2008). Host-TE fusions of essential genes may plausibly arise after gene duplications whereby one copy remains essential, and the other copy is under relaxed purifying selection to gain new functions including through TE domain fusion events. The TE content of genomes of Saccharomycotina (and other yeast species) is generally low compared to other fungal species yet several species retain TE activity (Holton *et al.*, 2001; Bleykasten-Grosshans & Neuvéglise, 2011; Zhu *et al.*, 2014; Potocki *et al.*, 2019). Identifying mechanisms creating host-TE fusions remains challenging as, regardless of genomic defenses, non-deleterious insertions of TEs into open reading frames of existing genes are likely very rare.

Dominance of helicase functions in host-TE fusions

Most detected host-TE fusions encode TE-derived helicase domains. The most common source appears to be TEs of the superfamilies *AcademH* and *KolobokH* belonging to DNA transposon with terminal inverted repeats. The specific *DEAD* and *Helicase_C* helicase domains were only recently recognized as of TE origin likely due to the recent discovery of *AcademH* and *KolobokH* TEs (Muszewska et al (2019). *AcademH* has been found as low-copy TEs in basidiomycetes, ascomycetes as well as Mucoromycota (Kojima, 2020). Helicases in general provide functions for the unwinding of DNA and are involved in DNA repair pathways (Croteau *et al.*, 2014). Helicases from *Helitrons* though are known to be able to capture neighboring regions during transposition events (Morgante *et al.*, 2005; Chellapan *et al.*, 2016). *Helitrons* might thus generate host-TE fusions through the capture of genes rather than their insertion into coding sequences. Whether the preponderance of host-TE fusions with *AcademH* helicases is related to such a promiscuous mechanism to capture neighboring genes remains unknown. Recurrent gene capture by helicase containing TEs could explain the high helicase diversity in fungi and their dominance among host-TE fusion genes (Chellapan *et al.*, 2016). Over evolutionary time, host-TE fusion genes are expected to abolish transposition-related functions as an element of the domestication process. Consistent with this expectation, most identified host-TE genes across fungi have low copy numbers.

Complex retention of a major host-TE fusion

DNA binding activity is predominant among fungal host-TE fusion genes and is also featured among the most phylogenetically conserved fusions. The SHREC host-TE fusion candidate shows a patchy distribution in some classes of ascomycetes, with sparse presence in other clades. Some classes of ascomycetes do not contain SHREC, which could be an indication that this host-TE fusion was lost. SHREC corresponds to the Snf2/Hdac repressive complex (SHREC), which leads to transcriptional silencing of genes and TEs in *S. pombe* (Sugiyama *et al.*, 2007). SHREC contains two TE-derived domains, *Helicase_C* and *Chromo*. In TEs, helicase domains are involved in nucleic acid and ATP binding (Kojima, 2020). *Chromo* domains are known in the superfamilies of *Maverick* (older naming Polinton) and chromoviruses (a group of *Gypsy*) and are located at the C-terminus of the integrase (Pritham *et al.*, 2007; Gao *et al.*, 2008; Quesneville, 2020). *Chromo* domains interact mostly with methylated histones (Brehm *et al.*, 2004; Gao *et al.*, 2008). The patchy distribution of SHREC and the prevalent loss of the *Chromo* domain indicate that the complex might not be functional in all fungi (Lei *et al.*, 2021). Chromatin remodeling is also controlled by the dicer domain,

which is present in all ascomycetes except Saccharomycotina (PF00270_Can55 and PF00271_Can80) (Sugiyama *et al.*, 2007; Marina *et al.*, 2013; Allshire & Madhani, 2018).

Fungal proteomes have been significantly shaped by TE insertions of various age, phylogenetic distribution and function. The exact mechanisms creating still functional proteins remain poorly documented, but intraspecies screens will improve our understanding. We suggest gene capturing by TEs (*i.e.*, *Helitron*) as an alternative mechanism to TE insertion into introns followed by alternative splicing to create host-TE fusion. Future research will shed a light on the ongoing evolution of host-TE fusions.

Acknowledgements

We thank Hadi Quesneville, Casey Bergman, Ksenia Krasileva and Anne Nakamoto for help with TE associated PFAM. We also thank Sabina Tralamazza for help with the biosynthetic clusters and Emile Gluck-Thaler for help with the multiple sequence alignment visualization.

Data availability

Supplementary Data available on https://zenodo.org/record/6029584.

Literature chapter 3

Agrawal Alka, Eastman Quinn M, Schatz David G. 1998. Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. Nature 394: 744–751.

Allshire Robin C, Madhani Hiten D. 2018. Ten principles of heterochromatin formation and function. Nature Reviews Molecular Cell Biology 19: 229–244.

Andersson Dan I, Jerlström-Hultqvist Jon, Näsvall Joakim. 2015. Evolution of New Functions De Novo and from Preexisting Genes. Cold Spring Harbor Perspectives in Biology 7: a017996.

Baucom Regina S, Estill James C, Leebens-Mack Jim, Bennetzen Jeffrey L. 2008. Natural selection on gene function drives the evolution of LTR retrotransposon families in the rice genome. Genome Research 19: 243–254.

Bleykasten-Grosshans Claudine, Neuvéglise Cécile. 2011. Transposable elements in yeasts. Comptes Rendus Biologies 334: 679–686.

Blin Kai, Wolf Thomas, Chevrette Marc G, Lu Xiaowen, Schwalen Christopher J, Kautsar Satria A, Suarez Duran Hernando G, de los Santos Emmanuel LC, Kim Hyun Uk, Nave Mariana, Dickschat Jeroen S, Mitchell Douglas A, Shelest Ekaterina, Breitling Rainer, Takano Eriko, Lee Sang Yup, Weber Tilmann, Medema Marnix H. 2017. antiSMASH 4.0—improvements in chemistry prediction and gene cluster boundary identification. Nucleic Acids Research 45: W36–W41.

Bourque Guillaume, Burns Kathleen H, Gehring Mary, Gorbunova Vera, Seluanov Andrei, Hammell Molly, Imbeault Michaël, Izsvák Zsuzsanna, Levin Henry L, Macfarlan Todd S, Mager Dixie L, Feschotte Cédric. 2018. Ten things you should know about transposable elements. Genome Biology 19: 199.

Brehm Alexander, Tufteland Katharina R, Aasland Rein, Becker Peter B. 2004. The many colours of chromodomains. BioEssays 26: 133–140.

Buchfink Benjamin, Xie Chao, Huson Daniel H. 2014. Fast and sensitive protein alignment using DIAMOND. Nature Methods 12: 59–60.

Cam Hugh P, Noma Ken Ichi, Ebina Hirotaka, Levin Henry L, Grewal Shiv IS. 2008. Host genome surveillance for retrotransposons by transposon-derived proteins. Nature 451: 431–436.

Capella-Gutiérrez Salvador, Silla-Martínez José M, Gabaldón Toni. 2009. trimAl: A tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25: 1972–1973.

Chellapan Biju Vadakkemukadiyil, Van Dam Peter, Rep Martijn, Cornelissen Ben JC, Fokkens Like. 2016. Non-canonical Helitrons in Fusarium oxysporum. Mobile DNA 7: 1–16.

Cosby Rachel L, Judd Julius, Zhang Ruiling, Zhong Alan, Garry Nathaniel, Pritham Ellen J, Feschotte Cédric. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. Science 371: eabc6405.

Croteau Deborah L, Popuri Venkateswarlu, Opresko Patricia L, Bohr Vilhelm A. 2014. Human RecQ Helicases in DNA Repair, Recombination, and Replication. Annual Review of Biochemistry 83: 519–552.

Dhillon Braham, Cavaletto Jessica R, Wood Karl V, Goodwin Stephen B. 2010. Accidental Amplification and Inactivation of a Methyltransferase Gene Eliminates Cytosine Methylation in Mycosphaerella graminicola. Genetics 186: 67-U139.

Eddy Sean R. 2011. Accelerated profile HMM searches. PLoS Computational Biology 7.

Ejima Yosuke, Yang Lichun. 2003. Trans mobilization of genomic DNA as a mechanism for retrotransposon-mediated exon shuffling. Human Molecular Genetics 12: 1321–1328.

Emms David M, Kelly Steven. 2018. OrthoFinder: Phylogenetic orthology inference for comparative genomics. bioRxiv: 1–14.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends in Genetics 20: 417–423.

Gao Xiang, Hou Yi, Ebina Hirotaka, Levin Henry L, Voytas Daniel F. 2008. Chromodomains direct integration of retrotransposons to heterochromatin. Genome Research 18: 359–369.

Gentleman Robert. 2021. Category: Category Analysis.

Gladyshev Eugene. 2017. Repeat-Induced Point Mutation (RIP) and Other Genome Defense Mechanisms in Fungi. Microbiology Spectrum 5.

González-Sayer Sandra, Oggenfuss Ursula, García Ibonne, Aristizabal Fabio. 2021. High-quality genome assembly of Pseudocercospora ulei the main threat to natural rubber trees. Genetics and Molecular Biology: 0–1.

Grabundzija Ivana, Messing Simon A, Thomas Jainy, Cosby Rachel L, Bilic Ilija, Miskey Csaba, Gogol-Doring Andreas, Kapitonov Vladimir, Diem Tanja, Dalda Anna, Jurka Jerzy, Pritham Ellen J, Dyda Fred, Izsvak Zsuzsanna, Ivics Zoltan. 2016. A Helitron transposon reconstructed from bats reveals a novel mechanism of genome shuffling in eukaryotes. Nature Communications 7.

Gupta Smriti, Gallavotti Andrea, Stryker Gabrielle A, Schmidt Robert J, Lal Shailesh K. 2005. A novel class of Helitron- related transposable elements in maize contain portions of multiple pseudogenes. Plant Molecular Biology 57: 115–127.

Hammond Scott M. 2005. Dicing and slicing: The core machinery of the RNA interference pathway. FEBS Letters 579: 5822–5829.

Hane James K, Paxman Jonathan, Jones Darcy AB, Oliver Richard P, de Wit Pierre. 2020. "CATAStrophy," a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There? Frontiers in Microbiology 10.

Hane James K, Williams Angela H, Taranto Adam P, Solomon Peter S, Oliver Richard P. 2015. Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. In: 55–68.

Hartmann Fanny E, Sánchez-Vallet Andrea, McDonald Bruce A, Croll Daniel. 2017. A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. The ISME Journal 11: 1189–1204.

Hess Jaqueline, Skrede Inger, Wolfe Benjamin E, Butti Kurt La, Ohm Robin A, Grigoriev Igor V., Pringle Anne. 2014. Transposable element dynamics among asymbiotic and ectomycorrhizal amanita fungi. Genome Biology and Evolution 6: 1564–1578.

Hoede Claire, Arnoux Sandie, Moisset Mark, Chaumier Timothée, Inizan Olivier, Jamilloux Véronique, Quesneville Hadi. 2014. PASTEC: An automatic transposable element classification tool. PLoS ONE 9: 1–6.

Hoen Douglas R, Bureau Thomas E. 2015. Discovery of novel genes derived from transposable elements using integrative genomic analysis. Molecular Biology and Evolution 32: 1487–1506.

Holton Nicholas J, Goodwin Timothy JD, Butler Margaret I, Poulter RTM. 2001. An active retrotransposon in Candida albicans. Nucleic Acids Research 29: 4014–4024.

Jangam Diwash, Feschotte Cédric, Betrán Esther. 2017. Transposable Element Domestication As an Adaptation to Evolutionary Conflicts. Trends in Genetics 33: 817–831.

Jiang Ning, Bao Zhirong, Zhang Xiaoyu, Eddy Sean R, Wessler Susan R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.

Kapitonov Vladimir V., Jurka Jerzy. 2005. RAG1 core and V(D)J recombination signal sequences were derived from Transib transposons. PLoS Biology 3: 0998–1011.

Katoh Kazutaka, Standley Daron M. 2013. MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. Molecular Biology and Evolution 30: 772–780.

Keeling Patrick J, Palmer Jeffrey D. 2008. Horizontal gene transfer in eukaryotic evolution. Nature Reviews Genetics 9: 605–618.

Kojima Kenji K. 2020. AcademH, a lineage of Academ DNA transposons encoding helicase found in animals and fungi. Mobile DNA 11: 1–11.

Kolkman Joost A, Stemmer Willem PC. 2001. Directed evolution of proteins by exon shuffling. Nature Biotechnology 19: 423–428.

Koonin Eugene V, Makarova Kira S, Wolf Yuri I, Krupovic Mart. 2020. Evolutionary entanglement of mobile genetic elements and host defence systems: guns for hire. Nature Reviews Genetics 21: 119–131.

Lanfear Robert, Calcott Brett, Ho Simon YW, Guindon Stephane. 2012. PartitionFinder: Combined selection of partitioning schemes and substitution models for phylogenetic analyses. Molecular Biology and Evolution 29: 1695–1701.

Lei Bingkun, Capella Matías, Montgomery Sean A, Borg Michael, Osakabe Akihisa, Goiser Malgorzata, Muhammad Abubakar, Braun Sigurd, Berger Frédéric. 2021. A Synthetic Approach to Reconstruct the Evolutionary and Functional Innovations of the Plant Histone Variant H2A.W. Current Biology 31: 182-191.e5.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li Heng, Handsaker Bob, Wysoker Alec, Fennell Tim, Ruan Jue, Homer Nils, Marth Gabor, Abecasis Goncalo, Durbin Richard. 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics 25: 2078–2079.

Li Yuanning, Steenwyk Jacob L, Chang Ying, Wang Yan, James Timothy Y, Stajich E, Spatafora Joseph W, Groenewald Marizeth, Dunn Casey W, Todd Chris. 2020. A genome-scale phylogeny of Fungi; insights into early evolution, radiations, and the relationship between taxonomy and phylogeny. : 1–51.

Manni Mosè, Berkeley Matthew R, Seppey Mathieu, Simão Felipe A, Zdobnov Evgeny M. 2021. BUSCO Update: Novel and Streamlined Workflows along with Broader and Deeper Phylogenetic Coverage for Scoring of Eukaryotic, Prokaryotic, and Viral Genomes. Molecular Biology and Evolution 38: 4647–4654.

Marina Diana B, Shankar Smita, Natarajan Prashanthi, Finn Kenneth J, Madhani Hiten D. 2013. A conserved ncrna-binding protein recruits silencing factors to heterochromatin through an RNAi-independent mechanism. Genes and Development 27: 1851–1856.

Martin Eliza C, Vicari Célia, Tsakou-Ngouafo Louis, Pontarotti Pierre, Petrescu Andrei J, Schatz David G. 2020. Identification of RAG-like transposons in protostomes suggests their ancient bilaterian origin. Mobile DNA 11: 1–20.

Mateo Lidia, González Josefa. 2014. Pogo-Like transposases have been repeatedly domesticated into CENP-B-Related Proteins. Genome Biology and Evolution 6: 2008–2016.

McLysaght Aoife, Guerzoni Daniele. 2015. New genes from non-coding sequence: The role of de novo protein-coding genes in eukaryotic evolutionary innovation. Philosophical Transactions of the Royal Society B: Biological Sciences 370.

Minh Bui Quang, Schmidt Heiko A, Chernomor Olga, Schrempf Dominik, Woodhams Michael D, Von Haeseler Arndt, Lanfear Robert, Teeling Emma. 2020. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. Molecular Biology and Evolution 37: 1530–1534.

Mistry Jaina, Chuguransky Sara, Williams Lowri, Qureshi Matloob, Salazar Gustavo A, Sonnhammer Erik LL, Tosatto Silvio CE, Paladin Lisanna, Raj Shriya, Richardson Lorna J, Finn Robert D, Bateman Alex. 2021. Pfam: The protein families database in 2021. Nucleic Acids Research 49: D412–D419.

Mitchell Alex, Chang Hsin Yu, Daugherty Louise, Fraser Matthew, Hunter Sarah, Lopez Rodrigo, McAnulla Craig, McMenamin Conor, Nuka Gift, Pesseat Sebastien, Sangrador-Vegas Amaia, Scheremetjew Maxim, Rato Claudia, Yong Siew Yit, Bateman Alex, Punta Marco, Attwood Teresa K, Sigrist Christian JA, Redaschi Nicole, et al. 2015. The InterPro protein families database: The classification resource after 15 years. Nucleic Acids Research 43: D213–D221.

Morgan Martin, Falcon Seth, Gentleman Robert. 2021. GSEABase: Gene set enrichment data structures and methods.

Morgante Michele, Brunner Stephan, Pea Giorgio, Fengler Kevin, Zuccolo Andrea, Rafalski Antoni. 2005. Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. Nature Genetics 37: 997–1002.

Mühlhausen Stefanie, Findeisen Peggy, Plessmann Uwe, Urlaub Henning, Kollmar Martin. 2016. A novel nuclear genetic code alteration in yeasts and the evolution of codon reassignment in eukaryotes. Genome Research 26: 945–955.

Muszewska Anna, Steczkiewicz Kamil, Stepniewska-Dziubinska Marta, Ginalski Krzysztof. 2019. Transposable Elements Contribute To Fungal Genes and Impact Fungal Lifestyle. Scientific Reports 9: 1–10.

Nakagawa Hiromi, Lee Joon Kyu, Hurwitz Jerard, Allshire Robin C, Nakayama Jun Ichi, Grewal Shiv IS, Tanaka Katsunori, Murakami Yota. 2002. Fission yeast CENP-B homologs nucleate centromeric heterochromatin by promoting heterochromatin-specific histone tail modifications. Genes and Development 16: 1766–1778.

Nishida Hiromi. 2006. Detection and characterization of fungal-specific proteins in Saccharomyces cerevisiae. Bioscience, Biotechnology and Biochemistry 70: 2646–2652.

Omrane Selim, Audéon Colette, Ignace Amandine, Duplaix Clémentine, Aouini Lamia, Kema Gert, Walker Anne-Sophie, Fillinger Sabine. 2017. Plasticity of the MFS1 promoter leads to multi drug resistance in the wheat pathogen Zymoseptoria tritici. mSphere: 1–42.

Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, Fillinger S. 2015. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. Environmental Microbiology 17: 2805–2823.

Van Oss Stephen Branden, Carvunis Anne Ruxandra. 2019. De novo gene birth. PLoS Genetics 15: 1–23.

Pagès Hervé, Carlson Marc, Falcon Seth, Li Nianhua. 2021. AnnotationDbi: Manipulation of SQLite-based annotations in Bioconductor.

Potocki Leszek, Kuna Ewelina, Filip Kamila, Kasprzyk Beata, Lewinska Anna, Wnuk Maciej. 2019. Activation of transposable elements and genetic instability during long-term culture of the human fungal pathogen Candida albicans. Biogerontology 20: 457–474.

Pritham Ellen J, Putliwala Tasneem, Feschotte Cédric. 2007. Mavericks, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene 390: 3–17.

Quesneville Hadi. 2020. Twenty years of transposable element analysis in the Arabidopsis thaliana genome. Mobile DNA 11: 1–13.

Quinlan Aaron R, Hall Ira M. 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics 26: 841–842.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology 10: 417–430.

Selker E. 2002. 15 Repeat-induced gene silencing in fungi. In: Homology Effects. Elsevier, 439–450.

Selker EU, Garrett PW. 1988. DNA sequence duplications trigger gene inactivation in Neurospora crassa. Proceedings of the National Academy of Sciences of the United States of America 85: 6870–6874.

Shen Xing Xing, Opulente Dana A, Kominek Jacek, Zhou Xiaofan, Steenwyk Jacob L, Buh Kelly V., Haase Max AB, Wisecaver Jennifer H, Wang Mingshuang, Doering Drew T, Boudouris James T, Schneider Rachel M, Langdon Quinn K, Ohkuma Moriya, Endoh Rikiya, Takashima Masako, Manabe Ri ichiroh, Čadež Neža, Libkind Diego, et al. 2018. Tempo and Mode of Genome Evolution in the Budding Yeast Subphylum. Cell 175: 1533-1545.e20.

Shen Xing-Xing, Steenwyk Jacob L, LaBella Abigail L, Opulente Dana A, Zhou Xiaofan, Kominek Jacek, Li Yuanning, Groenewald Marizeth, Hittinger Chris Todd, Rokas Antonis. 2020. Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. Science Advances 6.

Skamnioti Pari, Furlong Rebecca F, Gurr Sarah J. 2008. The fate of gene duplicates in the genomes of fungal pathogens. Communicative & Integrative Biology 1: 196–198.

Smit Arian FA, Riggs Arthur D. 1996. Tiggers and other DNA transposon fossils in the human genome. Proceedings of the National Academy of Sciences of the United States of America 93: 1443–1448.

Stamatakis Alexandros. 2014. RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30: 1312–1313.

Stritt Christoph, Gordon Sean P, Wicker Thomas, Vogel John P, Roulin Anne C. 2017. Recent activity in expanding populations and purifying selection have shaped transposable element landscapes across natural accessions of the Mediterranean grass Brachypodium distachyon. Genome Biology and Evolution 10: 1–38.

Sugiyama Tomoyasu, Cam Hugh P, Sugiyama Rie, Noma Ken ichi, Zofall Martin, Kobayashi Ryuji, Grewal Shiv IS. 2007. SHREC, an Effector Complex for Heterochromatic Transcriptional Silencing. Cell 128: 491–504.

Wang Jianhua, Han Guan Zhu. 2021. Unearthing LTR Retrotransposon gag Genes Co-opted in the Deep Evolution of Eukaryotes. Molecular Biology and Evolution 38: 3267–3278.

Wang Li Gen, Lam Tommy Tsan Yuk, Xu Shuangbin, Dai Zehan, Zhou Lang, Feng Tingze, Guo Pingfan, Dunn Casey W, Jones Bradley R, Bradley Tyler, Zhu Huachen, Guan Yi, Jiang Yong, Yu Guangchuang. 2020. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. Molecular Biology and Evolution 37: 599–603.

Wells Jonathan N, Feschotte Cédric. 2020. A Field Guide to Transposable Elements. Annual Review of Genetics 54: 7–34.

Wicker Thomas, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982.

van Wyk Stephanie, Wingfield Brenda D, De Vos Lieschen, van der Merwe Nicolaas A, Steenkamp Emma T. 2021. Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota. Frontiers in Microbiology 11.

Xu Shuangbin, Dai Zehan, Guo Pingfan, Fu Xiaocong, Liu Shanshan, Zhou Lang, Tang Wenli, Feng Tingze, Chen Meijun, Zhan Li, Wu Tianzhi, Hu Erqiang, Jiang Yong, Bo Xiaochen, Yu Guangchuang. 2021. GgtreeExtra: Compact Visualization of Richly Annotated Phylogenetic Data. Molecular Biology and Evolution 38: 4039–4042.

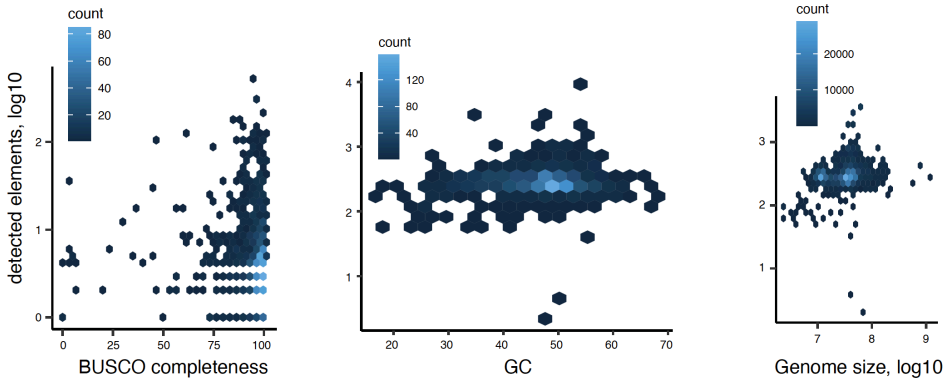
Yu Guangchuang, Smith David K, Zhu Huachen, Guan Yi, Lam Tommy Tsan Yuk. 2017. Ggtree: an R Package for Visualization and Annotation of Phylogenetic Trees With Their Covariates and Other Associated Data. Methods in Ecology and Evolution 8: 28–36.

Yurchenko Vyacheslav, Xue Zhu, Sadofsky Moshe. 2003. The RAG1 N-terminal domain is an E3 ubiquitin ligase. Genes and Development 17: 581–585.

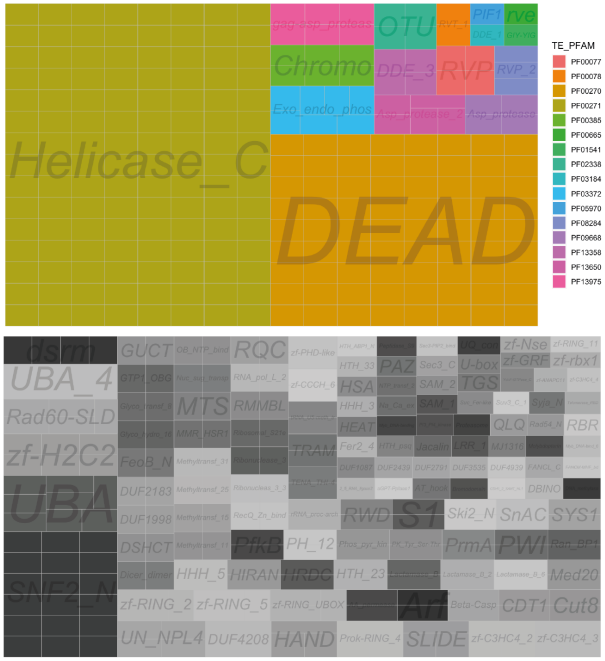
Zhang Han, Yohe Tanner, Huang Le, Entwistle Sarah, Wu Peizhi, Yang Zhenglu, Busk Peter K, Xu Ying, Yin Yanbin. 2018. DbCAN2: A meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Research 46: W95–W101.

Zhu Chun Xiang, Yan Lan, Wang Xiao Juan, Miao Qi, Li Xing Xing, Yang Feng, Cao Yong Bing, Gao Ping Hui, Bi Xin Ling, Jiang Yuan Ying. 2014. Transposition of the zorro2 retrotransposon is activated by miconazole in candida albicans. Biological and Pharmaceutical Bulletin 37: 37–43.

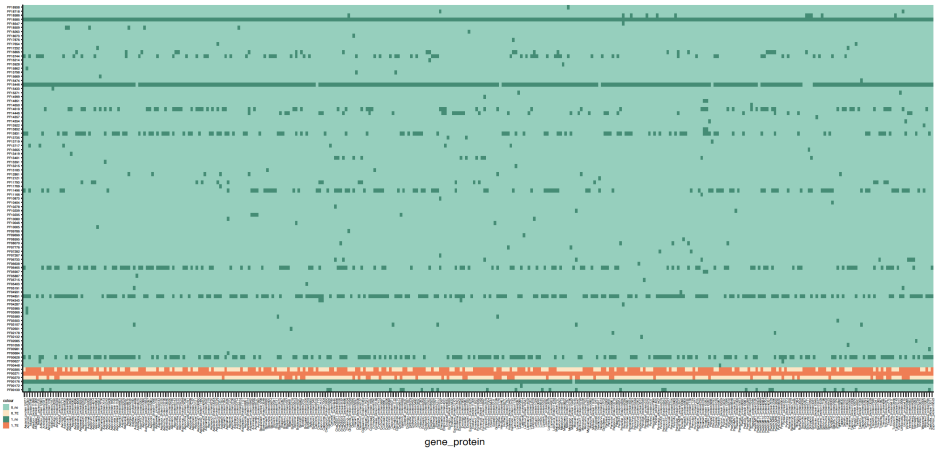
Supplementary Files



Supplementary Figure S1: Number of host-TE detection and genome characteristics: BUSCO completeness score, genome-wide GC content, genome size.



Supplementary Figure S2: Distribution of domains in the filtered set of candidates: After filtering for candidates that are present in at least 20 species, with at least 5 species being closely related. Distribution for the TE-derived domain and the non-TE derived domain.



Supplementary Figure S3: Additional domains in the host-TE fusion candidate SHREC: Dark orange indicates presence of a TE domain in a specific gene, bright orange indicates absence of a TE domain. Dark green indicates presence of a non-TE derived domain, bright green indicates its absence.

Supplementary Figure S4: Full multiple sequence alignment of all SCREC regions (only available via zenodo)

Supplementary File F1: Phylogenetic tree: phylogenetic tree of the fungal kingdom, based on 100 genes and in nexus format. The tree contains the following metadata: label (a short species ID), organism, species taxonomy ID, assembly name, assembly accession number, taxonomy ID, link to the genome in GenBank, phylum, class, order, family, genus, protein file, genome file, cds file, yeast (marked with 1 for yeast-like growing species), sequences counts, genome size (number of bases), average length, median length, maximum length, minimum length, N50, L50, BUSCO completeness score [%], BUSCO single copy genes [%], BUSCO fragmented genes [%], BUSCO missing genes [%], number of genes in the BUSCO code, BUSCO code, number of proteins detected, predicted lifestyle with CATAStrophy, phylum3 includes the differentiation between Saccharomycotina and the other classes of the phylum Ascomycota

General discussion

"In the shade of the sun
We'll bloom 'till we're young"
Kapitan Korsakov

In this doctoral thesis I investigated the impact of transposable elements (TEs) on the adaptive evolution of fungi. Genetic variation across the genome is a critical factor for the evolutionary potential, and TEs are important drivers of polymorphism within species. In the **first chapter**, we studied the population dynamics of TEs in the fungal plant pathogen *Zymoseptoria tritici*, using a large population genomics dataset. We found that TEs have recently been active in *Z. tritici*, with indications of strong purifying selection acting against most insertions. We found local bursts of new TE insertions strongly linked with a genome size expansion. While the additional length of TEs alone cannot solely explain the observed genome size expansion, we argue that TEs can also indirectly impact genome size evolution, for example by chromosomal rearrangements. We continued by analyzing expansion routes of TEs and the genomic environment within 19 *Z. tritici* telomere-to-telomere genomes for the **second chapter**. Using phylogenetic methods, we were able to reconstruct where bursts of proliferation started and in what types of genomic niches subsequent new TE copies inserted. While older TE copies are clustered in distinctive genomic niches with high TE contents, younger copies in recent bursts are more evenly distributed in the genome and often at a close distance to a gene. Simultaneously, TE copies in bursts show a reduced signature of defense mechanisms. We argue being close to a gene without having a deleterious impact might allow TEs to remain active and thus create new copies. For the **third chapter**, we used a dataset of 1,327 fungal species to uncover long-term impacts of TEs on fungal proteome evolution. We focused on fusions between existing genes and TE-derived domains. Helicases are the most prevalent TE-derived domains in host-TE fusions over the fungal kingdom. Functions of fusion genes include several already described genes involved in defense against viruses and TEs.

TE activity and short-term genome size expansion

TEs are important facilitators of short-term local adaptation. Most TEs in a genome are silenced and will not proliferate, but can be de-repressed under stress conditions (Horváth *et al.*, 2017). De-repression can be followed by bursts of proliferation. Most new TE insertions will have a neutral or deleterious effect, however, some new TE insertions can provide an adaptive value, including adaptation to a new environment (González *et al.*, 2008). Adaptive TE insertion loci will increase in frequency in the population. We used a short-read genome dataset covering 284 isolates from 6 globally distributed *Z. tritici* populations. Detecting TEs in short-read data is still challenging, as reads are not long enough to cover the full length of the TE. Consequently, most TE-containing reads cannot be annotated at the correct genomic position. In this project, we used an established method that is detecting insertions based on reads that cover both terminal regions of the TE and its surrounding region (Linheiro & Bergman, 2012; Nelson *et al.*, 2017). We validated the results with a new pipeline that confirms the presence or absence of TEs in a predefined locus in all isolates. Our pipeline provided us with a set of TE insertion loci that represent the most recently active elements, which allowed us to compare TE content between isolates and populations. Isolates from the center of origin of the species contain lower amounts of TEs, compared to more recently established populations. The short-term increase in TE content is strongly correlated with an increase in genome size. Ongoing TE activity most likely had both a direct and indirect impact on genome size expansion. Each TE insertion can directly increase the genome size to some extent, yet the total length of new TE copies cannot fully explain the increase in genome size we discovered. TEs can also indirectly increase genome size, by chromosomal rearrangements and partial duplications of chromosomes. Interestingly, TEs can also play a role in the decrease of genome size. By ectopic recombination, the region of a TE or between two TEs can be deleted (Devos *et al.*, 2002). Ectopic recombination might be important to remove TEs or larger regions in *Z. tritici* as well and could be a counter-balance to ongoing TE activity. While recent TE bursts can produce a large number of new copies in a short time, ectopic recombination is a random process that will not remove TEs at the same speed as insertions emerge. Additionally, purifying selection might act against most occurrences of ectopic recombination, as deletion of random regions can lead to deletion of essential genes. It is possible that the most recent populations with large genome size will have reduced genome sizes in the absence of ongoing TE bursts.

Expansion routes of TEs in the genomic ecosystem

The mechanisms of how TEs gain activity, create new copies and insert into new sites is already well established for the most prevalent TE superfamilies (Wicker *et al.*, 2007; Sultana *et al.*, 2017; Wells & Feschotte, 2020). Which TEs are activated and where they insert into is less well known. In a study of the invasion of a TE family after horizontal transfer, a short burst of TEs was detected after specific environmental conditions, while transposition rates were much lower in other conditions (Kofler *et al.*, 2018). Environmental stresses can have an impact on the de-repression of a TE, but the genomic environment might be important as well. The genome can be seen as an ecosystem consisting of different niches (Venner *et al.*, 2009). Genomic niches can differ in TE content, gene content, distance to the closest gene or in chromatin state. In some niches, the impact of new TE insertions might be close to neutral, yet the TE might be silenced more easily. Other niches might allow TEs to remain active, with the cost that strong purifying selection will remove the insertion from the population. In the first chapter we detected indication of strong purifying selection, as most TE insertions are only present in one isolate. In addition to purifying selection, TEs can also be removed by deletion, for example by ectopic recombination. Looking at only one reference genome would thus not provide the full picture of a TE expansion route. We analyzed 19 telomere-to-telomere genomes that are covering the global distribution of *Z. tritici* (Badet *et al.*, 2020). As described above, stress conditions can induce bursts of TE proliferation. Whether a burst is started by a general de-repression of TEs or a single TE that will then create new copies in a cascade-like way is not known. We used phylogenetic tools to establish the relationships between the individual copies of each TE family. Even for TE families with high copy numbers we detected few recent bursts clades per family. We compared different characteristics of both the TE and the niches they were inserted into, and reconstructed the ancestral state to see if and how a characteristic emerged. Most TEs are located on a different chromosome than their closest relatives, indicating the insertion site is independent from the original copy. We found most older TEs placed in specific niches with high TE content and few genes. In contrast to older elements, TEs in recent burst clades show an increase in GC content compared to all other elements of the TE family. An increase in GC content could indicate the absence of an Ascomycete specific defense mechanism called RIP (repeat-induced point mutations), which introduces directed point mutations, mostly CpG -> TpA (Galagan & Selker, 2004). RIP might

not be triggered in recent bursts yet, which is surprising, as most bursts predate the separation of the 19 isolates, and RIP has been described to have a very strong impact from the first round of sexual recombination in *Neurospora crassa* (Wang *et al.*, 2020b). In contrast, RIP left strong signals on older TE copies, and the genes that are thought to be necessary for RIP are present in *Z. tritici* (van Wyk *et al.*, 2021). Yet there are some indications, that RIP might have been lost (Lorrain *et al.*, 2021a), or RIP might simply not be as effective as in *N. crassa* (Thomas Badet, personal communication). Escape of RIP might be a reason why we found strong TE bursts and genome size expansions in the first chapter as well.

In addition to a tendency for a higher GC content, TEs at the start of a burst are closer located to genes. A close distance of a TE to a gene can in many cases be deleterious, either by disrupting coding regions or by changing the expression of the gene. Generally, TE copies close to genes will strongly be selected against. Yet, a close distance to a gene might prevent the TE from being silenced or deleted, as a defense against the TE can impact the gene as well (Fudal *et al.*, 2009). We mostly detected small non-autonomous MITEs close to genes, which correlates with the idea that smaller TEs are less deleterious (Barrón *et al.*, 2014). Consequently, niches close to genes might even serve as protective habitat for certain TE families to proliferate. And as only TEs that provide a beneficial impact will survive over a long period of time, the very TEs that help their host in local adaptation might be responsible for potentially devastating bursts of proliferation as well.

Domestication of TEs in fungi

The proteome of a species evolves continuously, and novel gene functions can be gained. Gene evolution is predominantly based on the recombination and re-shuffling of pre-existing open reading frames (ORFs) (Andersson *et al.*, 2015). TEs play many roles in the evolution of genes, and can provide ORFs themselves as well, leading to fusions between existing genes and TEs. Host-TE fusions have been detected in animals and plants (Bennetzen, 2000; Cosby *et al.*, 2021), and there are indications of host-TE fusions in the fungal kingdom as well (Muszewska *et al.*, 2019). The exact mechanisms in how TEs fuse with genes is not entirely clear, but TEs likely either insert into existing genes or capture gene fragments, for example by the TE superfamilies *Helitron* and *Pack-Mule* (Jiang *et al.*, 2004; Barbaglia *et al.*, 2012). Gene fragment capture could serve the TE evolution rather than gene evolution, leading to new TE

subfamilies, but they might also become host-TE fusions. Both TE insertion and gene capturing might happen rather frequently on species-level, but are likely to have a neutral or even negative impact. Only beneficial host-TE fusions will remain over longer evolutionary times, and their presence in several species that share an ancestor indicate a benefit.

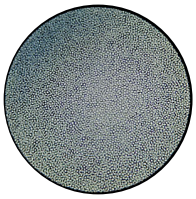
To detect TE-derived fusions with a beneficial impact beyond the species level, we analyzed the annotated genomes of 1,327 fungal species. We filtered for genes that include TE-derived domains and domains that have no relation with TEs that are present in at least 20 fungal species. We detected 241 very stable host-TE fusion candidates. Predominantly, the TE-derived domain is a helicase, derived either from the superfamily *AcademH* or *KolobokH* (Muszewska *et al.*, 2019). We could not confirm or deny the true TE origin of helicase domains. The helicase domain has a binding function to either DNA or RNA and is thus a universally useful function. On interspecies level, many fusions containing a helicase domain show strong presence/absence polymorphism for additional domains. Yet, most host-TE fusions show no or only one duplication event per species, indicating they are not functional TEs anymore. Containing a helicase domain might enable the host-TE fusion to continue capturing genes, gain additional functions, and to continue to diversify as a gene. Helicases that still continue capturing genes would not be stable host-TE fusions, but could lead to a high variability of new helicase derived host-TE fusions. Some of the detected host-TE fusions have already been described in the literature, either on fungal level or in plant or animals, including dicer dimerization and the SHREC complex, which are both involved in defense against viruses and TE silencing, and contain the respective TE-derived domains (Hammond, 2005; Sugiyama *et al.*, 2007). As these genes not only exist in fungi, the origin of host-TE fusions might be older than the separation of the fungal kingdom.

Yet, both dicer dimerization and SHREC complex were not detected in the Ascomycete class of Saccharomycotina. So far described species in Saccharomycotina are also missing other defense mechanisms against TEs, including RIP (van Wyk *et al.*, 2021). Saccharomycotina also have a rather low number of TEs, a reduced number of genes generally small genome sizes (Bleykasten-Grosshans & Neuvéglise, 2011; Shen *et al.*, 2020). Intriguingly, we found the proportion of genes that contain a TE-derived domain to be increased in Saccharomycotina compared to other fungi. With a strong selection pressure on both TEs and genes, it is likely that host-TE fusions are better suited as genes in Saccharomycotina. Saccharomycotina might have evolved individual defense mechanisms against TEs.

Outlook

This thesis underlines the importance of TEs in short-term adaptation and long-term evolution of fungal genomes. Our methods included several novel approaches in combining large datasets with many individuals per species, many TE copies per genome or many species per kingdom to compare. We improved the understanding of the role that TEs play on different scales, from local adaptation to TE evolution in a species to the impact of TEs on the evolution of the fungal kingdom. At the same time, this thesis opens many new questions in how TEs evolve in fungi.

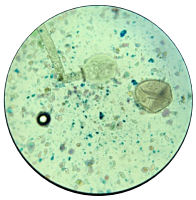
TE diversity and TE content and the evolution of the host



TE diversity and TE content in fungi are species specific and can even vary between isolates of the same species (Raffaele & Kamoun, 2012; Badet *et al.*, 2020). TE activity has been hypothesized to be involved in speciation by inducing reproductive isolation (Serrato-Capuchina & Matute, 2018). It is likely that TE content and diversity will diverge after speciation, and random processes can have a strong impact in the fate of TEs in a genome. By analyzing phylogenetic relationships, we detected divergences of sequences of recently active TE families in the second chapter. Low copy TE families can become extinct, remain at low copies or can rapidly increase in copy numbers after de-repression, while the sequences continue to evolve until they are not recognized as the same family anymore (Bleykasten-Grosshans *et al.*, 2021). Environmental stressors either after the migration of the species to different climates or after changes in the climate can impact fungi via the host. A change of host species or cultivars can also have dramatic effects, and forces the fungus to evolve new effectors in plant pathogens, or adapt to the high core body temperature for mammalian hosts (Leach *et al.*, 2012; Richards *et al.*, 2019). Other stressors target the fungus directly, for example fungicides. All these stressors can induce TE activity, most likely leading to many new deleterious and neutral TE insertions (Horváth *et al.*, 2017). Yet, in some cases new TE insertions can lead to local adaptation. Some TE-derived adaptations have been shown to be

triggered by insertion of different TEs in the same locus, as seen in fungicide resistance in *Z. tritici* (Omrane *et al.*, 2015, 2017; Mäe *et al.*, 2020). Artificial de-repression and activation of TEs is a very helpful source for future plant breeding (Thieme *et al.*, 2017; Thieme & Bucher, 2018), but could also be used to better understand fungal pathogens.

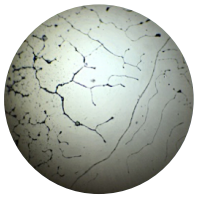
How do Saccharomycotina deal with TEs?



Different defense mechanisms are acting against TEs and viruses. Defenses include RNA interference, where siRNA is leading to cleavage of the RNA, silencing by RdRp, DNA methylation, histone methylation or repeat induced point (RIP) mutation, a fungus-specific defense that introduces a higher mutation rate during sexual recombination (Galagan & Selker, 2004; Slotkin & Martienssen, 2007). TEs have been shown to be involved in viral and TE regulation as well, for example via the domestication of *gag* (capsid protein) and *env* (envelope protein) of retrotransposons or by copy number control (Saha *et al.*, 2015; Naville *et al.*, 2016). In the 3rd chapter, we found more potential examples of TEs being involved in viral and TE regulation as part of host-TE fusions. TEs are involved as part of host-TE fusions in Dicer dimerization and SHREC, which are potential defense mechanisms against viruses and TEs, and present beyond the fungal kingdom (Hammond, 2005; Sugiyama *et al.*, 2007). Interestingly, the presence of these potential defense mechanism is very patchy along the fungal phylogeny, especially in Ascomycetes, indicating independent losses. If the mechanisms are still functional is another question. As many Ascomycetes contain RIP, other defense mechanisms might have become redundant. RIP itself was also actively involved in the destruction of a defense in *Z. tritici* by mutating all copies of a methyltransferase after duplication, rendering cytosine methylation unfunctional (Dhillon *et al.*, 2010). Saccharomycotina, a class of Ascomycetes not only is lacking most of the generally known defense mechanisms against TEs, but also has no indication of RIP in any of the species (van Wyk *et al.*, 2021). The absence of most known defense mechanisms did not lead to burst of TEs in Saccharomycotina. On the contrary, species of Saccharomycotina contain very low TE contents, while there is still indication that TEs are still active and even induce adaptation (Zhu *et al.*, 2014; Maxwell, 2020). These species might be very resistant against TEs, via not yet described defense mechanisms (Fouché *et al.*, 2021). Some Saccharomycetes are using the

genetic code slightly different (Butler *et al.*, 2009), which might act as a defense mechanism against viruses, and subsequently against TEs. Research on the activity and regulation of TEs in Saccharomycotina will help to detect new defense strategies against TEs.

Literature discussion and outlook



Andersson Dan I, Jerlström-Hultqvist Jon, Näsvall Joakim. 2015. Evolution of New Functions De Novo and from Preexisting Genes. Cold Spring Harbor Perspectives in Biology 7: a017996.

Badet Thomas, Oggenfuss Ursula, Abraham Leen, McDonald Bruce A, Croll Daniel. 2020. A 19-isolate reference-quality global pangenome for the fungal wheat pathogen Zymoseptoria tritici. BMC Biology 18: 12.

Barbaglia Allison M, Klusman Katarina M, Higgins John, Shaw Janine R, Hannah L Curtis, Lal Shailesh K. 2012. Gene capture by Helitron transposons reshuffles the transcriptome of maize. Genetics 190: 965–975.

Barrón Maite G, Fiston-Lavier Anna-Sophie, Petrov Dmitri A, González J. 2014. Population Genomics of Transposable Elements in Drosophila (BL Bassler, Ed.). Annual Review of Genetics 48: 561–581.

Bennetzen Jeffrey L. 2000. Transposable element contributions to plant gene and genome evolution. Plant Molecular Biology 42: 251–269.

Bleykasten-Grosshans Claudine, Fabrizio Romeo, Friedrich Anne, Schacherer Joseph. 2021. Species-Wide Transposable Element Repertoires Retrace the Evolutionary History of the Saccharomyces cerevisiae Host . Molecular Biology and Evolution 38: 4334–4345.

Bleykasten-Grosshans Claudine, Neuvéglise Cécile. 2011. Transposable elements in yeasts. Comptes Rendus Biologies 334: 679–686.

Butler Geraldine, Rasmussen Matthew D, Lin Michael F, Santos Manuel AS, Sakthikumar Sharadha, Munro Carol A, Rheinbay Esther, Grabherr Manfred, Forche Anja, Reedy Jennifer L, Agrafioti Ino, Arnaud Martha B, Bates Steven, Brown Alistair JP, Brunke Sascha, Costanzo Maria C, Fitzpatrick David A, de Groot Piet WJ, Harris David, et al. 2009. Evolution of pathogenicity and sexual reproduction in eight Candida genomes. Nature 459: 657–662.

Cosby Rachel L, Judd Julius, Zhang Ruiling, Zhong Alan, Garry Nathaniel, Pritham Ellen J, Feschotte Cédric. 2021. Recurrent evolution of vertebrate transcription factors by transposase capture. Science 371: eabc6405.

Devos Katrien M, Brown James KM, Bennetzen Jeffrey L. 2002. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. Genome Research 12: 1075–1079.

Dhillon Braham, Cavaletto Jessica R, Wood Karl V, Goodwin Stephen B. 2010. Accidental Amplification and Inactivation of a Methyltransferase Gene Eliminates Cytosine Methylation in Mycosphaerella graminicola. Genetics 186: 67-U139.

Fouché Simone, Oggenfuss Ursula, Chanclud Emilie, Croll Daniel. 2021. A devil's bargain with transposable elements in plant pathogens. Trends in Genetics: 1–9.

Fudal Isabelle, Ross Simon, Brun Hortense, Besnard Anne Laure, Ermel Magali, Kuhn Marie Line, Balesdent Marie Hélène, Rouxel Thierry. 2009. Repeat-Induced Point Mutation (RIP) as an alternative mechanism of evolution toward virulence in leptosphaeria maculans. Molecular Plant-Microbe Interactions 22: 932–941.

Galagan JE, Selker EU. 2004. RIP: the evolutionary cost of genome defense. Trends in Genetics 20: 417–423.

González J, Lenkov K, Lipatov M, Macpherson JM, Petrov DA. 2008. High Rate of Recent Transposable Element-Induced Adaptation in Drosophila melanogaster. Plos Biology 6: 2109–2129.

Hammond Scott M. 2005. Dicing and slicing: The core machinery of the RNA interference pathway. FEBS Letters 579: 5822–5829.

Horváth Vivien, Merenciano Miriam, González J. 2017. Revisiting the Relationship between Transposable Elements and the Eukaryotic Stress Response. Trends in Genetics 33: 832–841.

Jiang Ning, Bao Zhirong, Zhang Xiaoyu, Eddy Sean R, Wessler Susan R. 2004. Pack-MULE transposable elements mediate gene evolution in plants. Nature 431: 569–573.

Kofler Robert, Senti Kirsten-Andre, Nolte Viola, Tobler Ray, Schlötterer Christian. 2018. Molecular dissection of a natural transposable element invasion. Genome Research: gr.228627.117.

Leach Michelle D, Tyc Katarzyna M, Brown Alistair JP, Klipp Edda. 2012. Modelling the regulation of thermal adaptation in candida albicans, a major fungal pathogen of humans. PLoS ONE 7: 1–14.

Linheiro Raquel S, Bergman Casey M. 2012. Whole Genome Resequencing Reveals Natural Target Site Preferences of Transposable Elements in Drosophila melanogaster (Jason E Stajich, Ed.). PLOS ONE 7: e30008.

Lorrain Cecile, Feurtey Alice, Ller Mareike Mo, Haueisen Janine, Stukenbrock Eva. 2021. Dynamics of transposable elements in recently diverged fungal pathogens: Lineage-specific transposable element content and efficiency of genome defenses. G3: Genes, Genomes, Genetics 11: 0–12.

Mäe Andres, Fillinger Sabine, Sooväli Pille, Heick Thies Marten. 2020. Fungicide Sensitivity Shifting of Zymoseptoria tritici in the Finnish-Baltic Region and a Novel Insertion in the MFS1 Promoter. Frontiers in Plant Science 11: 1–10.

Maxwell Patrick H. 2020. Diverse transposable element landscapes in pathogenic and nonpathogenic yeast models: The value of a comparative perspective. Mobile DNA 11: 1–26.

Muszewska Anna, Steczkiewicz Kamil, Stepniewska-Dziubinska Marta, Ginalski Krzysztof. 2019. Transposable Elements Contribute To Fungal Genes and Impact Fungal Lifestyle. Scientific Reports 9: 1–10.

Naville M, Warren IA, Haftek-Terreau Z, Chalopin D, Brunet F, Levin P, Galiana D, Volff JN. 2016. Not so bad after all: retroviruses and long terminal repeat retrotransposons as a source of new genes in vertebrates. Clinical Microbiology and Infection 22: 312–323.

Nelson Michael G, Linheiro Raquel S, Bergman Casey M. 2017. McClintock: An Integrated Pipeline for Detecting Transposable Element Insertions in Whole-Genome Shotgun Sequencing Data. G3: Genes|Genomes|Genetics 7: 2763–2778.

Omrane Selim, Audéon Colette, Ignace Amandine, Duplaix Clémentine, Aouini Lamia, Kema Gert, Walker Anne-Sophie, Fillinger Sabine. 2017. Plasticity of the MFS1 promoter leads to multi drug resistance in the wheat pathogen Zymoseptoria tritici. mSphere: 1–42.

Omrane S, Sghyer H, Audeon C, Lanen C, Duplaix C, Walker AS, Fillinger S. 2015. Fungicide efflux and the MgMFS1 transporter contribute to the multidrug resistance phenotype in Zymoseptoria tritici field isolates. Environmental Microbiology 17: 2805–2823.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nature Reviews Microbiology 10: 417–430.

Richards Jonathan K, Stukenbrock EH, Carpenter Jessica, Liu Zhaohui, Cowger Christina, Faris Justin D, Friesen Timothy L. 2019. Local adaptation drives the diversification of effectors in the fungal wheat pathogen Parastagonospora nodorum in the United States (Daniel R Matute, Ed.). PLOS Genetics 15: e1008223.

Saha Agniva, Mitchell Jessica A, Nishida Yuri, Hildreth Jonathan E, Ariberre Joshua A, Gilbert Wendy V., Garfinkel David J. 2015. A trans -Dominant Form of Gag Restricts Ty1 Retrotransposition and Mediates Copy Number Control . Journal of Virology 89: 3922–3938.

Serrato-Capuchina Antonio, Matute Daniel R. 2018. The role of transposable elements in speciation. Genes 9.

Shen Xing-Xing, Steenwyk Jacob L, LaBella Abigail L, Opulente Dana A, Zhou Xiaofan, Kominek Jacek, Li Yuanning, Groenewald Marizeth, Hittinger Chris Todd, Rokas Antonis. 2020. Genome-scale phylogeny and contrasting modes of genome evolution in the fungal phylum Ascomycota. Science Advances 6.

Slotkin RK, Martienssen R. 2007. Transposable elements and the epigenetic regulation of the genome. Nature Reviews Genetics 8: 272–285.

Sugiyama Tomoyasu, Cam Hugh P, Sugiyama Rie, Noma Kenichi, Zofall Martin, Kobayashi Ryuji, Grewal Shiv IS. 2007. SHREC, an Effector Complex for Heterochromatic Transcriptional Silencing. Cell 128: 491–504.

Sultana T, Zamborlini A, Cristofari G, Lesage P. 2017. Integration site selection by retroviruses and transposable elements in eukaryotes. Nat Rev Genet 18: 292–308.

Thieme Michael, Bucher Etienne. 2018. Transposable Elements as Tool for Crop Improvement. Elsevier Ltd.

Thieme Michael, Lanciano Sophie, Balzergue Sandrine, Daccord Nicolas, Mirouze Marie, Bucher Etienne. 2017. Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. Genome Biology 18: 1–10.

Venner Samuel, Feschotte Cédric, Biémont Christian. 2009. Dynamics of transposable elements: towards a community ecology of the genome. Trends in Genetics 25: 317–323.

Wang Long, Sun Yingying, Sun Xiaoguang, Yu Luyao, Xue Lan, He Zhen, Huang Ju, Tian Dacheng, Tian Dacheng, Hurst Laurence D, Yang Sihai, Yang Sihai. 2020. Repeat-induced point mutation in Neurospora crassa causes the highest known mutation rate and mutational burden of any cellular life. Genome Biology 21: 1–23.

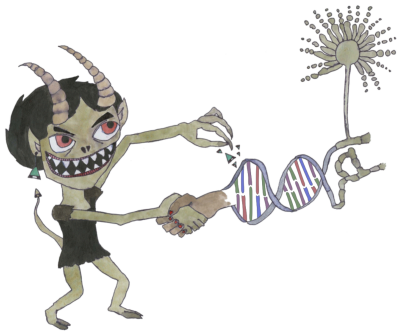
Wells Jonathan N, Feschotte Cédric. 2020. A Field Guide to Transposable Elements. Annual Review of Genetics 54: 7–34.

Wicker Thomas, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, Paux E, SanMiguel P, Schulman AH. 2007. A unified classification system for eukaryotic transposable elements. Nature Reviews Genetics 8: 973–982.

van Wyk Stephanie, Wingfield Brenda D, De Vos Lieschen, van der Merwe Nicolaas A, Steenkamp Emma T. 2021. Genome-Wide Analyses of Repeat-Induced Point Mutations in the Ascomycota. Frontiers in Microbiology 11.

Zhu Chun Xiang, Yan Lan, Wang Xiao Juan, Miao Qi, Li Xing Xing, Yang Feng, Cao Yong Bing, Gao Ping Hui, Bi Xin Ling, Jiang Yuan Ying. 2014. Transposition of the zorro2 retrotransposon is activated by miconazole in candida albicans. Biological and Pharmaceutical Bulletin 37: 37–43.

Acknowledgements



I am deeply grateful to all the amazing people that accompanied me on my PhD journey, as supervisors, co-workers, collaborators, friends, students. Science cannot be done in isolation, and I was lucky to be surrounded by the most inspiring people. My sincerest thanks go to:

My supervisor Prof. Dr Daniel Croll. Thank you for taking a chance with me since our first discussion whether viruses are alive or not. I would not have been able to go through this journey without your inspiration, guidance, motivation, kindness, trust and openness to new ideas. Working in your lab was a privilege. Thank you for being the most amazing supervisor and mentor I could have ever asked for!

Prof. Dr Anna Selmecki, Prof. Dr Anne Roulin and Prof. Dr Pilar Junier for being an amazing PhD committee! I enjoyed talking with you about my project and beyond. Thank you for the inspiration and encouragement! A special thanks to Anna Selmecki for giving me the chance to work with you! Thinking about this new project was an important motivation to finish my PhD. Thanks also to PD Dr Thomas Wicker and Dr Andrea Sanchez-Vallet for being part of my mid-thesis committee. Special thanks to PD Dr Thomas Wicker for all the help with TE classification and inspiring talks about Helitrons and RIP, and a different perspective on bioinformatics.

Dr Carolina Cornejo and Dr Martina Peter for your wonderful mentoring, trust and support on my way to start a PhD and beyond. Without your feedback, I would not only have lost a great master project and internship, but I would not have considered doing a PhD. Your mentoring helped me to become more independent and confident.

Dr Simone Fouché, my TE-sister, and biggest inspiration during my PhD! You helped me to always find my enthusiasm again. I miss our TE-tea in the office or train, but look forward to continuing talking about science with you. Thank you also for being there for me in times that were not so easy, and for always finding the right words and solutions!

Dr Vinciane Mossion for being the most amazing co-teacher and inspiring naturalist. “C’est belle!”. I learned so much from you, about interacting with students, histology, botany, and becoming more resilient in complicated situations. Thank you for making our lab eco-friendlier, and also thank you for entrusting me your fern babies for a summer!

Dr Thomas Badet for being an amazing co-author and mentor, thank you for all the great discussions, for always quickly sharing mountains of data for my 3rd chapter, and all the feedback! Your calm and kindness, and huge knowledge and curiosity made everything so much easier for me!

All the past and present members of the laboratory of Evolutionary Genetics who created a welcoming and inspiring environment that allowed me as an introvert to get out of my shell. Especially to Dr Nikhil Kumar Singh, my great office mate for all the discussions on science, philosophy and life. To Leen Abraham, Luzia Stadler and Dr Sabina Tralamazza for creating a wonderful lab atmosphere. To Dr Emilie Chanclud, working and discussing with you was always an inspiration and I always walked away smarter and sparks for new ideas. To Dr Sandra Milena González-Sayer for involving me in your amazing project, for the enthusiasm and to be a friend beyond continents. To Sarai Reyes for your huge enthusiasm, contagious cheerfulness and science inspiration, I'm looking up to you in so many ways. To Dr Alice Feurtey, Dr Cecile Lorrain and Dr Emile Gluck-Thaler for great discussions about weird TEs and much appreciated feedback. To Guido Pucchetti for inspiring TEs across Europe and talking about fish. To Hadjer Bellah, for all your kindness and support. To Hanna Glad for bringing new perspectives in both bioinformatics and imaging to the lab, and for being a great and inspiring office mate. To Tina Schwendener for creative new insights into my work.

To my (main) collaborators Dr Danilo Pereira, Dr Sabina Tralamazza, David E Torres, Dr Sandra Milena González-Sayer, Leen Abraham and Luzia Stadler who invited me to work on fascinating projects, each opening a new TE world or a different fungal pathogen genome and interesting new questions. Thank you for your enthusiasm!

To the TE community as a whole. Thank you for being so welcoming, kind and genuinely curious. It is amazing to be part of this community from early on in my PhD, but having virtual support with easy access to knowledge and discussions was amazing, especially during the pandemic! Special thanks to Ilya Kirov for organizing the TExJC, Alexander Suh and team for the Uppsala Transposon Symposium and the TEhub initiative for helping us to keep up with literature and methods.

To Dennis for always listening, never getting tired of endless fish-watching and saving my glasses from the bottom of the lake! To my friends, especially to Lea and Liyas, Blina, Paula and Carmen, Mia and Lex - you make my world a bright, colorful and fantastic place, you inspire me and give me comfort! Danke, ihr liebste Liebelottas! Uf wiiteri Jahr voller Glücksmomänt mit ünzne, tagediebe und umelümmle.

Annex: Figure sources

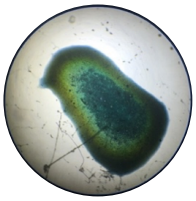


Figure 1 Introduction: Photography of Barbara McClintock by David Miklos, Cold Spring Harbor Laboratory Archives.

Figure 2 introduction: Figure taken from Wells & Feschotte, 2020

All other figures, spore prints, microscopic pictures and photographs are made by Ursula Oggenfuss

"Whether in forests, labs or kitchens, fungi have changed my understanding of how life happens"

Merlin Sheldrake, Entangled Life