



**HAL**  
open science

# All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity

Nathan Papon, Pauline Lasserre-Zuber, H el ene Rimbart, Romain de Oliveira, Etienne Paux, Fr ed eric Choulet

## ► To cite this version:

Nathan Papon, Pauline Lasserre-Zuber, H el ene Rimbart, Romain de Oliveira, Etienne Paux, et al.. All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity. *Plant Genome*, 2023, 10.1002/tpg2.20347 . hal-04143737

**HAL Id: hal-04143737**

**<https://hal.inrae.fr/hal-04143737v1>**

Submitted on 27 Jun 2023

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.



L'archive ouverte pluridisciplinaire **HAL**, est destin ee au d ep ot et  a la diffusion de documents scientifiques de niveau recherche, publi es ou non,  emanant des  tablissements d'enseignement et de recherche fran ais ou  trangers, des laboratoires publics ou priv es.



Distributed under a Creative Commons Attribution 4.0 International License

## ORIGINAL ARTICLE

# All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity

Nathan Papon | Pauline Lasserre-Zuber | H el ene Rimbert  | Romain De Oliveira | Etienne Paux | Fr ed eric Choulet 

INRAE, GDEC, Universit e Clermont Auvergne, Clermont-Ferrand, France

## Correspondence

Fr ed eric Choulet, INRAE, GDEC, Universit e Clermont Auvergne, 63000, Clermont-Ferrand, France.  
Email: [frederic.choulet@inrae.fr](mailto:frederic.choulet@inrae.fr)

## Present address

Romain De Oliveira, Gencoverly, Lyon, France.

## Present address

Etienne Paux, VetAgro Sup Campus agronomique, Lempdes, France.

Assigned to Associate Editor Nils Stein.

## Funding information

French Ministry of Higher Education, Research and Innovation (MESRI)

## Abstract

Bread wheat (*Triticum aestivum* L.) is a major crop and its genome is one of the largest ever assembled at reference-quality level. It is 15 Gb, hexaploid, with 85% of transposable elements (TEs). Wheat genetic diversity was mainly focused on genes and little is known about the extent of genomic variability affecting TEs, transposition rate, and the impact of polyploidy. Multiple chromosome-scale assemblies are now available for bread wheat and for its tetraploid and diploid wild relatives. In this study, we computed base pair-resolved, gene-anchored, whole genome alignments of A, B, and D lineages at different ploidy levels in order to estimate the variability that affects the TE space. We used assembled genomes of 13 *T. aestivum* cultivars (6x = AABBDD) and a single genome for *Triticum durum* (4x = AABB), *Triticum dicoccoides* (4x = AABB), *Triticum urartu* (2x = AA), and *Aegilops tauschii* (2x = DD). We show that 5%–34% of the TE fraction is variable, depending on the species divergence. Between 400 and 13,000 novel TE insertions per subgenome were detected. We found lineage-specific insertions for nearly all TE families in di-, tetra-, and hexaploids. No burst of transposition was observed and polyploidization did not trigger any boost of transposition. This study challenges the prevailing idea of wheat TE dynamics and is more in agreement with an equilibrium model of evolution.

## 1 | INTRODUCTION

Transposable elements (TEs) are key factors of genome evolution and their contribution to plant phenotypic variations and adaptation were shown in many studies (for review, see Baduel & Quadrana, 2021; Lisch, 2013). They are particularly prominent in the genomes of *Triteae*, a tribe of monocot plants encompassing important crops like wheat, barley, and rye, which diverged from a common ancestor

**Abbreviations:** Aet, *Aegilops tauschii*; CDS, coding sequence; CS, Chinese Spring; HC, high confidence; ISBP, insertion site-based polymorphism; LTR, long terminal repeat; MYA, million years ago; Myr, million year; nt, nucleotide; oIGR, orthologous intergenic region; PAV, presence-absence variation; SNP, single nucleotide polymorphism; SV, structural variation; Tae, *Triticum aestivum*; Tdi, *Triticum dicoccoides*; Tdu, *Triticum durum*; TE, transposable element; TIP, transposon insertion polymorphism; TSD, target site duplication; Tur, *Triticum urartu*.

This is an open access article under the terms of the [Creative Commons Attribution](https://creativecommons.org/licenses/by/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

  2023 The Authors. *The Plant Genome* published by Wiley Periodicals LLC on behalf of Crop Science Society of America.

~13 million years ago (MYA). *Triticeae* genomes contain millions of copies of TEs, making them a good model to understand the dynamics of TEs in complex genomes, their contribution to structural variations (SVs), and species adaptation. Bread wheat (*Triticum aestivum*) is the most widely grown crop on earth. Its genome is hexaploid and was shaped by two successive events of allopolyploidization (reviewed recently in Levy & Feldman, 2022) that brought together three related diploid subgenomes called A, B, and D, originated from three diploid species whose common ancestor was estimated around 6 MYA (Avni et al., 2022; Glémin et al., 2019; Li et al., 2022; Marcussen et al., 2014; Middleton et al., 2014). They share a common karyotype with seven pairs of chromosomes representing around 5 Gb each. A first allopolyploidization event occurred ~0.8 MYA between A and B lineages, and a second event occurred with the D lineage ~0.01 MYA during the period of wheat domestication. The closest living representative diploid species of AA, BB, and DD are *Triticum urartu* (AA), *Aegilops speltoides* (BB), and *Aegilops tauschii* (DD) whose divergence time with A, B, and D wheat subgenomes were estimated to 1.3, 4.4, and 0.01 MYA, respectively. Tetraploid wild emmer wheat is *Triticum turgidum* ssp. *dicoccoides* (AABB). It evolved to domesticated emmer wheat which is at the origin of cultivated tetraploid *Triticum turgidum* ssp. *durum* (AABB) and also hybridized with *Ae. tauschii* 0.01 MYA to give rise to cultivated hexaploid *T. aestivum* (AABBDD).

The large size of its genome (~5 Gb per haploid subgenome) is mainly explained by a massive TE content, 85%, meaning that the intergenic regions are composed of large clusters of TEs inserted into each other. This feature has limited for decades our capability to assemble them and, thus, characterize their content, organization, and diversity. Before assessing a complete genome sequence, studies showed high variation and sub-genome specificity of TEs (Yaakov, Ben-David, et al., 2013; Yaakov, Meyer, et al., 2013). Evidence of TE mobilization in newly formed wheat polyploids was described for a few families as well as changes in the epigenetic status (Kraitshtein et al., 2010; Yaakov & Kashkush, 2011, 2012). These studies tended to conclude that polyploidy was a genomic shock and suggest that TEs evolve by bursts where a few families escape to silencing and expand massively in a short time period, leading to rapid genome diversification. In 2018, a reference-quality genome assembly was produced for the hexaploid cultivar Chinese Spring (IWGSC, 2018) and a deep comparative analysis of the ~4 million TE copies shaping the A, B, and D subgenomes led to unexpected conclusions about the evolutionary dynamics of TEs (Wicker et al., 2018). Since their divergence 6 MYA, the TE turnover has fully replaced ancestral TEs by more recent ones, independently in the evolution of diploid lineages while maintaining a stable genome size, and no burst of TE transposition was

### Core Ideas

- Base pair-resolved whole genome alignments of *Triticeae* A, B, and D subgenomes were analyzed.
- Structural variations affect 5%–15% of the transposable element (TE) content.
- Traces of transposition were detected for almost all TE families.
- TE insertions and deletions are balanced.
- There was a constant rate of transposition per TE family across the lineages.

observed after polyploidization events. Surprisingly, despite the TE turnover, the wide majority of the ~500 TE families are still present in similar proportions in each subgenome, although they evolved into subgenome-specific variants, that is, subfamilies. Abundant families are still the same while low-copy families persist at low-copy numbers. Hypotheses raised were that TE turnover is highly regulated and may follow evolutionary constraints. This conclusion challenged the burst-centered view of plant TE dynamics and was rather in line with an alternative scenario of equilibrium as observed, for instance, in *Brachypodium distachyon* natural populations where TE activity is “remarkably constant” (Stritt et al., 2018) or as exemplified by the Alesia family (Stritt et al., 2021) which maintains low-copy numbers in the Angiosperms across the generations, suggesting self-regulatory mechanisms (Cosby et al., 2019). TEs might maintain an equilibrium between deletion and amplification in *Triticeae* but it is still a matter of debate (Bariah et al., 2020).

The limit with comparing A-B-D genomes is that all ancestral TEs have been erased so it is not possible to trace recent deletion/transposition events. For that, it is necessary to compare genomes that have diverged much more recently to identify structural variations. This is what was performed by producing resequencing data on flow-sorted chromosomes 3B in a panel of 45 diverse *Triticeae* (De Oliveira et al., 2020). The extent of TE presence-absence variations (PAVs) was estimated to 7%–8% per *T. aestivum* accession and up to 24% in more distant species and, again, no burst of any specific TE family was observed. In contrast, recent TE insertions were detected for a wide diversity of families, confirming that most of families were active and transposed. Several *Triticeae* genomes assembled at reference-quality level are now available, offering the opportunity to analyze TE dynamics at a resolution never reached so far using whole genome alignments. Analysis of five full-length long terminal repeat (LTR) retrotransposons across *T. aestivum* assembled genomes confirmed that distinct subfamilies were active in

diploids mainly, in waves lasting hundred thousand years, with only sparse evidence of recent insertions in polyploids (Wicker et al., 2022). However, the extent of variability due to TEs deletion/amplification is still unknown.

In this study, we developed a method in order to compute *Triticeae* whole-genome sequence alignments guided by anchor-genes and retrieve the complete landscape of TE variability originated from deletions and transpositions. We thus compared the fully assembled sequences of the A genomes, the B genomes, and also the D genomes between di-, tetra-, and hexaploids: *T. urartu* (AA; Ling et al., 2013), *Ae. tauschii* (DD; Jia et al., 2013), *T. dicoccoides* (AABB; Avni et al., 2017), *T. durum* (AABB; Maccaferri et al., 2019), and 13 *T. aestivum* (AABBDD; Aury et al., 2022; IWGSC, 2018; Walkowiak et al., 2020). The evolutionary time scale studied here is thus limited to the recent evolution corresponding to the early stages of TE turnover: 0.01 million year (Myr) for comparisons within hexaploids and between *T. durum* and *T. aestivum*, ~0.8 Myr between *T. dicoccoides* and *T. aestivum*, and 1.3 Myr for *T. urartu* which diverged from the A genome donor ~0.5 Myr before tetraploidization. We show that variability is affected by 5%–34% of the TEs depending on the species compared. We identified 51,928 recent transposition events involving 346 different families that were active recently in all species whatever their ploidy level. We show that transposition rate is similar in di-, tetra-, and hexaploids, confirming the equilibrium we observed previously, and confirming that polyploidy did not disturb this equilibrium.

## 2 | MATERIALS AND METHODS

### 2.1 | Genome sequence data

The reference-quality assembled genome sequences of *T. urartu* cultivar G1812 v2.0 (PRJNA337888) (Ling et al., 2018), *Ae. tauschii subsp. strangulata* cultivar AL8/78 v4.0 (PRJNA182898) (Luo et al., 2017), *T. turgidum subsp. durum* cultivar Svevo v1.0 (PRJEB22687) (Maccaferri et al., 2019), *T. turgidum subsp. dicoccoides* isolate Atlit2015 ecotype Zavitan v2.0 (PRJNA310175) (Zhu et al., 2019) and *T. aestivum* RefSeq v1.0 (PRJEB27788) (IWGSC, 2018) were downloaded from NCBI (<https://www.ncbi.nlm.nih.gov/>). Reference-quality assembled genome sequences of *T. aestivum* accessions ArinaLrFor, CDC Landmark, CDC Stanley, Jagger, Julius, LongReach Lancer, Mace, Norin61, Spelt, and SY\_Mattis were downloaded from <https://wheat.ipk-gatersleben.de/> (Walkowiak et al., 2020). We also used the Renan genome sequence that we published previously (GCA\_937894285) (Aury et al., 2022), and that of Tibetan wheat Zang1817 (Guo et al., 2020).

### 2.2 | TE annotation and comparison of family proportions

We used the available TE annotations that we computed previously for Chinese Spring (Aury et al., 2022; Wicker et al., 2018) and Renan (Aury et al., 2022; Wicker et al., 2018). For all the other genome sequences compared in this study, we did not use the available annotation but rather produced a TE annotation using CLARITE and the ClariTeRep library (Daron et al., 2014) with the same parameters as for Chinese Spring and Renan. CLARITE uses RepeatMasker (Smit et al., 1996–2004) for similarity search, applies a step of defragmentation in order to merge adjacent predictions that describe a single element, and eventually applies a step of reconstruction of nested insertions. The abundance of each TE (sub)family in a genome was calculated by cumulating the length of all fragments assigned a given (sub)family divided by the total length of TEs. To investigate differences of TE family abundance between genomes, and potential enrichment in the variable fraction of the genome, we calculated proportions of each TE family (cumulated length assigned to a given family divided by the subgenome size) and computed log<sub>2</sub> ratios. Only families accounting for more than 100 kb in the analyzed subgenome were considered.

### 2.3 | Estimation of the extent of genomic variability using insertion site-based polymorphism (ISBP) markers

Based on CLARITE predictions of TEs, we extracted 150 bps encompassing the 5' and 3' junctions of each TE extremity with its insertion site with 75 bps on each side of the junction as previously described (De Oliveira et al., 2020). These 150 bps tags correspond to insertion site-based polymorphism (ISBP) markers (Paux et al., 2006) that are expected to be unique kmers at the whole genome level. For each genome studied, we extracted all ISBPs and discarded those containing Ns. In case two ISBPs overlap by 100 bps or more, we kept only one of both. We also discarded ISBPs that are non-unique in the genome from which they were designed: mapped with Minimap2 (Li, 2018) at multiple loci with cutoff of 98% identity and 100% query overlap. Presence/absence of ISBPs were searched in all compared genomes using Minimap2 option -xsr -w5. An ISBP was considered absent (PAV) if no match was found considering at least 95% identity and 95% coverage (maximum 7 bases soft-clipped). Only pseudomolecule sequences were considered for this analysis, meaning that unanchored scaffolds ("ChrUn") were excluded. To determine nucleotide positions defining the limits between distal (R1 and R3), proximal (R2), and (peri)-centromeric (C) regions of each chromosome (as previously defined in Chinese Spring;



IWGSC, 2018) in all species/accessions analyzed, we used the Chinese Spring ISBP mapping data. Hence, the chromosomal position of the ISBP that was the closest to a border (between R1 and R2 of chr1A for instance) defined in Chinese Spring was used as border in the compared species/accessions.

## 2.4 | Estimation of the extent of genomic variability using orthologous intergenic regions

To estimate the extent of variability that is, proportion of variable versus conserved orthologous TEs, we split the genome into small orthologous intervals in order to avoid analyzing whole-chromosome alignments with spurious matches between repeated TEs. Thus, we used the 105,200 High Confidence (HC) genes with a position along the 21 pseudomolecules of RefSeq v1.1 (IWGSC, 2018) as anchors to find orthologous intergenic regions in other genomes. The nucleotide sequences of the corresponding CDSs were mapped using GMAP (v18.05.11; Wu & Watanabe, 2005, options: `gmapl-cross_species`) on the homeologous chromosome for each genome compared. Only best hit with at least 90% identity and 90% coverage were considered and kept for the subsequent analyses. We developed a Python script (`Get_Collinear_Region.py`) in order to retrieve all orthologous intergenic regions (oIGRs) that were flanked by the same pair of neighbor orthologs (collinearity) in Chinese Spring and in the compared genome. Cases of tandem duplicated gene copies that mapped at the same positions in a compared genome were dealt with by the script so that it specifically determined which copies delimitate an oIGR. Chromosome positions of these pairs of neighbor orthologs were used to extract the corresponding genomic segments (including the genes at both extremities of oIGRs) in both compared genomes. oIGRs were then aligned with BLASTN (v2.11.0+; Camacho et al., 2009, threshold value:  $1e-5$ ) and we filtered out HSPs (High Scoring Pairs) with a cutoff at 90% identity. We excluded HSPs whose coordinates were included in a larger one. This happened in case of lineage-specific tandem duplications, or when a novel copy of an element shared homology with another TE present within the aligned region. Coordinates of HSPs were then used to create BED files and we used “Bedtools merge” (bedtools/2.26; Quinlan & Hall, 2010) to merge overlapping conserved segments for each genome. We then used “Bedtools complement” to create BED files describing the variable (i.e., specific) segments between genomes that is, sequences that are subject to presence-absence variations (PAVs) between two compared oIGRs. PAV candidates corresponding to gaps in the assembly (stretches of Ns) were identified and discarded from PAVs. We eventually used “Bedtools intersect” to retrieve TE annotation of the conserved/variable sequences.

## 2.5 | Detection of recent TE insertions and estimation of the insertion dates

BED files describing the positions of PAVs (see above) were analyzed with the objective of finding PAV coordinates that fit nearly perfectly (i.e., with a tolerance of 10 bps at 5' and 3' extremities) with the coordinates of a TE, with status “complete” annotated by CLARITE, as evidence for recent transposition (or possible excision for class 2 elements). For each candidate of novel insertion, we searched for the presence of a target site duplication (TSD) as molecular evidence of insertion. TSD are short motifs repeated at 5' and 3' ends of a TE and immediately flanking the terminal motifs that determine the exact borders of the TE. Since the predicted extremities of TEs may not correspond exactly to the terminal nucleotides, TSDs were searched within a subsequence of 20 nucleotides (nts) overlapping both TE extremities: 10 nts on each side of the predicted extremity for superfamilies RLG, RLC, RLX, and DTC; 5 nts inside+10 nts outside for superfamilies DTM, DTT, DTH, and DTA. We developed Python scripts (1 per superfamily; available at [https://forgemia.inra.fr/umr-gdec/scripts\\_files/](https://forgemia.inra.fr/umr-gdec/scripts_files/)) to identify TSDs of variable size, depending on the superfamily considered (Wicker et al., 2007), immediately flanking terminal motifs: 5'-TG and CA-3' for RLGs, RLCs, and RLXs; 5'-CACT[AG] and [CT]AGTG-3' for DTCs. For the other superfamilies with undefined terminal motifs, the entire subsequence of 20 nts was scanned for the presence of a TSD. The expected TSD size was 5 nts for RLGs, RLCs, and RLXs, 3 nts for DTCs and DTHs, 8 nts for DTAs, varying between 9 and 11 nts for DTM, and a TA duplication was expected for DTTs. We tolerated one single nucleotide polymorphism (SNP) between 5' and 3' TSDs. No TSD was searched for LINES/SINEs. Insertion dates of LTR retrotransposons were estimated by aligning the two LTRs of an element with BLASTN and we used a mutation rate of  $1.3 \times 10^{-8}$  substitutions/site/year (Ma & Bennetzen, 2004) as previously described (Wicker et al., 2018). Distances of newly inserted TE from the nearest predicted gene were computed with “Bedtools closest.”

## 3 | RESULTS

### 3.1 | TE annotation and comparison of orthologous subgenomes between di, tetra-, and hexaploid *Triticeae*

To avoid biases due to different TE annotation approaches, we predicted TEs in the genome sequences of *T. aestivum* (13 accessions), *T. urartu*, *Ae. tauschii*, *T. dicoccoides*, and *T. durum* (abbreviated *Tae*, *Tur*, *Aet*, *Tdi*, and *Tdu*,

**TABLE 1** Proportions of transposable element (TE) superfamilies in the A, B, and D subgenomes of Triticeae (sub)genomes annotated with CLARITE.

| (Sub)genome                  | <i>T. urartu</i> (%) |      |      | <i>T. dicoccoides</i> (%) |      | <i>T. durum</i> (%) |      | <i>Ae. tauschii</i> (%) | <i>T. aestivum</i> (Chinese Spring) (%) |      |  |
|------------------------------|----------------------|------|------|---------------------------|------|---------------------|------|-------------------------|---|------|--|
|                              | A                    | A    | B    | A                         | B    | D                   | D    | A                       | B                                       | D    |  |
| All TEs                      | 86.8                 | 85.4 | 84.4 | 86.9                      | 86.0 | 83.4                | 83.4 | 86.1                    | 84.7                                    | 83.1 |  |
| Class I: Retrotransposons    | 71.8                 | 70.7 | 66.5 | 72.0                      | 67.9 | 62.0                | 62.0 | 71.9                    | 67.4                                    | 62.4 |  |
| RLG (Gypsy)                  | 49.7                 | 49.7 | 46.0 | 50.8                      | 47.0 | 40.7                | 40.7 | 50.9                    | 46.8                                    | 41.4 |  |
| RLC (Copia)                  | 18.6                 | 17.3 | 16.0 | 17.6                      | 16.3 | 16.4                | 16.4 | 17.5                    | 16.2                                    | 16.3 |  |
| RLX (unclassified)           | 2.5                  | 2.7  | 3.4  | 2.7                       | 3.5  | 3.7                 | 3.7  | 2.6                     | 3.5                                     | 3.7  |  |
| RIX (LINE)                   | 0.94                 | 0.95 | 1.10 | 0.95                      | 1.13 | 1.08                | 1.08 | 0.82                    | 0.96                                    | 0.93 |  |
| RSX (SINE)                   | 0.01                 | 0.01 | 0.01 | 0.01                      | 0.01 | 0.01                | 0.01 | 0.01                    | 0.01                                    | 0.01 |  |
| Class II: DNA transposons    | 14.4                 | 14.2 | 17.0 | 14.3                      | 17.2 | 20.7                | 20.7 | 13.7                    | 16.5                                    | 20.1 |  |
| DTC (CACTA)                  | 13.4                 | 13.2 | 15.9 | 13.3                      | 16.1 | 19.4                | 19.4 | 12.8                    | 15.5                                    | 19.0 |  |
| DTA (hAT)                    | 0.01                 | 0.01 | 0.01 | 0.01                      | 0.01 | 0.01                | 0.01 | 0.01                    | 0.01                                    | 0.01 |  |
| DTM (Mutator)                | 0.35                 | 0.34 | 0.43 | 0.35                      | 0.44 | 0.55                | 0.55 | 0.30                    | 0.38                                    | 0.48 |  |
| DTT (Mariner)                | 0.15                 | 0.15 | 0.17 | 0.15                      | 0.18 | 0.19                | 0.19 | 0.14                    | 0.16                                    | 0.17 |  |
| DTH (Harbinger)              | 0.17                 | 0.17 | 0.17 | 0.17                      | 0.18 | 0.19                | 0.19 | 0.15                    | 0.16                                    | 0.18 |  |
| DTX (unclassified with TIRs) | 0.23                 | 0.23 | 0.23 | 0.23                      | 0.23 | 0.24                | 0.24 | 0.21                    | 0.20                                    | 0.22 |  |
| DHH (Helitron)               | 0.05                 | 0.05 | 0.08 | 0.05                      | 0.08 | 0.06                | 0.06 | 0.05                    | 0.08                                    | 0.05 |  |
| DXX (unclassified class II)  | 0.01                 | 0.01 | 0.01 | 0.01                      | 0.01 | 0.01                | 0.01 | 0.00                    | 0.00                                    | 0.00 |  |
| Unclassified repeats XXX     | 0.59                 | 0.61 | 0.89 | 0.60                      | 0.85 | 0.74                | 0.74 | 0.51                    | 0.81                                    | 0.59 |  |
| Genes and unannotated DNA    | 13.2                 | 14.6 | 15.6 | 13.1                      | 14.0 | 16.6                | 16.6 | 13.9                    | 15.3                                    | 16.9 |  |

Note: Proportions are expressed as the percentage of sequences assigned to each superfamily relatively to the (sub)genome size.

Abbreviation: TIRs, terminal inverted repeats.

respectively) with the same method and criteria, using CLARITE (Daron et al., 2014). TEs represent between 86% and 87% for the A lineage, between 84% and 85% for B, and between 82% and 83% for D. We previously showed that A-B-D diverged ~5–6 MYA, a time during which the TE turnover has erased ancestral TEs so that there is (almost) no TE conserved between orthologous loci (homeologous in polyploids). Here, we focused our work on a much more recent evolutionary scale, by comparing A subgenomes together, B together, and D together, at different ploidy levels. TEs shaping the intergenic space are, thus, conserved because they were present in the common ancestor of the genomes compared: 0.01 MYA for *Tae*<sup>ABD</sup> accessions, *Tae*<sup>AB</sup>/*Tdu*<sup>AB</sup> and *Tae*<sup>D</sup>/*Aet*<sup>D</sup>, 0.8 MYA for *Tae*<sup>AB</sup>/*Tdi*<sup>AB</sup>, and at maximum 1.3 MYA for *Tur*<sup>A</sup> divergence from tetra- and hexaploids (Levy & Feldman, 2022).

The amount of TEs appeared very similar, even for each TE superfamily, when we compared A (sub)genomes, so did B, and D (Table 1; Table S1). *T. urartu*<sup>A</sup> appeared the most different, which fits with its earlier divergence. To get a first flavor about the extent of variability, we started by comparing (sub)genome-wide proportions of TE families between lineages. Using Chinese Spring (hereafter abbreviated CS) as a reference, we distinguished 301 TE families (those accounting for more than 100 kb in at least one

subgenome), with the 20 most abundant representing >84% of the TE fraction, meaning that most families are present at low-copy number. We showed that 97% (292/301) of the families were present in similar proportions in the compared genomes (fold-change < 2), suggesting that none of these *Triticeae* have experienced a massive transposition “burst” of any family since they diverged, neither before, nor after, the polyploidization events. Rare cases of differential abundance were observed for nine very low-copy families (DTM\_famc13, RIX\_famc25 XXX\_famc10-150-46-53-57-60, DHH\_famc1) whose abundance varies around the 100 kb cutoff applied here and for which the extent of the difference may not be associated with TE amplification but rather to methodological limits (partial genome assembly and anchoring, TE annotation). The only remarkable case is family XXX\_famc10, which is not a TE per se but rather a subtelomeric satellite repeat specific to the B lineage. It exhibited strong differences of copy number affecting the A subgenome between diploid *Tur*<sup>A</sup> and tetraploid *Tdi*<sup>A</sup>, which is explained by the previously characterized 4A/7B translocation (Dvorak et al., 2018) that brought a part of the B genome onto the 4A chromosome in the tetraploid ancestor. Such global conservation of family proportions may hide higher levels of structural variations between those genomes, which we investigated using uniquely mappable TE-derived markers.

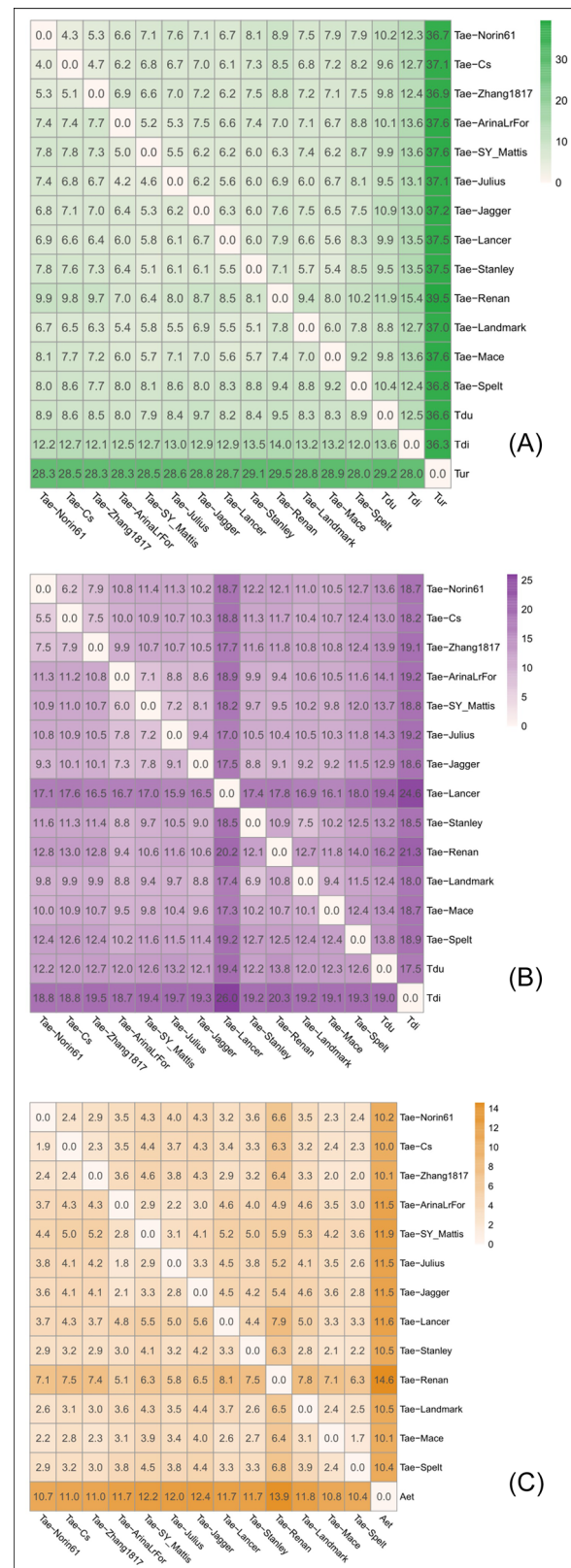
### 3.2 | Extent of structural variations affecting TEs through the mapping of ISBP markers

Although TEs are repeated, each copy is inserted into a different locus, which makes the 5' and 3' junctions between TE extremities and the insertion site, unique kmers at the whole genome level. We used this valuable feature to address presence/absence variations (PAVs), also called transposon insertion polymorphism (TIPs) or ISBPs among the wheat research community, between orthologous loci. We extracted 150 bps encompassing each TE extremity of all genomes which provided a dataset of, on average, 1.7, 1.9, and 1.5 million uniquely mappable ISBP markers from the A, B, and D subgenomes, respectively. ISBPs mapped with at least 95% identity over 95% of their length were considered present while no match revealed a PAV. Proportions of markers subject to PAVs in all pairwise comparisons are presented in Figure 1.

At the interspecific level, considering CS as a reference for *T. aestivum*, we show that TE PAVs ranged from 10% for the D subgenome (*Tae<sup>D</sup>* vs. *Aet<sup>D</sup>*) to 37% for the A subgenome (*Tae<sup>A</sup>-Tdi<sup>A</sup>-Tdu<sup>A</sup>* vs. *Tur<sup>A</sup>*), reflecting the earlier divergence of *T. urartu*. Comparing tetraploids to hexaploids showed that *T. aestivum* is closer to *T. durum* than to *T. dicoccoides*. In addition, the variability is higher for the B than for A subgenome: proportions were estimated to 10% (A) and 13% (B) for *Tae* compared to *Tdu*, and to 13% (A) and 18% (B) compared to *Tdi*. We observed the same shift between A and B genome variability when comparing tetraploids together (*Tdi* vs. *Tdu*). Altogether, these data highlighted that around 10%–18% of the TE space is variable (specific), in terms of presence/absence, between species *T. aestivum*, *T. durum*, *T. dicoccoides*, and *Ae. tauschii*, while it is much more variable compared to *T. urartu*.

At the intraspecific level, PAV detection using 13 fully assembled genomes of *T. aestivum* revealed a similar pattern: D subgenome is the least variable (4% on average), A is intermediate (7%), and B is the more variable (11%). Outliers were observed for the B subgenome of CDC Lancer (16%) due to the presence of an alien chr2B originating from *Triticum timopheevii* (Walkowiak et al., 2020). A slight increase (1%–2%) of variability was observed for Renan compared to all other European wheat accessions, due to the sequencing method that was different (Oxford Nanopore vs. Illumina) and may have led to 1%–2% ISBPs with sequencing errors (indels preventing from aligning 95% of the ISBP length).

PAV levels detected here were in agreement with the existence of two phylogenetically distinct groups corresponding to the Asian (CS, Norin61, Zang1817) and European wheat genetic pools. Variability was slightly lower within each group (A: 4.8%; B: 7.1%; D: 2.4% on average for the Asian pool; A: 6.8%; B: 10.0%; D: 4.2% on average for the European pool



**FIGURE 1** Matrix of the levels of variability presence-absence variations (PAVs) estimated using insertion site-based polymorphism (ISBP) mapping. Values represent the percentages of ISBP markers that are under PAVs in pairwise comparisons that is, present in the query (in rows) and absent in the reference (in columns) genome.



excluding CDC Lancer) than between groups (A: 7.4%; B: 11.1%; D: 4.2% on average).

These global estimates at the whole subgenome level may hide strong local differences. Chromosome extremities (distal regions) were, on average, four times more variable than the central regions of chromosomes (borders as defined in IWGSC, 2018). For instance, between the closely related Asian accessions CS and Norin61 (Figure 1), 4% of TEs are variable, however, it reached up to 9% in the fast-evolving distal regions whereas it is limited to 2% for the rest of the genome. Hence, TE PAVs can be high (>10%) in the distal regions even between accessions that are closely related. These results confirmed, at a short evolutionary scale, previous assessments suggesting accelerated evolution in the recombinogenic distal regions compared to the rest of the genome.

This ISBP-based PAV detection method provided an unbiased genome-wide view of the extent of the variable versus conserved parts of *Triticum/Aegilops*. Roughly, it represents 5%–10% and 10%–20% of the TE-derived markers at the intra- and inter-specific levels, respectively, in pairwise comparisons, with substantial differences between the B (more variable), A (intermediate), and D (least variable) subgenomes. After getting this first flavor of genome-wide structural variability using ISBPs as proxy, we established a method to characterize at a bp resolution which TEs comprise the variable and conserved parts through pairwise alignments of orthologous intergenic regions.

### 3.3 | TE variability assessed by whole genome alignments

Aligning Gb-sized genomes containing 85% of TEs is not trivial. In this regard, we developed a dedicated approach aiming at identifying collinear orthologous genes in order to target the pairwise alignment of pre-identified orthologous intergenic regions (oIGRs). Therefore, we mapped the 105,200 predicted genes of *T. aestivum* Chinese Spring (35,345, 35,643, and 34,212 carried by subgenomes A, B, and D, respectively) on related genomes with high stringency and found unambiguous orthologs for 79%, 96%, 94%, and 94% of them in *T. urartu*<sup>A</sup>, *Ae. tauschii*<sup>D</sup>, *T. dicoccoides*<sup>AB</sup>, and *T. durum*<sup>AB</sup>, respectively. These genes were used as anchors to guide all genome sequence alignments. From this, we extracted all intergenic regions flanked by collinear neighbor orthologs, representing a dataset of 17,904, 30,623, 59,048, and 59,812 oIGRs, respectively. Their cumulated length represents 35/34% (1.6/1.7 Gb) of the *Tur*<sup>A</sup>/*Tae*<sup>A</sup> subgenomes, 90/90% (3.6/3.5 Gb) of the *Aet*<sup>D</sup>/*Tae*<sup>D</sup> subgenomes, 83/83% (8.6/8.4 Gb) of the *Tdi*<sup>AB</sup>/*Tae*<sup>AB</sup> genomes, and 86/84% (8.6/8.5 Gb) of the *Tdu*<sup>AB</sup>/*Tae*<sup>AB</sup> genomes. We then retrieved the positions of the conserved versus variable (specific)

sequences from pairwise oIGR alignments as illustrated in Figure 2. This fine-tuned approach allowed us to get a bp-resolved view of TE copies that are conserved or affected by PAVs (due to insertions, duplications, or deletions) among >90% of the A-B-D subgenomes, except with *Tur*<sup>A</sup> which exhibited lower collinearity and lower assembly quality.

The variable fraction of the A genome represents 34% of CS *Tae*<sup>A</sup> subgenome when compared to diploid *Tur*<sup>A</sup> (Table 2). Tetraploids are much more closely related: the variable fraction represents 7% and 10% of the A subgenome compared to *Tdu*<sup>A</sup> and *Tdi*<sup>A</sup>, respectively. The B subgenome exhibits higher variability level: 9% and 13%, respectively. Compared to diploid *Aet*<sup>D</sup>, 9% of the *Tae*<sup>D</sup> subgenome is under PAVs. These values are in agreement with the ISBP-based estimates described in the above paragraph although oIGR analysis tended to slightly underestimate the level of variability because we did not sample regions where genes are noncollinear. Considering that A, B, and D subgenomes are shaped by 1.2 million TEs on average, our results revealed that roughly 120k TEs (~10%) are non-conserved while ~90% of the TE space is still conserved in pairwise comparisons.

Intraspecific pairwise comparisons of 13 *T. aestivum* accessions revealed that the variable TE fraction is in the same range than compared to *T. durum*, representing on average 6% of A and 8% of B subgenomes, although it is lower, 4% and 5%, when comparing CS with closer Asian accessions Norin61 and Zang1817. Thus, we did not observe a strong difference in terms of TE turnover when comparing CS with hexaploids and with *T. durum*. *T. dicoccoides* appeared more distantly related. For the D subgenome, only 4% of CS<sup>D</sup> TEs were affected by PAVs compared to other accessions which is half that observed compared to *Ae. tauschii*<sup>D</sup> (9%).

Alignments of oIGRs of CDC Lancer revealed a higher level of TE PAVs for the B subgenome due to the presence of an alien chromosome 2B, as commented above. Except for such introgressions, TE variability appeared quite stable and in agreement with SNP-estimated divergence. We did not observe cases where TE activity would have triggered accelerated TE turnover.

Another conclusion we could draw from Table 2 is that the proportions of variable regions are similar in both the query and reference aligned genomes: for instance, 320 Mb (7.5%) of *Tdu*<sup>A</sup> TEs are absent in CS<sup>A</sup> while, reversely, 286 Mb (6.7%) of the CS<sup>A</sup> TEs are absent in *Tdu*<sup>A</sup>. Between A subgenomes of accessions CS and Norin61, specific TEs account for 186 Mb and 197 Mb, respectively. This shows that none of the genomes analyzed here evolved toward expansion or contraction of the TE space. In contrast, the rate of TE turnover is globally conserved in all lineages analyzed, suggesting that insertions of novel TEs compensate TE loss by deletions.



**TABLE 2** Proportions of the variable transposable element (TE) fraction identified by orthologous intergenic region (oIGR) alignments in each pairwise comparison with Chinese spring (CS).

| Pairwise comparison query versus CS | Sub genome | No. of oIGRs         | oIGR length (Mb)  |                   | oIGR %   |          | Variable region length (Mb) |                | Variable regions (%) |         |
|-------------------------------------|------------|----------------------|-------------------|-------------------|----------|----------|-----------------------------|----------------|----------------------|---------|
|                                     |            |                      | Query             | CS                | Query    | CS       | Query                       | CS             | Query                | CS      |
| <i>T. urartu</i> versus CS          | A/—/—      | 17,904/—/—           | 1616/—/—Mb        | 1683/—/—Mb        | 35/—/—   | 34/—/—   | 511/—/—Mb                   | 572/—/—Mb      | 32/—/—               | 34/—/—  |
| <i>T. durum</i> versus CS           | A/B/—      | 30,169/29,643/—      | 4282/4288/—Mb     | 4249/4255/—Mb     | 88/84/—  | 86/82/—  | 320/434/—Mb                 | 286/399/—Mb    | 7/10/—               | 7/9/—   |
| <i>T. dicoccoides</i> versus CS     | A/B/—      | 30,098/28,950/—      | 4289/4326/—Mb     | 4201/4174/—Mb     | 85/81/—  | 85/81/—  | 526/732/—Mb                 | 429/561/—Mb    | 12/17/—              | 10/13/— |
| <i>Ae. tauschii</i> versus CS       | —/—/—D     | —/—/30,623           | —/—/3614 Mb       | —/—/3536 Mb       | —/—/90   | —/—/90   | —/—/355 Mb                  | —/—/305 Mb     | —/—/10               | —/—/9   |
| <i>Tae-Spelt</i> versus CS          | A/B/D      | 31,077/30,604/32,346 | 4448/4501/3777 Mb | 4415/4444/3761 Mb | 91/88/95 | 89/86/95 | 320/458/142 Mb              | 283/396/123 Mb | 7/10/4               | 6/9/3   |
| <i>Tae-Mace</i> versus CS           | A/B/D      | 31,732/30,672/32,401 | 4451/4487/3768 Mb | 4450/4446/3766 Mb | 91/88/96 | 90/86/95 | 268/392/134 Mb              | 263/347/130 Mb | 6/9/4                | 6/8/3   |
| <i>Tae-Landmark</i> versus CS       | A/B/D      | 31,849/31,053/32,241 | 4514/4553/3766 Mb | 4458/4475/3727 Mb | 91/87/95 | 90/86/94 | 315/428/182 Mb              | 264/353/149 Mb | 7/9/5                | 6/8/4   |
| <i>Tae-Renan</i> versus CS          | A/B/D      | 30,630/29,934/30,583 | 4266/4303/3718 Mb | 4283/4342/3524 Mb | 86/82/93 | 87/84/89 | 331/412/273 Mb              | 281/365/207 Mb | 8/10/7               | 7/8/6   |
| <i>Tae-Stanley</i> versus CS        | A/B/D      | 31,548/30,829/32,200 | 4535/4560/3786 Mb | 4470/4460/3743 Mb | 91/87/95 | 91/86/95 | 323/464/170 Mb              | 263/367/133 Mb | 7/10/4               | 6/8/4   |
| <i>Tae-Lanceer</i> versus CS        | A/B/D      | 32,058/30,003/32,001 | 4495/4214/3744 Mb | 4485/4252/3736 Mb | 92/84/95 | 91/82/95 | 253/560/140 Mb              | 240/594/130 Mb | 6/13/4               | 5/14/3  |
| <i>Tae-Jagger</i> versus CS         | A/B/D      | 31,223/30,583/31,405 | 4477/4520/3687 Mb | 4412/4447/3660 Mb | 90/87/93 | 89/86/93 | 306/401/172 Mb              | 251/334/153 Mb | 7/9/5                | 6/8/4   |
| <i>Tae-Julius</i> versus CS         | A/B/D      | 31,870/30,872/32,138 | 4551/4595/3776 Mb | 4497/4513/3736 Mb | 92/88/95 | 91/87/95 | 301/441/179 Mb              | 255/365/146 Mb | 7/10/5               | 6/8/4   |
| <i>Tae-SY_Mattis</i> versus CS      | A/B/D      | 31,715/30,489/31,906 | 4526/4467/3730 Mb | 4497/4436/3721 Mb | 92/87/95 | 91/86/94 | 274/364/161 Mb              | 243/327/150 Mb | 6/8/4                | 5/7/4   |
| <i>Tae-ArinaLrFor</i> versus CS     | A/B/D      | 32,041/30,674/32,258 | 4558/4501/3789 Mb | 4510/4441/3752 Mb | 92/86/95 | 91/86/95 | 293/404/177 Mb              | 244/344/140 Mb | 6/9/5                | 5/8/4   |
| <i>Tae-Zang1817</i> versus CS       | A/B/D      | 32,400/31,922/32,551 | 4654/4752/3789 Mb | 4625/4699/3774 Mb | 94/92/96 | 94/91/96 | 228/339/138 Mb              | 198/283/122 Mb | 5/7/4                | 4/6/3   |
| <i>Tae-Norin61</i> versus CS        | A/B/D      | 32,898/32,253/32,839 | 4613/4760/3793 Mb | 4605/4733/3791 Mb | 94/92/96 | 93/91/96 | 197/265/116 Mb              | 186/235/114 Mb | 4/6/3                | 4/5/3   |

Abbreviation: *Tae*, *Triticum aestivum*.



**FIGURE 2** Example of an orthologous intergenic region on chr7A compared using Minimap2 across the four species with (sub)genome A (*T. urartu*, *T. dicoccoides*, *T. durum*, *T. aestivum*), and 13 *T. aestivum* accessions. Conserved transposable element (TEs) are covered by grey areas which illuminates regions subject to presence-absence variations (PAVs). PAV coordinates may fit with TE coordinates suggesting recent TE insertion/excision.

We then wondered about the composition of the variable fraction of the TE space in order to investigate which families have impacted the recent *Triticeae* genome evolution. Indeed, recent TE amplification of active families could be detected because they are enriched in the variable fraction. Thus, we searched for families whose abundance is substantially enriched in the variable fraction of the genome compared to its genome average. We calculated enrichment ratios for the 100, 113, and 98 most abundant families of the A, B, and D subgenomes, respectively (those representing at least 1 Mb per subgenome in CS). Together they represent 99% of the TE content, the others being low-copy TEs. Fold changes are shown as heatmaps in Figure 3. The main result here is that such enrichment is rare. The composition of the variable fraction is quite similar to the genome average. However, we found 3 TE families (2 CACTAs and 1 Gypsy) and 2 satellite repeats that were enriched ( $\log_2$  fold change  $\geq 2.0$ ) in the variable TE fraction of some genomes: RLG\_famc8 (Cereba, Quinta), DTC\_famc4 (Clifford/Mandrake/Byron), DTC\_famc9 (Isaac), and satellites XXX\_famc1 (Tail) and XXX\_famc10 (unnamed). RLG\_famc8 (Cereba/Quinta) is a centromeric gypsy family which represents on average 1% of the oIGRs but accounts for 5% of the variable sequences

detected. Centromeric TEs and telomeric satellites have been, thus, main components of the recent (intraspecific) genome structure diversification. But if these examples are striking, together they were responsible for at maximum 5%–6% (in bps) of the PAVs, showing that the observed variability does not originate from amplification bursts of a few very active families. In contrast, the composition of the recently shaped TE space resembles the ancestral one.

### 3.4 | Estimation of recent transposition rate and impact of polyploidy

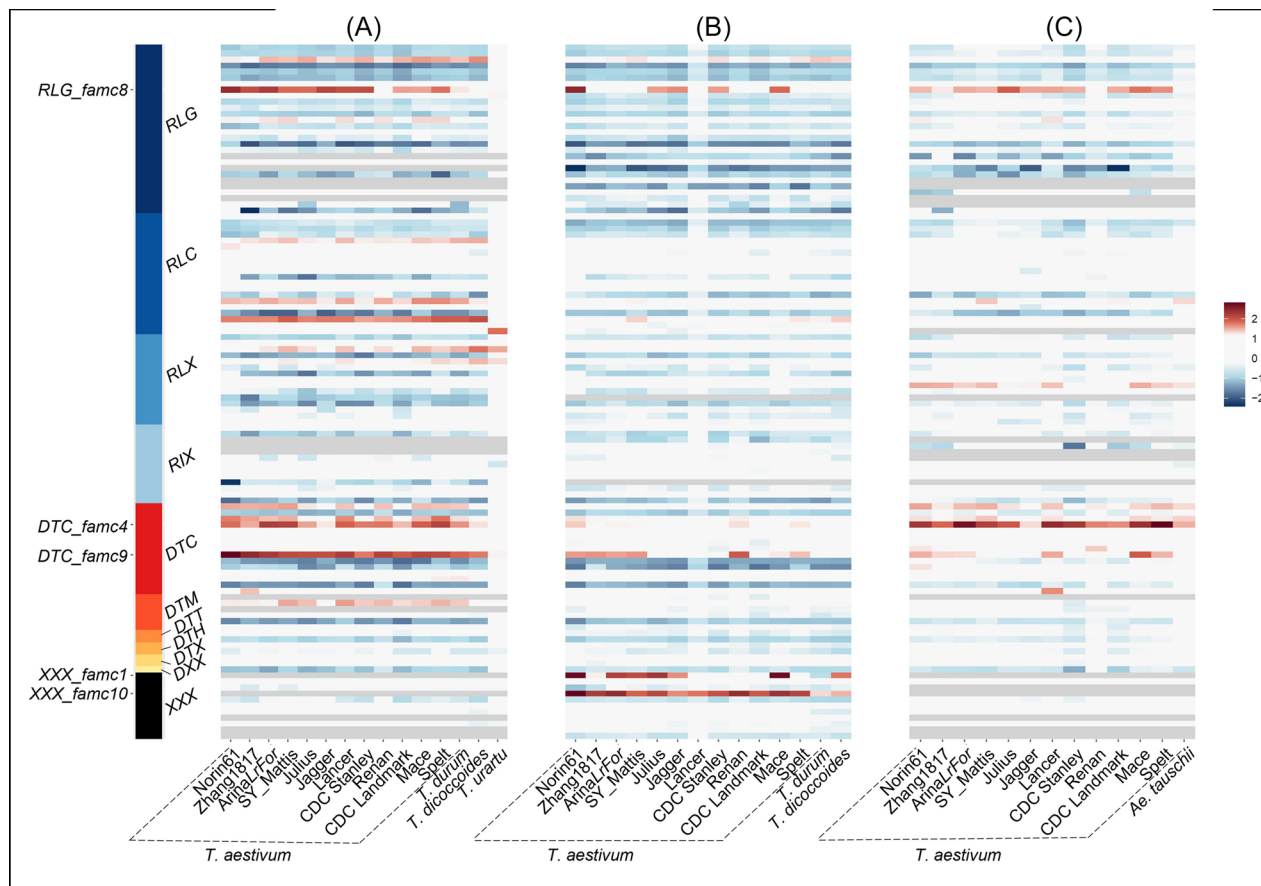
Variable regions originated from both deletions and insertions. To get deeper insights into the dynamics of transposition, we searched for PAVs (identified by oIGR pairwise BLAST alignments) whose borders fit with borders of a TE, as potential traces of transposition (or potentially excision for class 2 elements). The number of such events is summarized in Table 3. Pairwise comparisons with CS revealed between 433 and 12,491 recent TE insertions per subgenome, depending on the species/accession considered. We found the presence of TSDs (target site duplications,

TABLE 3 Number of specific transposable element (TE) insertions detected in pairwise orthologous intergenic region (oIGR) alignments.

| Pairwise comparison       | (Sub) genome | QUERY   CS  |                          | Total           |
|---------------------------|--------------|-------------|--------------------------|-----------------|
|                           |              | Class I     | Class II (+unclassified) |                 |
| <i>T. urartu</i> /CS      | A            | 6632   8200 | 1188   2233              | 7820   10,433   |
| <i>T. dicoccoides</i> /CS | A            | 8582   8662 | 2232   2374              | 10,814   11,036 |
|                           | B            | 9835   9444 | 2906   3047              | 12,741   12,491 |
| <i>T. durum</i> /CS       | A            | 4411   4202 | 1510   1504              | 5921   5706     |
|                           | B            | 5703   5232 | 1792   1812              | 7495   7044     |
| <i>Ae. tauschii</i> /CS   | D            | 3122   3438 | 1123   1011              | 4245   4449     |
| <i>Tae-Spelt</i> /CS      | A            | 4787   4387 | 1469   1399              | 6256   5786     |
|                           | B            | 5960   5240 | 1782   1725              | 7742   6965     |
|                           | D            | 359   360   | 223   162                | 582   522       |
| <i>Tae-Mace</i> /CS       | A            | 3911   3571 | 1201   1138              | 5112   4709     |
|                           | B            | 4436   4083 | 1583   1472              | 6019   5555     |
|                           | D            | 347   353   | 185   174                | 532   527       |
| <i>Tae-Landmark</i> /CS   | A            | 3556   3484 | 1180   1154              | 4736   4638     |
|                           | B            | 4212   4038 | 1491   1464              | 5703   5502     |
|                           | D            | 539   456   | 246   253                | 785   709       |
| <i>Tae-Renan</i> /CS      | A            | 3870   3561 | 1277   1219              | 5147   4780     |
|                           | B            | 4172   3796 | 1547   1447              | 5719   5243     |
|                           | D            | 575   578   | 387   354                | 962   932       |
| <i>Tae-Stanley</i> /CS    | A            | 3578   3374 | 1242   1180              | 4820   4554     |
|                           | B            | 4840   4578 | 1620   1599              | 6460   6177     |
|                           | D            | 473   374   | 208   232                | 681   606       |
| <i>Tae-Lancer</i> /CS     | A            | 3281   3006 | 1176   1043              | 4457   4049     |
|                           | B            | 6341   5663 | 2256   2329              | 8597   7992     |
|                           | D            | 367   394   | 211   195                | 578   589       |
| <i>Tae-Jagger</i> /CS     | A            | 3170   3166 | 1155   1005              | 4325   4171     |
|                           | B            | 3669   3544 | 1344   1377              | 5013   4921     |
|                           | D            | 541   448   | 290   301                | 831   749       |
| <i>Tae-Julius</i> /CS     | A            | 3553   3393 | 1171   1143              | 4724   4536     |
|                           | B            | 4088   3853 | 1544   1520              | 5632   5373     |
|                           | D            | 509   423   | 273   251                | 782   674       |
| <i>Tae-SY_Mattis</i> /CS  | A            | 3794   3376 | 1282   1071              | 5076   4447     |
|                           | B            | 4569   3986 | 1564   1487              | 6133   5473     |
|                           | D            | 585   597   | 425   314                | 1010   911      |
| <i>Tae-ArinaLrFor</i> /CS | A            | 3900   3449 | 1260   1064              | 5160   4513     |
|                           | B            | 4450   4102 | 1172   1497              | 5622   5599     |
|                           | D            | 539   482   | 349   264                | 888   746       |
| <i>Tae-Zang1817</i> /CS   | A            | 2046   1906 | 817   749                | 2863   2655     |
|                           | B            | 2743   2675 | 1107   1055              | 3850   3730     |
|                           | D            | 280   285   | 166   148                | 446   433       |
| <i>Tae-Norin61</i> /CS    | A            | 1750   1780 | 662   532                | 2412   2312     |
|                           | B            | 2073   2073 | 827   801                | 2900   2874     |
|                           | D            | 320   336   | 193   173                | 513   509       |

Note: Number of TEs inserted in the query | subject genomes are indicated relatively to each other.

Abbreviations: CS, Chinese spring; *Tae*, *Triticum aestivum*.



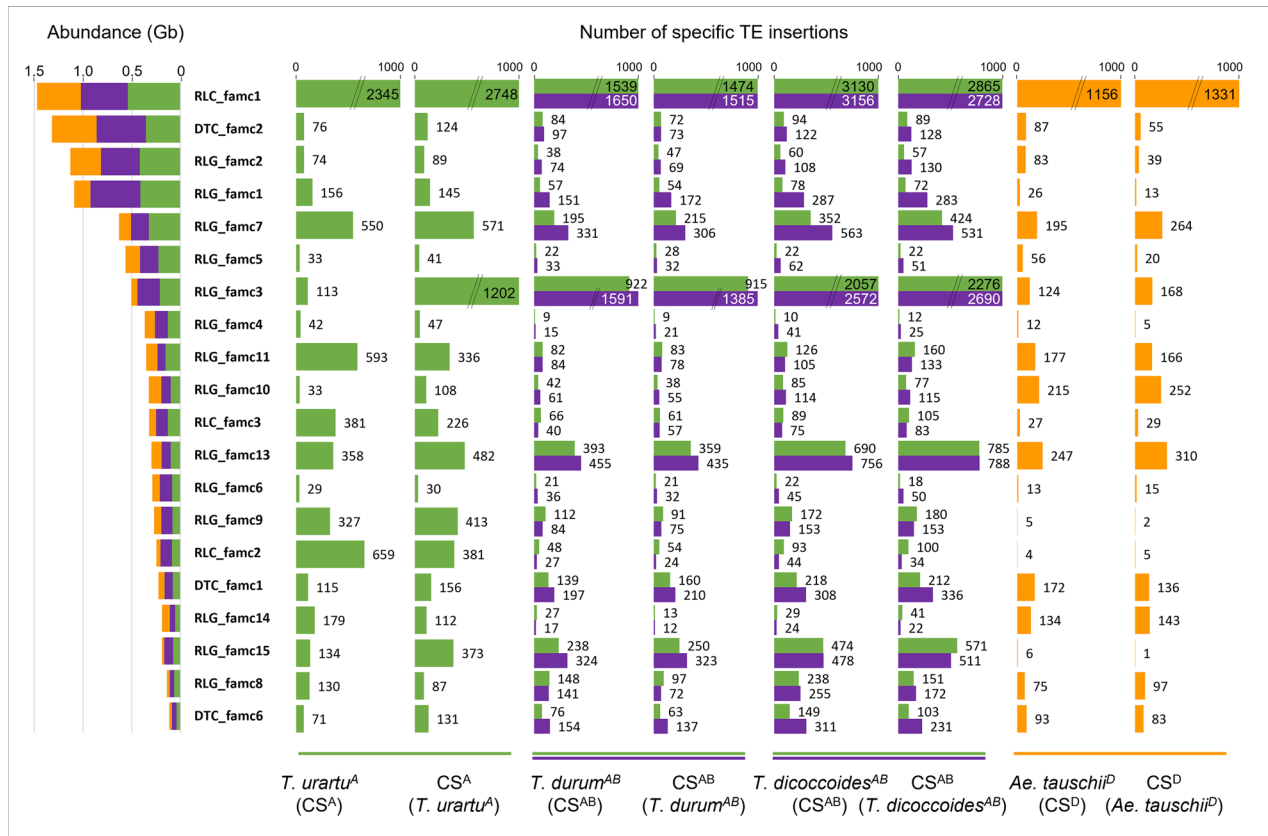
**FIGURE 3** Transposable element (TE) family enrichment in the variable fraction of the genome represented as heatmaps for the subgenomes A, B, and D. Enrichment ratios were calculated for the 100, 113, and 98 most abundant families of the A, B, and D subgenomes, respectively (representing at least 1 Mb per subgenome in Chinese Spring [CS]). Abundance of each family (in bps) was retrieved at the whole genome scale and compared to that in the variable fraction of the genome (identified from pairwise comparisons with the reference cultivar CS). Log<sub>2</sub> ratios between these two proportions were then calculated and represented as heatmaps with red showing families that could be considered as enriched in the variable fraction compared to their genome average. Families were ordered in rows according to their superfamily classification represented by a color code on the left panel and labelled with a 3-letter code. Names of five families showing log<sub>2</sub> ratios  $\geq 2$  are indicated on the left.

i.e., a molecular evidence of transposition) for 78% of them. Density of newly transposed elements follows species divergence time: 6 insertions per Mb in CS<sup>A</sup> compared to *T. urartu*<sup>A</sup>; 2.6 and 3.0 insertions/Mb compared to *T. dicoccoides* A and B subgenomes, respectively; 1.3 and 1.7 insertions/Mb compared to *T. durum* A and B, respectively; 1.3 insertions/Mb compared to *Ae. tauschii*<sup>D</sup>. These novel insertions accounted 10%–15% of the size of the specific regions identified in each subgenome. At the intraspecific level, we found that A, B, and D subgenomes of CS carry on average 4261, 5460, and 659 specific insertions compared to the 12 other accessions, representing a density of 1.0, 1.2, and 0.2/Mb, respectively. Novel insertions were more frequent in the distal regions (as defined in IWGSC, 2018) than in the central part of chromosomes for example, 8.5 versus 5.4 novel insertions/Mb in CS<sup>A</sup> compared to *T. urartu*<sup>A</sup>; 1.9/2.0/0.5 versus 0.7/1.0/0.1 novel insertions/Mb in CS A/B/D subgenomes compared to other *T. aestivum*.

A striking result is that the numbers of newly inserted TEs were quite similar between the two genomes aligned, whatever the comparison considered (Table 3). For instance, 5706 novel insertions were found in CS compared to the orthologous loci in *T. durum* and, reversely, we detected 5921 novel insertions in *T. durum* compared to CS. Similarly, comparing CS<sup>D</sup> against *Ae. tauschii*<sup>D</sup> revealed 4245 and 4449 specific insertions, respectively. This was the case for every species/accession compared to the reference. This shows that, first, transposition is not silenced in any species analyzed here, with thousands of recent events discovered, and, second, that transposition rate is somehow constant since the divergence of these genomes. We found no occurrence of enhanced TE amplification in any lineage explored here and found no impact of ploidy on TE activity. In other terms, there were no more transpositions in polyploids than in diploids.

We then wondered which TE families were the most active in this short evolutionary time frame. We found traces of





**FIGURE 4** Histograms of the number of recent transposable element (TE) insertions discovered by whole genome alignments for the 20 most abundant wheat TE families. Abundance (in Gb) of TE families annotated in Chinese spring (CS) A, B, and D subgenomes (IWGSC RefSeq v1.1) is represented on the left panel. Specific TE insertions were identified in pairwise orthologous intergenic region (oIGR) alignments and the 8 histograms represent the number of TEs that are present the query genome and absent at the orthologous locus in the compared genome which is mentioned in parentheses. Numbers of insertions are provided by subgenome with the following color code: A: green; B: violet; D: orange.

recent transposition for 346 different families, most (79%) of them having transposed in A, B, and D subgenomes. Together these active families represent 99.7% of the whole genome TE content because, although 505 families were distinguished in ClariTeRep, many families were only poorly characterized, with truncated elements, spurious predictions, or misclassified repeats, and we cannot find newly inserted copies for such families. Thus, we applied a 100 kb threshold per subgenome analyzed in order to estimate the proportion of active families in wheat. This retained 301 high confidence families and 89% of them were active. We conclude that virtually all families were active recently and gave rise to newly inserted copies in the recent *Triticeae* evolution. Even at the intraspecific level, we found traces of transposition for 328 families. This situation cannot be explained by cycles of silencing/bursts but is rather in favor of an equilibrium model of evolution.

To go further, we wondered if the level of activity of a given family was different between species/accessions and if the rate of transposition was correlated or not with the abundance of the family. Figure 4 represents the number of specific insertions per pair of aligned subgenomes for the 20

most abundant families. RLC\_famc1 (*Angela-WIS*) was the most active family representing around 25% of the recent insertions discovered in all A, B, and D genome comparisons and it is, actually, the most abundant family. But other abundant families DTC\_famc2 (*Jorge*), RLG\_famc2 (*Sabrina/Derami/Egug*), and RLG\_famc1 (*Fatima*) were much less active, representing only ~1% of the recent insertions. In contrast, RLG\_famc3 (*Laura*) and RLG\_famc13 (*Latidu*) were among the most active families (reaching 22% of the insertions) while they are less abundant. We conclude that there is no positive nor negative correlation between the recent transposition rate and the family abundance. Again, a striking result is that this pattern is conserved in the compared genomes. For instance, since the divergence of *T. urartu*<sup>A</sup> and CS<sup>A</sup>, 156 copies of RLG\_famc1 (*Fatima*) transposed specifically in *T. urartu*<sup>A</sup> and 145 in CS<sup>A</sup>. Comparing *Ae. tauschii*<sup>D</sup> with CS<sup>D</sup> revealed 26 and 13 novel *Fatima* copies in CS<sup>D</sup> and *Ael*<sup>D</sup>. These values were also quite similar when comparing CS with tetraploids (Figure 4). For instance, *T. dicoccoides*<sup>AB</sup> carries 474 (on A) and 478 (on B) novel copies of RLG\_famc15 (*Jeli*) absent at orthologous loci

CS<sup>AB</sup>, but, reversely, CS<sup>AB</sup> carries 571 and 511 novel *Jeli* copies that are absent at orthologous loci in *T. dicoccoides*<sup>AB</sup>. Such unexpected similarity is not limited to *Fatima* and *Jeli* but rather true for almost all the other families in all species. However, we found a case, RLG\_famc3 (*Laura*), that deviates from this pattern since it transposed 10 times more in CS<sup>A</sup> than in *T. urartu*<sup>A</sup> (1202 vs 113 specific insertions, respectively). This makes us conclude that, if the global transposition rate appears constant during the recent *Triticeae* history, each family amplified at a specific rate, which is not simply explained by its abundance, but this rate tends to remain constant across speciation events. Polyploidy is not associated with more copies accumulated over time and transposition rate of each family is not even disturbed following polyploidization events that is, diploids, tetraploids, and hexaploids accumulated novel TEs at similar rates. Globally, all TE families generate novel copies, independently, at different genomic locations, but at approximately the same rate, in the different lineages.

### 3.5 | Proximity to genes and insertion dates

In total, our bp-resolved alignments revealed the presence of 51,928 elements (39,966 class 1, 11,048 class 2, and 914 unclassified) in the Chinese Spring genome that are absent from the corresponding insertion site in one or more compared genomes. They represent a large and clean dataset of novel insertions, although some of the class 2 elements detected here may be traces of recent excision in the compared genome. We used this dataset in order to investigate the global distribution of new insertions and the potential preferential insertion in gene vicinity. Among the 102,601 analyzed oIGRs, 26,862 contained one or more new insertions, showing that there was no insertion hotspot but rather that novel insertions spread the genome homogeneously. In wheat, genes tend to be clustered into small islands separated by large TE clusters, so it was important to distinguish small IGRs ( $\leq 40$  kb; 51,120 regions corresponding to gene islands) from large ones ( $>40$  kb; 51,481 regions corresponding to large TE clusters). Small IGRs represent only 4% of the TE space but accumulated 10% of the novel insertions: 5414 recent insertions scattered into 4530 small IGRs. There were also 46,514 novel insertions in 22,332 distinct large IGRs with, again, a number of novel insertions in the close vicinity of genes higher than expected by random insertion: 11% of insertions occurred within 5 kb around genes, that is, 4% of target sequences. Affinity with genic regions depends on the TE family considered. We thus analyzed the insertional behavior for 44 TE families for which we had at least 100 novel insertions detected in CS. Half of them exhibit preferential insertion around genes (at least three times more insertions close to genes than expected randomly). They belong to class

two transposons and LINEs retrotransposon families that were previously shown to be enriched in gene promoters (Wicker et al., 2018). In contrast, Gypsy and Copia insertions tend to be excluded from the gene vicinity and insert preferentially in the core of large IGRs.

Finally, we estimated the age of these novel insertions with the approach, that is, widely used by the community: aligning 5' and 3' LTRs of LTR-retrotransposons and applying a molecular clock, considering that both LTRs are replicated (and thus identical) during transposition process. The purpose was to validate the reliability of this approach since we have sampled here the largest dataset of recently inserted LTR-RTs. The average insertion time estimated for the 88,269 newly inserted LTR-RTs identified in all our comparisons, is 590,000 years. Only 4% are estimated to be younger than 100,000 years ago. This estimate is surprisingly older than expected and not in accordance with the divergence time estimated for these species/accessions. Thus, we selected 209 LTR-RTs that are strictly specific to CS while absent from all other species/accessions, to ensure we collected very recent ones (hundreds/thousands of years). Actually, only five out of 209 carry identical LTRs while all others already exhibit sequence differences. Moreover, 95% (198/209) are estimated to be older than 100,000 years, an inconsistency with the fact that they transposed specifically in CS. This raises questions about the error-less replication of LTRs upon insertion in *Triticeae*, which may have implications for our understanding of TE evolution.

## 4 | DISCUSSION

Assembling the wheat genome has long been a challenge but we have now reached the pangenomics area with multiple high-quality genome assemblies available. SNP diversity was intensively characterized in order to get a world-wide view of the *Triticum* population structure, impact of selection, introgressions (Balfourier et al., 2019; Zhou et al., 2020), and to even build haplotype maps for genotype imputations (Brinton et al., 2020; Jordan et al., 2022). SNPs are easy to discover and to genotype because bioinformatics pipelines that handle short-reads are well established and because technology advances tackle the complexity of the genome. However, lots remains to be done to go beyond the type of diversity we could investigate with SNPs which is basically no more than allele combinations. This is the goal of pangenomics, which relies on discovering the hidden diversity with loci under presence/absence variations. For that, it is important to go beyond the “uniquely mappable area” of short-read-based bioinformatics, especially when studying complex genomes. Structural variations (SVs) were only poorly characterized for wheat genes and partially addressed for TEs (De Oliveira et al., 2020; Montenegro et al., 2017). SVs are, however, of

major importance to understand the molecular factors responsible for phenotypic variations and to understand evolutionary rules governing TE dynamics. Here, we faced the challenge of characterizing variability affecting Gbs of repeated elements in one of the most complex genomes ever assembled. Addressing this question required dedicated tools, strategies, and expertise in TE classification and annotation. We homogenized TE annotations of all available genome sequences using CLARITE and ClariTeRep library that were specifically developed for modeling wheat TEs (Daron et al., 2014) and previously used to comparing A-B-D TE content (Wicker et al., 2018). TE annotations are actually very dependent on the tools and library used so it is hazardous to compare annotations performed by different groups. Most striking difference comes from CACTA transposons because the ClariTeRep library is enriched in CACTAs manually curated (Choulet et al., 2010) that are generally absent from other wheat-specific TE libraries. Indeed, several *Triticeae* sequencing projects concluded that CACTAs represent 5%–6% of the genome (Jia et al., 2013; Ling et al., 2013), while their proportion is around 15% when using ClariTeRep.

We established a workflow in order to compute base pair-resolved whole-genome sequence alignments for Gb-sized genomes, by taking advantage of the high level of gene collinearity of *Triticeae* genomes to anchor the alignments. We split the subgenomes into ~30,000 intervals corresponding to individual orthologous intergenic regions flanked by pairs of collinear orthologs, thus reducing the alignment space so that most TEs are not repeated within the aligned interval. This, however, excluded noncollinear regions from the analyses and it was important to check whether such a filter introduces bias by excluding regions that may be more variable than the genome average. This is why we checked with unbiased estimators of variability: comparing genome-wide family proportions and mapping of all ISBP markers. Results from oIGR alignments were consistent with these unbiased estimates. First, TE families showing copy number differences at the whole genome level were also the ones enriched in the variable TE fraction defined by the alignments. Second, the extent of variability defined by oIGR alignments was fully in line with the unbiased estimates based on ISBP mapping although slightly (2% on average) lower, confirming that noncollinear (non-aligned) regions contain more SVs than the genome average. However, this quality control demonstrated that we can be confident that the conclusions reached by interval-based whole genome alignments are sound.

Our conclusions are in favor of an equilibrium model (Cosby et al., 2019; Stritt et al., 2018) for the *Triticeae* which recent genome diversification was not governed by dramatic changes due to cycles of TE bursts/silencing (Rey et al., 2016). First pieces of evidence that wheat TE dynamics is a continuous process was brought by comparing A with B and D subgenomes (Wicker et al., 2018). These lineages diverged

millions of years ago so that TE turnover is nearly complete, that is, there are (almost) no more orthologous TEs at that scale. Although TEs evolved independently in diploids, many striking features appeared: the wide majority of TE families maintained their ancestral copy-number with abundant families remaining abundant, and low-copy families remaining at low-copy number in the three lineages. Moreover, families that tend to be enriched close to genes kept this behavior in all lineages, although novel copies did not target the same genes. All features led to conclude that TE dynamics was highly regulated and suggested evolutionary constraints to maintain global equilibrium. This is this model that we tentatively challenged here by characterizing the initial events of the ongoing TE turnover in three *Triticeae* lineages and the impact of polyploidy on TE activity. Our data revealed that the TE composition of the variable TE fraction resembles the ancestral conserved one. This confirms that TE turnover does not modify the global TE landscape. In contrast, it occurs while maintaining an equilibrium with unchanged TE family proportions. We did not see any burst of a given family nor families that would have decreased in proportion because of being completely silenced.

We wondered how the availability of one single genotype per tetraploid and diploid species impacted our ability to conclude. We consider that frontiers between *Triticeae* species are not clear and more related to ploidy than to sequence divergence per se. For instance, the extent of variability between CS and *T. durum* is not substantially higher than between two *T. aestivum* accessions. The important point in our study was to explore variability at different time scales across the tree, whatever the species definition, as long as one can align orthologous TEs to retrieve structural variations corresponding to obvious events of TE insertions and deletions. The four genotypes of tetraploid and diploid species might be not the best representative and one cannot exclude that some genotypes may not follow the evolutionary rules we commented here. Intraspecific level was assessed for *T. aestivum* only and it would be interesting to estimate the extent of intraspecific variability of the wild species also when more data will be available.

Rare cases of families that contribute more than others to recent genome diversification were observed. Interestingly, this was the case for subtelomeric satellite repeats (XXX\_famc10) and centromeric retrotransposons Cereba/Quinta (RLG\_famc8) (Presting et al., 1998). This suggests that such elements that play a major structural role in shaping chromosome architecture comprise the most rapidly evolving elements of the genome.

Specific regions originate from both TE deletions and insertions. An important outcome is that, whatever the genomes compared and the ploidy levels, the proportion of specific sequences was similar in both the query and reference. We did not see a species that would have accumulated more novel

insertions than others. The TE turnover seems to occur at a conserved rate in the different lineage which is, again, in favor of the equilibrium model with a controlled genome size. This strongly suggests that TE deletions compensate genome expansion by novel insertions over time. This observation is in concordance with the experimentally observed balance between deletions and insertions during DNA double-strand break repair at targeted DSBs in barley (Vu et al., 2017). Since we reached a bp-resolution by BLAST alignments, we were able to search for SVs that are traces of a recent insertion/excision. Although we found thousands of such insertions, their accumulated size reached only 10% of the specific fraction (in bps), while we would expect around 50% since insertions and deletions were likely balanced. The reason for this is most probably because criteria to call a novel insertion were too strict to avoid false positives. In addition, subsequent events of deletions/rearrangements either in the query or reference may have erased molecular traces of such recent insertions. Such TE dynamics is not a generality for grasses. Comparing *O. brachyantha* and *O. sativa* genomes revealed substantial size variations due to recent (<2 Myr) TE bursts in rice and a higher deletion rate in *O. brachyantha* (Chen et al., 2013). Comparing 13 genomes of wild and cultivated *Oryza* species also revealed rapid TE-driven diversification via lineage-specific amplifications and preference of deletion over insertion (Stein et al., 2018), confirming previous observations (Piegu et al., 2006; Vitte et al., 2007). Differences between *Triticeae* and *Oryza* cannot be associated with genome size since studies in *Brachypodium* were more in agreement with what we observed in *Triticeae* (Stritt et al., 2018).

Another interesting result is that polyploids have not experienced more transpositions than their diploid relatives. Diploids and polyploids accumulated a very similar number of TE insertions per subgenome, confirming there was no genomic shock that would have been followed by TE deregulation. In contrast, our results tend to show that the rate of TE transposition and, thus, TE turnover, is stable over time and whole genome duplications did not destabilize this equilibrium. However, this model is not in agreement with the conclusions of a differential accumulation of TE families in *Triticum/Aegilops* (Keidar-Friedman et al., 2018; Yaakov et al., 2013). One of the main sources of such discrepancies is the definition of what we call a family. Previous studies observed lineage-specific accumulations of families whereas we observed an equilibrium. In fact, we agree that families evolved into variants that is, subfamilies, that have accumulated independently in different lineages. But many families that were called by different names are actually related, sometimes can even be considered members of the same families. Deciphering these phylogenetic relations between copies is crucial to understand that, even if it is possible to see variants overrepresented in one lineage, the important and striking

result is that its family remained at an equilibrium in all lineages. The importance of connecting elements was also underlined in (Stritt et al., 2021) where it has been shown that a rare TE family tend to maintained vertically at low-copy number throughout angiosperms, suggesting an alternative scenario prevailing the usual view of TE dynamics. Also similar to what we found, TE insertion polymorphisms explored in 53 *Brachypodium* accessions did not reveal any lineage-specific TE activations but rather a homogeneous activity and a stable transposition rate among populations, suggesting a conserved regulatory mechanism (Stritt et al., 2018). TE proliferation in *Triticeae* does not appear to be the result of an induction by the environment but rather a highly regulated process playing a role in basic genome functioning. This challenges the previous view of TEs as “invasive” elements as it becomes more and more suggested in the literature by the perspective of the genome ecology where TEs would persist by creating their own niche (Kremer et al., 2020). This has been recently commented in maize where the analysis of family-level ecology of the genome led to conclude that, like in our study, most families have been active recently, each family having its own survival strategy representing the evolution of a distinct ecological niche (Stitzer et al., 2021). Despite the short evolutionary time frame studied here, we discovered traces of novel insertions for 346 families, that is, meaning that virtually all families are active actually. Same conclusions were reached in rice where ca. 11,000 new insertions of 251 families were detected in genomes of 3000 natural rice accessions (Liu et al., 2021). This highlighted the discrepancies between artificial stress-induced transposition and activity under natural conditions which provide two different views. Similarly, we wondered why our results are not in agreement with TE behavior observed in synthetic newly formed polyploid wheats where some TEs were mobilized in response to polyploidy (Yaakov & Kashkush, 2012; Yaakov et al., 2013b). As suggested earlier, instability following polyploidy is unlikely to be selected for under natural conditions (Zhao et al., 2011) and we support the hypothesis that only polyploids for which the equilibrium remained stable maintained across the generations.

Finally, the fact that recently inserted LTR-RTs did not carry identical LTRs made us underline some doubts about the method used to estimate dates of insertions. Similar inconsistencies were reported based on the analysis of ca. 25,000 LTR-RTs in 15 plant species which revealed that one-fourth of nested retrotransposons exhibited older insertion date than pre-existing elements in which they had been inserted (Jedlicka et al., 2020). Gene conversion was suggested to be involved in erasing divergence over time. Another possible explanation would be that replication of LTRs during the insertion mechanism is error-prone. This may have strong impact on the way the community uses the molecular clock to date insertion events. An alternative explanation could be that



cryptic polymorphisms in the population of wild progenitors have been conserved and appear now as old polymorphisms among wheat cultivars. Finally, recent introgressions from wild species may have contributed to identify novel insertions that actually occurred earlier in the history of the donor.

## AUTHOR CONTRIBUTIONS

**Nathan Papon:** Conceptualization; data curation; formal analysis; investigation; methodology; resources; software; validation; visualization; writing—original draft; writing—review & editing. **Pauline Lasserre-Zuber:** Formal analysis; writing—review & editing. **Hélène Rimbart:** Formal analysis; software; writing—review & editing. **Romain De Oliveira:** Formal analysis; writing—review & editing. **Etienne Paux:** Conceptualization; funding acquisition; project administration; supervision; writing—review & editing. **Frédéric Choulet:** Conceptualization; data curation; formal analysis; methodology; project administration; supervision; writing—original draft; writing—review & editing.

## ACKNOWLEDGMENTS

The authors are grateful to the Mésocentre Clermont-Auvergne and the platform AuBi of the Université Clermont Auvergne for providing help, computing, and storage resources. NP PhD was financed by a grant from the French Ministry of Higher Education, Research and Innovation (MESRI).

## CONFLICT OF INTEREST STATEMENT

The authors declare no conflicts of interests.

## DATA AVAILABILITY STATEMENT

Data generated (GFFs files of TE annotation, BED files of positions of orthologous regions aligned, positions of variable regions, and positions of novel insertions) were deposited in <https://entrepot.recherche.data.gouv.fr> under the <https://doi.org/10.57745/RCTOQM>. Scripts used are available on Gitlab at [https://forgemia.inra.fr/umr-gdec/scripts\\_files/](https://forgemia.inra.fr/umr-gdec/scripts_files/).

## ORCID

Hélène Rimbart  <https://orcid.org/0000-0002-2288-6864>

Frédéric Choulet  <https://orcid.org/0000-0003-1788-7288>

## REFERENCES

- Aury, J.-M., Engelen, S., Istace, B., Monat, C., Lasserre-Zuber, P., Belser, C., Cruaud, C., Rimbart, H., Leroy, P., Arribat, S., Dufau, I., Bellec, A., Grimichler, D., Papon, N., Paux, E., Ranoux, M., Alberti, A., Wincker, P., & Choulet, F. (2022). Long-read and chromosome-scale assembly of the hexaploid wheat genome achieves high resolution for research and breeding. *Gigascience*, *11*, giac034. <https://doi.org/10.1093/gigascience/giac034>
- Avni, R., Lux, T., Minz-Dub, A., Millet, E., Sela, H., Distelfeld, A., Deek, J., Yu, G., Steuernagel, B., Pozniak, C., Ens, J., Gundlach, H., Mayer, K. F. X., Himmelbach, A., Stein, N., Mascher, M., Spannagl, M., Wulff, B. B. H., & Sharon, A. (2022). Genome sequences of three *Aegilops* species of the section *Sitopsis* reveal phylogenetic relationships and provide resources for wheat improvement. *Plant Journal*, *110*, 179–192. <https://doi.org/10.1111/tbj.15664>
- Avni, R., Nave, M., Barad, O., Baruch, K., Twardziok, S. O., Gundlach, H., Hale, I., Mascher, M., Spannagl, M., Wiebe, K., Jordan, K. W., Golan, G., Deek, J., Ben-Zvi, B., Ben-Zvi, G., Himmelbach, A., Maclachlan, R. P., Sharpe, A. G., Fritz, A., ... Distelfeld, A. (2017). Wild emmer genome architecture and diversity elucidate wheat evolution and domestication. *Science*, *357*, 93–97. <https://doi.org/10.1126/science.aan0032>
- Baduel, P., & Quadrana, L. (2021). Jumpstarting evolution: How transposition can facilitate adaptation to rapid environmental changes. *Current Opinion in Plant Biology*, *61*, 102043. <https://doi.org/10.1016/j.cpb.2021.102043>
- Balfourier, F., Bouchet, S., Robert, S., De Oliveira, R., Rimbart, H., Kitt, J., Choulet, F., & Paux, E. (2019). Worldwide phylogeography and history of wheat genetic diversity. *Science Advances*, *5*, eaav0536. <https://doi.org/10.1126/sciadv.aav0536>
- Bariah, I., Keidar-Friedman, D., & Kashkush, K. (2020). Where the wild things are: Transposable elements as drivers of structural and functional variations in the wheat genome. *Frontiers in Plant Science*, *11*, 585515. <https://doi.org/10.3389/fpls.2020.585515>
- Brinton, J., Ramirez-Gonzalez, R. H., Simmonds, J., Wingen, L., Orford, S., Griffiths, S., Haberer, G., Spannagl, M., Walkowiak, S., Pozniak, C., & Uauy, C. (2020). A haplotype-led approach to increase the precision of wheat breeding. *Communications Biology*, *3*, 712. <https://doi.org/10.1038/s42003-020-01413-2>
- Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*, 421. <https://doi.org/10.1186/1471-2105-10-421>
- Chen, J., Huang, Q., Gao, D., Wang, J., Lang, Y., Liu, T., Li, B., Bai, Z., Luis Goicoechea, J., Liang, C., Chen, C., Zhang, W., Sun, S., Liao, Y., Zhang, X., Yang, L., Song, C., Wang, M., Shi, J., ... Chen, M. (2013). Whole-genome sequencing of *Oryza brachyantha* reveals mechanisms underlying *Oryza* genome evolution. *Nature Communications*, *4*, 1595. <https://doi.org/10.1038/ncomms2596>
- Choulet, F., Wicker, T., Rustenholz, C., Paux, E., Salse, J., Leroy, P., Schlub, S., Le Paslier, M.-C., Magdelenat, G., Gonthier, C., Couloux, A., Budak, H., Breen, J., Pumphrey, M., Liu, S., Kong, X., Jia, J., Gut, M., Brunel, D., ... Feuillet, C. (2010). Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*, *22*, 1686–1701. <https://doi.org/10.1105/tpc.110.074187>
- Cosby, R. L., Chang, N.-C., & Feschotte, C. (2019). Host-transposon interactions: Conflict, cooperation, and cooption. *Genes & Development*, *33*, 1098–1116. <https://doi.org/10.1101/gad.327312.119>
- Daron, J., Glover, N., Pingault, L., Theil, S., Jamilloux, V., Paux, E., Barbe, V., Mangenot, S., Alberti, A., Wincker, P., Quesneville, H., Feuillet, C., & Choulet, F. (2014). Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biology*, *15*, 546. <https://doi.org/10.1186/s13059-014-0546-4>
- De Oliveira, R., Rimbart, H., Balfourier, F., Kitt, J., Dynamant, E., Vrána, J., Doležel, J., Cattonaro, F., Paux, E., & Choulet, F. (2020). Structural variations affecting genes and transposable elements of chromosome 3B in wheats. *Frontiers in Genetics*, *11*, 891. <https://doi.org/10.3389/fgene.2020.00891>

- Dvorak, J., Wang, L., Zhu, T., Jorgensen, C. M., Luo, M.-C., Deal, K. R., Gu, Y. Q., Gill, B. S., Distelfeld, A., Devos, K. M., Qi, P., & Mcguire, P. E. (2018). Reassessment of the evolution of wheat chromosomes 4A, 5A, and 7B. *Theoretical and Applied Genetics*, *131*, 2451–2462. <https://doi.org/10.1007/s00122-018-3165-8>
- Glémin, S., Scornavacca, C., Dainat, J., Burgarella, C., Viader, V., Ardisson, M., Sarah, G., Santoni, S., David, J., & Ranwez, V. (2019). Pervasive hybridizations in the history of wheat relatives. *Science Advances*, *5*, eaav9188. <https://doi.org/10.1126/sciadv.aav9188>
- Guo, W., Xin, M., Wang, Z., Yao, Y., Hu, Z., Song, W., Yu, K., Chen, Y., Wang, X., Guan, P., Appels, R., Peng, H., Ni, Z., & Sun, Q. (2020). Origin and adaptation to high altitude of Tibetan semi-wild wheat. *Nature Communications*, *11*, 5085. <https://doi.org/10.1038/s41467-020-18738-5>
- IWGSC. (2018). Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*, *361*, eaar7191.
- Jedlicka, P., Lexa, M., & Kejnovsky, E. (2020). What can long terminal repeats tell us about the age of Itr retrotransposons, gene conversion and ectopic recombination? *Frontiers in Plant Science*, *11*, 644. <https://doi.org/10.3389/fpls.2020.00644>
- Jia, J., Zhao, S., Kong, X., Li, Y., Zhao, G., He, W., Appels, R., Pfeifer, M., Tao, Y., Zhang, X., Jing, R., Zhang, C., Ma, Y., Gao, L., Gao, C., Spannagl, M., Mayer, K. F. X., Li, D., Pan, S., ... Consortium, I.W.G.S. (2013). *Aegilops tauschii* draft genome sequence reveals a gene repertoire for wheat adaptation. *Nature*, *496*, 91–95. <https://doi.org/10.1038/nature12028>
- Jordan, K. W., Bradbury, P. J., Miller, Z. R., Nyine, M., He, F., Fraser, M., Anderson, J., Mason, E., Katz, A., Pearce, S., Carter, A. H., Prather, S., Pumphrey, M., Chen, J., Cook, J., Liu, S., Rudd, J. C., Wang, Z., Chu, C., ... Akhunov, E. D. (2022). Development of the wheat practical haplotype graph database as a resource for genotyping data storage and genotype imputation. *G3 (Bethesda)*, *12*, jkab390.
- Keidar-Friedman, D., Bariah, I., & Kashkush, K. (2018). Genome-wide analyses of miniature inverted-repeat transposable elements reveals new insights into the evolution of the *Triticum-Aegilops* group. *PLoS One*, *13*, e0204972. <https://doi.org/10.1371/journal.pone.0204972>
- Kraitshtein, Z., Yaakov, B., Khasdan, V., & Kashkush, K. (2010). Genetic and epigenetic dynamics of a retrotransposon after allopolyploidization of wheat. *Genetics*, *186*, 801–812. <https://doi.org/10.1534/genetics.110.120790>
- Kremer, S. C., Linnquist, S., Saylor, B., Elliott, T. A., Gregory, T. R., & Cottenie, K. (2020). Transposable element persistence via potential genome-level ecosystem engineering. *BMC Genomics*, *21*, 367. <https://doi.org/10.1186/s12864-020-6763-1>
- Levy, A. A., & Feldman, M. (2022). Evolution and origin of bread wheat. *Plant Cell*, *34*, 2549–2567. <https://doi.org/10.1093/plcell/koac130>
- Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*, 3094–3100. <https://doi.org/10.1093/bioinformatics/bty191>
- Li, L.-F., Zhang, Z.-B., Wang, Z.-H., Li, N., Sha, Y., Wang, X.-F., Ding, N., Li, Y., Zhao, J., Wu, Y., Gong, L., Mafessoni, F., Levy, A. A., & Liu, B. (2022). Genome sequences of five Sitopsis species of *Aegilops* and the origin of polyploid wheat B subgenome. *Molecular Plant*, *15*, 488–503. <https://doi.org/10.1016/j.molp.2021.12.019>
- Ling, H.-Q., Ma, B., Shi, X., Liu, H., Dong, L., Sun, H., Cao, Y., Gao, Q., Zheng, S., Li, Y., Yu, Y., Du, H., Qi, M., Li, Y., Lu, H., Yu, H., Cui, Y., Wang, N., Chen, C., ... Liang, C. (2018). Genome sequence of the progenitor of wheat A subgenome *Triticum urartu*. *Nature*, *557*, 424–428. <https://doi.org/10.1038/s41586-018-0108-0>
- Ling, H.-Q., Zhao, S., Liu, D., Wang, J., Sun, H., Zhang, C., Fan, H., Li, D., Dong, L., Tao, Y., Gao, C., Wu, H., Li, Y., Cui, Y., Guo, X., Zheng, S., Wang, B., Yu, K., Liang, Q., ... Wang, J. (2013). Draft genome of the wheat A-genome progenitor *Triticum urartu*. *Nature*, *496*, 87–90. <https://doi.org/10.1038/nature11997>
- Lisch, D. (2013). How important are transposons for plant evolution? *Nature Reviews Genetics*, *14*, 49–61. <https://doi.org/10.1038/nrg3374>
- Liu, Z., Zhao, H., Yan, Y., Wei, M.-X., Zheng, Y.-C., Yue, E.-K., Alam, M. S., Smartt, K. O., Duan, M.-H., & Xu, J.-H. (2021). Extensively current activity of transposable elements in natural rice accessions revealed by singleton insertions. *Frontiers in Plant Science*, *12*, 745526. <https://doi.org/10.3389/fpls.2021.745526>
- Luo, M.-C., Gu, Y. Q., Puiu, D., Wang, H., Twardziok, S. O., Deal, K. R., Huo, N., Zhu, T., Wang, L., Wang, Y., Mcguire, P. E., Liu, S., Long, H., Ramasamy, R. K., Rodriguez, J. C., Van, S. L., Yuan, L., Wang, Z., Xia, Z., ... Dvořák, J. (2017). Genome sequence of the progenitor of the wheat D genome *Aegilops tauschii*. *Nature*, *551*, 498–502. <https://doi.org/10.1038/nature24486>
- Ma, J., & Bennetzen, J. L. (2004). Rapid recent growth and divergence of rice nuclear genomes. *PNAS*, *101*, 12404–12410. <https://doi.org/10.1073/pnas.0403715101>
- Maccaferri, M., Harris, N. S., Twardziok, S. O., Pasam, R. K., Gundlach, H., Spannagl, M., Ormanbekova, D., Lux, T., Prade, V. M., Milner, S. G., Himmelbach, A., Mascher, M., Bagnaresi, P., Faccioli, P., Cozzi, P., Lauria, M., Lazzari, B., Stella, A., Manconi, A., ... Cattivelli, L. (2019). Durum wheat genome highlights past domestication signatures and future improvement targets. *Nature Genetics*, *51*, 885–895. <https://doi.org/10.1038/s41588-019-0381-3>
- Marcussen, T., Sandve, S. R., Heier, L., Spannagl, M., Pfeifer, M., Jakobsen, K. S., Wulff, B. B. H., Steuernagel, B., Mayer, K. F. X., Olsen, O.-A., Rogers, J., Doležel, J., Pozniak, C., Eversole, K., Feuillet, C., Gill, B., Friebe, B., Lukaszewski, A. J., Sourdille, P., ... Praud, S. (2014). Ancient hybridizations among the ancestral genomes of bread wheat. *Science*, *345*, 1250092. <https://doi.org/10.1126/science.1250092>
- Middleton, C. P., Senerchia, N., Stein, N., Akhunov, E. D., Keller, B., Wicker, T., & Kilian, B. (2014). Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PLoS One*, *9*, e85761. <https://doi.org/10.1371/journal.pone.0085761>
- Montenegro, J. D., Golicz, A. A., Bayer, P. E., Hurgobin, B., Lee, H., Chan, C.-K. K., Visendi, P., Lai, K., Doležel, J., Batley, J., & Edwards, D. (2017). The pangenome of hexaploid bread wheat. *Plant Journal*, *90*, 1007–1013. <https://doi.org/10.1111/tpj.13515>
- Paux, E., Roger, D., Badaeva, E., Gay, G., Bernard, M., Sourdille, P., & Feuillet, C. (2006). Characterizing the composition and evolution of homoeologous genomes in hexaploid wheat through BAC-end sequencing on chromosome 3B. *Plant Journal*, *48*, 463–474. <https://doi.org/10.1111/j.1365-313X.2006.02891.x>
- Piegu, B., Guyot, R., Picault, N., Roulin, A., Saniyal, A., Kim, H., Collura, K., Brar, D. S., Jackson, S., Wing, R. A., & Panaud, O. (2006). Doubling genome size without polyploidization: Dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Research*, *16*, 1262–1269. <https://doi.org/10.1101/gr.5290206>
- Presting, G. G., Malysheva, L., Fuchs, J., & Schubert, I. (1998). A *Ty3/gypsy* retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant Journal*, *16*, 721–728. <https://doi.org/10.1046/j.1365-313x.1998.00341.x>

- Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. <https://doi.org/10.1093/bioinformatics/btq033>
- Rey, O., Danchin, E., Mirouze, M., Loot, C., & Blanchet, S. (2016). Adaptation to global change: A transposable element-epigenetics perspective. *Trends in Ecology & Evolution*, *31*, 514–526. <https://doi.org/10.1016/j.tree.2016.03.013>
- Smit, A. F. A., Hubley, R., & Green, P. (1996–2004). *RepeatMasker Open-3.0*. <http://www.repeatmasker.org>
- Stein, J. C., Yu, Y., Copetti, D., Zwickl, D. J., Zhang, L., Zhang, C., Chougule, K., Gao, D., Iwata, A., Goicoechea, J. L., Wei, S., Wang, J., Liao, Y., Wang, M., Jacquemin, J., Becker, C., Kudrna, D., Zhang, J., Londono, C. E. M., ... Wing, R. A. (2018). Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus *Oryza*. *Nature Genetics*, *50*, 285–296. <https://doi.org/10.1038/s41588-018-0040-0>
- Stitzer, M. C., Anderson, S. N., Springer, N. M., & Ross-Ibarra, J. (2021). The genomic ecosystem of transposable elements in maize. *Plos Genetics*, *17*, e1009768. <https://doi.org/10.1371/journal.pgen.1009768>
- Stritt, C., Gordon, S. P., Wicker, T., Vogel, J. P., & Roulin, A. C. (2018). 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*, 304–318. <https://doi.org/10.1093/gbe/evx276>
- Stritt, C., Thieme, M., & Roulin, A. C. (2021). Rare transposable elements challenge the prevailing view of transposition dynamics in plants. *American Journal of Botany*, *108*, 1310–1314. <https://doi.org/10.1002/ajb2.1709>
- Vitte, C., Panaud, O., & Quesneville, H. (2007). LTR retrotransposons in rice (*Oryza sativa*, L.): Recent burst amplifications followed by rapid DNA loss. *BMC Genomics*, *8*, 218. <https://doi.org/10.1186/1471-2164-8-218>
- Vu, G. T. H., Cao, H. X., Reiss, B., & Schubert, I. (2017). Deletion-bias in DNA double-strand break repair differentially contributes to plant genome shrinkage. *New Phytologist*, *214*, 1712–1721. <https://doi.org/10.1111/nph.14490>
- Walkowiak, S., Gao, L., Monat, C., Haberer, G., Kassa, M. T., Brinton, J., Ramirez-Gonzalez, R. H., Kolodziej, M. C., Delorean, E., Thabugala, D., Klymiuk, V., Byrns, B., Gundlach, H., Bandi, V., Siri, J. N., Nilsen, K., Aquino, C., Himmelbach, A., Copetti, D., ... Pozniak, C. J. (2020). Multiple wheat genomes reveal global variation in modern breeding. *Nature*, *588*, 277–283. <https://doi.org/10.1038/s41586-020-2961-x>
- Wicker, T., Gundlach, H., Spannagl, M., Uauy, C., Borrill, P., Ramirez-González, R. H., De Oliveira, R., Mayer, K. F. X., Paux, E., & Choulet, F. (2018). Impact of transposable elements on genome structure and evolution in bread wheat. *Genome Biology*, *19*, 103. <https://doi.org/10.1186/s13059-018-1479-0>
- Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., Sanmigué, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*, 973–982. <https://doi.org/10.1038/nrg2165>
- Wicker, T., Stritt, C., Sotiropoulos, A. G., Poretti, M., Pozniak, C., Walkowiak, S., Gundlach, H., & Stein, N. (2022). Transposable element populations shed light on the evolutionary history of wheat and the complex co-evolution of autonomous and non-autonomous retrotransposons. *Advanced Genetics*, *3*, 2100022. <https://doi.org/10.1002/ggn2.202100022>
- Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*, 1859–1875. <https://doi.org/10.1093/bioinformatics/bti310>
- Yaakov, B., Ben-David, S., & Kashkush, K. (2013). Genome-wide analysis of stowaway-like MITEs in wheat reveals high sequence conservation, gene association, and genomic diversification. *Plant Physiology*, *161*, 486–496. <https://doi.org/10.1104/pp.112.204404>
- Yaakov, B., & Kashkush, K. (2011). Massive alterations of the methylation patterns around DNA transposons in the first four generations of a newly formed wheat allohexaploid. *Genome*, *54*, 42–49. <https://doi.org/10.1139/G10-091>
- Yaakov, B., & Kashkush, K. (2012). Mobilization of stowaway-like MITEs in newly formed allohexaploid wheat species. *Plant Molecular Biology*, *80*, 419–427. <https://doi.org/10.1007/s11103-012-9957-3>
- Yaakov, B., Meyer, K., Ben-David, S., & Kashkush, K. (2013). Copy number variation of transposable elements in *Triticum-Aegilops* genus suggests evolutionary and revolutionary dynamics following allopolyploidization. *Plant Cell Reports*, *32*, 1615–1624. <https://doi.org/10.1007/s00299-013-1472-8>
- Zhao, N., Zhu, B., Li, M., Wang, L., Xu, L., Zhang, H., Zheng, S., Qi, B., Han, F., & Liu, B. (2011). Extensive and heritable epigenetic remodeling and genetic stability accompany allohexaploidization of wheat. *Genetics*, *188*, 499–510. <https://doi.org/10.1534/genetics.111.127688>
- Zhou, Y., Zhao, X., Li, Y., Xu, J., Bi, A., Kang, L., Xu, D., Chen, H., Wang, Y., Wang, Y.-G., Liu, S., Jiao, C., Lu, H., Wang, J., Yin, C., Jiao, Y., & Lu, F. (2020). *Triticum* population sequencing provides insights into wheat adaptation. *Nature Genetics*, *52*, 1412–1422. <https://doi.org/10.1038/s41588-020-00722-w>
- Zhu, T., Wang, L., Rodriguez, J. C., Deal, K. R., Avni, R., Distelfeld, A., McGuire, P. E., Dvorak, J., & Luo, M.-C. (2019). Improved genome sequence of wild emmer wheat *avitan* with the aid of optical maps. *G3 (Bethesda)*, *9*, 619–624. <https://doi.org/10.1534/g3.118.200902>

## SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Papon, N., Lasserre-Zuber, P., Rimbart, H., De Oliveira, R., Paux, E., & Choulet, F. (2023). All families of transposable elements were active in the recent wheat genome evolution and polyploidy had no impact on their activity. *The Plant Genome*, e20347. <https://doi.org/10.1002/tpg2.20347>