

# Impact of transposable elements on genome structure and evolution in bread wheat

International Wheat Genome Sequencing Consortium (IWGSC)

DOI:

[10.1186/s13059-018-1479-0](https://doi.org/10.1186/s13059-018-1479-0)

License:

Creative Commons: Attribution (CC BY)

*Document Version*

Publisher's PDF, also known as Version of record

*Citation for published version (Harvard):*

International Wheat Genome Sequencing Consortium (IWGSC) 2018, 'Impact of transposable elements on genome structure and evolution in bread wheat', *Genome Biology*, vol. 19, no. 1, pp. 103.

<https://doi.org/10.1186/s13059-018-1479-0>

[Link to publication on Research at Birmingham portal](#)

## **Publisher Rights Statement:**

Checked for eligibility: 26/09/2018

First published in:

Genome Biology 2018;19:103

<https://doi.org/10.1186/s13059-018-1479-0>

## **General rights**

Unless a licence is specified above, all rights (including copyright and moral rights) in this document are retained by the authors and/or the copyright holders. The express permission of the copyright holder must be obtained for any use of this material other than for purposes permitted by law.

- Users may freely distribute the URL that is used to identify this publication.
- Users may download and/or print one copy of the publication from the University of Birmingham research portal for the purpose of private study or non-commercial research.
- User may use extracts from the document in line with the concept of 'fair dealing' under the Copyright, Designs and Patents Act 1988 (?)
- Users may not further distribute the material nor use it for the purposes of commercial gain.

Where a licence is displayed above, please note the terms and conditions of the licence govern your use of this document.

When citing, please reference the published version.

## **Take down policy**

While the University of Birmingham exercises care and attention in making items available there are rare occasions when an item has been uploaded in error or has been deemed to be commercially or otherwise sensitive.


If you believe that this is the case for this document, please contact [UBIRA@lists.bham.ac.uk](mailto:UBIRA@lists.bham.ac.uk) providing details and we will remove access to the work immediately and investigate.

RESEARCH

Open Access



# Impact of transposable elements on genome structure and evolution in bread wheat

Thomas Wicker<sup>1†</sup>, Heidrun Gundlach<sup>2†</sup>, Manuel Spannagl<sup>2</sup>, Cristobal Uauy<sup>3</sup>, Philippa Borrill<sup>3</sup>, Ricardo H. Ramírez-González<sup>3</sup>, Romain De Oliveira<sup>4</sup>, International Wheat Genome Sequencing Consortium<sup>5</sup>, Klaus F. X. Mayer<sup>2,6</sup>, Etienne Paux<sup>4</sup> and Frédéric Choulet<sup>4\*</sup> 

## Abstract

**Background:** Transposable elements (TEs) are major components of large plant genomes and main drivers of genome evolution. The most recent assembly of hexaploid bread wheat recovered the highly repetitive TE space in an almost complete chromosomal context and enabled a detailed view into the dynamics of TEs in the A, B, and D subgenomes.

**Results:** The overall TE content is very similar between the A, B, and D subgenomes, although we find no evidence for bursts of TE amplification after the polyploidization events. Despite the near-complete turnover of TEs since the subgenome lineages diverged from a common ancestor, 76% of TE families are still present in similar proportions in each subgenome. Moreover, spacing between syntenic genes is also conserved, even though syntenic TEs have been replaced by new insertions over time, suggesting that distances between genes, but not sequences, are under evolutionary constraints. The TE composition of the immediate gene vicinity differs from the core intergenic regions. We find the same TE families to be enriched or depleted near genes in all three subgenomes. Evaluations at the subfamily level of timed long terminal repeat-retrotransposon insertions highlight the independent evolution of the diploid A, B, and D lineages before polyploidization and cases of concerted proliferation in the AB tetraploid.

**Conclusions:** Even though the intergenic space is changed by the TE turnover, an unexpected preservation is observed between the A, B, and D subgenomes for features like TE family proportions, gene spacing, and TE enrichment near genes.

**Keywords:** Transposable elements, Wheat genome, Genome evolution, LTR retrotransposons, Polyploidy, *Triticum aestivum*

## Background

Transposable elements (TEs) are ubiquitous components of genomes and one of the major forces driving genome evolution [1]. They are classified into two classes: retrotransposons (class 1), transposing via reverse transcription of their messenger RNA (mRNA), and DNA transposons (class 2), representing all other types of elements [2]. TEs are small genetic units with the ability to make copies of themselves or move around in the genome. They do not encode a function that would allow them to be maintained by selection across generations; rather, their strategy relies

on their autonomous or non-autonomous amplification. TEs are subject to rapid turnover, are the main contributors of intraspecific genomic diversity, and are the main factor explaining genome size variations. Thus, TEs represent the dynamic reservoir of the genomes. They are epigenetically silenced [3], preventing them from long-term massive amplification that could be detrimental. The dynamics of TEs in genomes remains unclear, and it was supposed that they may escape silencing and experience bursts of amplification followed by rapid silencing. Their impact on gene expression has also been documented in many species (for a review, see [4]). In addition, they play a role at the structural level, as essential components of centromeric chromatin in plants [3, 5]. Plant genomes are generally dominated by a small number of highly repeated

\* Correspondence: [frederic.choulet@inra.fr](mailto:frederic.choulet@inra.fr)

<sup>†</sup>Thomas Wicker and Heidrun Gundlach contributed equally to this work.

<sup>4</sup>GDEC, INRA, UCA (Université Clermont Auvergne), Clermont-Ferrand, France  
Full list of author information is available at the end of the article



families, especially class I Gypsy and Copia long terminal repeat retrotransposons (LTR-RTs) [6–10]. Most of our knowledge about TE dynamics and their impact on gene expression in complex plant genomes comes from maize [10–14]. At the whole genome level, Makarevitch et al. have shown that four to nine maize TE families, including all major class I superfamilies (Gypsy, Copia, long interspersed nuclear elements (LINEs)), and DNA transposons, are enriched (more than twofold) in promoters of genes being up-regulated in response to different abiotic stresses [15]. This study also suggested that TEs are a major source of allelic variations explaining differential response to stress between accessions.

The genome of bread wheat (*Triticum aestivum* L.), one of the most important crop species, has also undergone massive TE amplification with more than 85% of it being derived from such repeat elements. It is an allohexaploid comprising three subgenomes (termed A, B, and D) that have diverged from a common ancestor around 2–3 million years ago (Mya) (according to molecular dating of chloroplast DNA [16]) and hybridized within the last half million years. This led to the formation of a complex, redundant, and allohexaploid genome. These characteristics make the wheat genome by far the largest and most complex genome that has been sequenced and assembled into near-complete chromosomes so far. They, however, also make wheat a unique system in which to study the impact of TE activity on genome structure, function, and organization.

Previously only one reference sequence quality wheat chromosome was available, which we annotated using our automated TE annotation pipeline (CLARITE) [17, 18]. However, it was unknown whether the TE content of chromosome 3B was typical of all wheat chromosomes and how TE content varied between the A, B, and D subgenomes. Therefore, in this study, we address the contribution of TEs to wheat genome evolution on a chromosome-wide scale. We report on the comparison of the three A-B-D subgenomes in terms of TE content and proliferation dynamics. We show that, although rounds of TE insertions/deletions have completely modified the TE space since A-B-D diverged, the proportion of each TE family remained stable between subgenomes. In addition, the specific TE landscape in the direct vicinity of genes is very similar between the three subgenomes. Our results strongly suggest that TEs play a role at the structural level likely under selection pressure. We also identified TE families that are over-represented in promoters compared to the rest of the genome but did not reveal a strong association between particular TE families and nearby gene expression pattern or a strong stress-response association.

## Results and discussion

### TE content and distribution along the 21 bread wheat chromosomes

Building from a decade-long effort from the wheat genomics community, we used the accumulated knowledge about TEs to precisely delineate the TE repertoire of the 21 chromosomes based on a similarity search with a high-quality TE databank: ClariTeRep [17] which includes TREP [19]. This represents 3050 manually annotated and curated TEs carried by the three subgenomes and mainly identified on bacterial artificial chromosome (BAC) sequences obtained during map-based cloning or survey sequencing projects, especially on chromosome 3B [20]. CLARITE was used to model TEs in the sequence and their nested insertions when possible [17]. This led to the identification of 3,968,974 TE copies, belonging to 505 families, and representing 85% of RefSeq\_v1.0. Overall, the TE proportion is very similar in the A, B, and D subgenomes, as they represented 86%, 85%, and 83% of the sequence, respectively. However, the sizes of the subgenomes differ: with 5.18 Gb, the B subgenome has the largest assembly size, followed by the A subgenome (4.93 Gb) and the smaller D subgenome (3.95 Gb). The repetitive fraction is mostly dominated by TEs of the class I Gypsy and Copia and class II CACTA superfamilies; other superfamilies contribute very little to overall genome size (Table 1, Fig. 1a).

At the superfamily level, the A, B, and D subgenomes have similar TE compositions (Fig. 1a). The smaller size of the D subgenome (~ 1 Gb smaller than A and B) is mainly due to a smaller amount of Gypsy (~ 800 Mb less; Fig. 1a). The A and B subgenomes differ in size by only 245 Mb (~ 5%), and nearly half of this (106 Mb) is not due to known TEs but rather to low copy sequences. Since the amount of coding DNA is very conserved (43, 46, and 44 Mb, respectively), this difference is mainly due to parts of the genome that remained un-annotated so far. This un-annotated portion of the genome may contain degenerated and unknown weakly repeated elements.

Similar to other complex genomes, only six highly abundant TE families represent more than half of the TE content: RLC\_famc1 (Angela), DTC\_famc2 (Jorge), RLG\_famc2 (Sabrina), RLG\_famc1 (Fatima), RLG\_famc7 (Sumana/Sumaya), and RLG\_famc5 (WHAM), while 486 families out of 505 (96%) each account for less than 1% of the TE fraction. In terms of copy number, 50% (253) of the families are repeated in fewer than 1000 copies at the whole genome level, while more than 100,000 copies were detected for each of the seven most repeated families (up to 420,639 Jorge copies).

Local variations of the TE density were observed following a pattern common to all chromosomes: the TE proportion is lower (on average 73%) in the distal

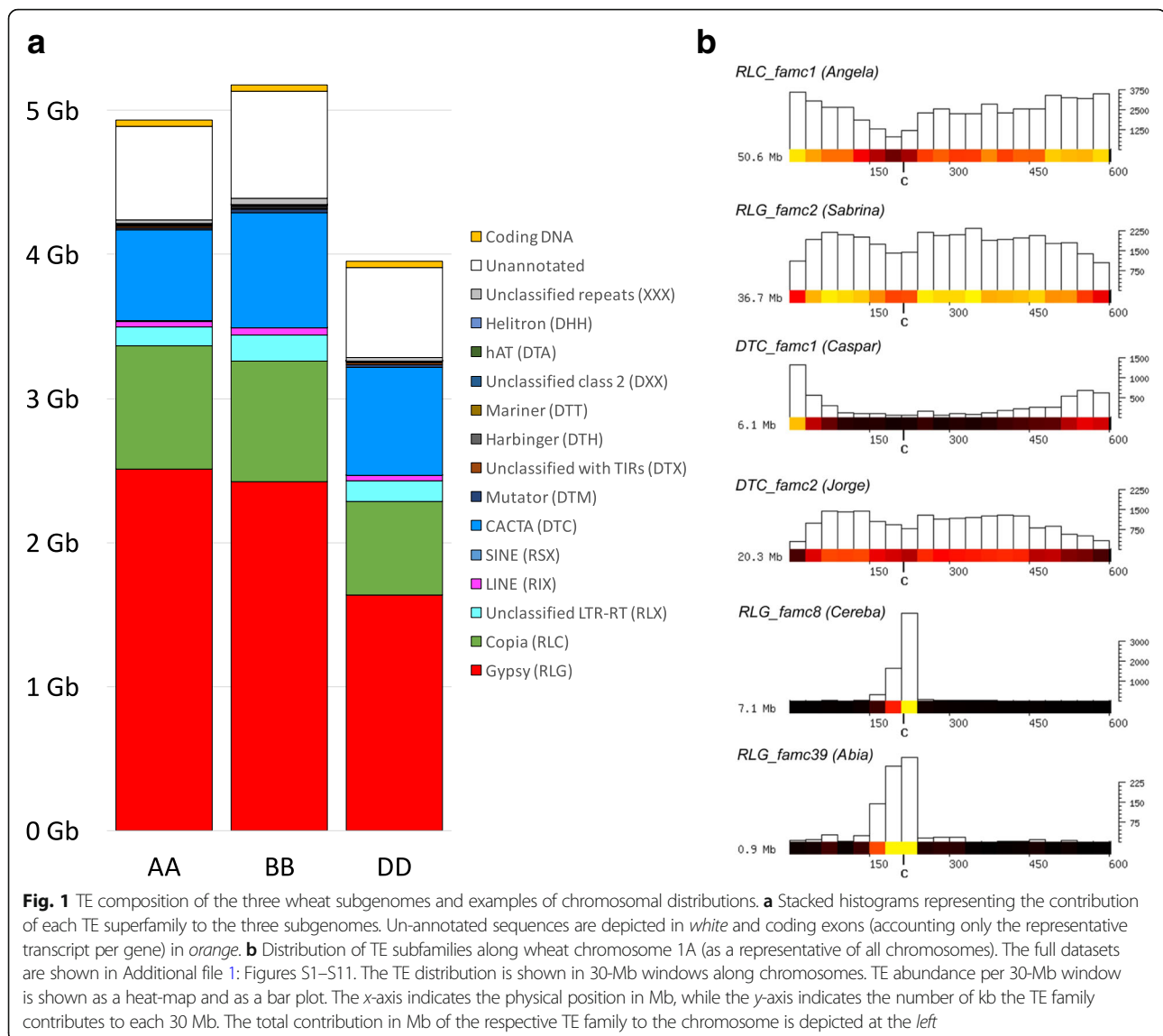
**Table 1** Proportion of TE superfamilies in the A, B, and D subgenomes and at the whole genome level. Proportions are expressed as the percentage of sequences assigned to each superfamily relatively to the genome size. *TIR* terminal inverted repeat

	A	B	D	Complete genome
Class 1 LTR retrotransposons				
Gypsy (RLG)	50.9%	46.8%	41.4%	46.7%
Copia (RLC)	17.5%	16.2%	16.3%	16.7%
Unclassified LTR-RT (RLX)	2.6%	3.5%	3.7%	3.2%
Non-LTR retrotransposons				
LINE (RIX)	0.82%	0.96%	0.93%	0.90%
SINE (RSX)	0.01%	0.01%	0.01%	0.01%
Class 2 DNA transposons				
CACTA (DTC)	12.8%	15.5%	19.0%	15.5%
Mutator (DTM)	0.30%	0.38%	0.48%	0.38%
Unclassified with TIRs (DTX)	0.21%	0.20%	0.22%	0.21%
Harbinger (DTH)	0.15%	0.16%	0.18%	0.16%
Mariner (DTT)	0.14%	0.16%	0.17%	0.16%
Unclassified class 2 (DXX)	0.05%	0.08%	0.05%	0.06%
hAT (DTA)	0.01%	0.01%	0.01%	0.01%
Helitrons (DHH)	0.00%	0.00%	0.00%	0.00%
Unclassified repeats (XXX)	0.55%	0.85%	0.63%	0.68%
Genes and non TE-related DNA	13.9%	15.3%	16.8%	15.2%

regions than in the proximal and interstitial regions (on average 89%). However, much stronger local variations were observed when distributions of individual TE families were studied. Figure 1b shows TE distributions using chromosome 1A as a representative example. Distributions for selected TE families on all chromosomes are shown in Additional file 1: Figures S1–S11. The most abundant TE family, RLC\_famc1 (Angela) was enriched towards telomeres and depleted in proximal regions. In contrast, highly abundant Gypsy retrotransposons RLG\_famc2 (Sabrina, Fig. 1b) and RLG\_famc5 (WHAM, not shown) were enriched in central parts of chromosome arms and less abundant in distal regions. CACTA TEs also showed a variety of distribution patterns. They can be grouped into distinct clades depending on their distribution pattern, as suggested earlier based on chromosome 3B TE analyses [17]. Families of the Caspar clade [21] are highly enriched in telomeric regions, as is shown for the example of the DTC\_famc1 (Caspar) whereas DTC\_famc2 (Jorge) showed the opposite pattern (Fig. 1b).

Centromeres have a specific TE content. Previous studies on barley and wheat reported that the Gypsy family RLG\_famc8.3 (Cereba) is enriched in centromeres [22, 23]. It was speculated that Cereba integrase can target centromere-specific heterochromatin due to the presence of a chromodomain that binds specifically to centromeric histones [24]. We found that wheat Cereba elements are

concentrated in centromeric regions but absent from the rest of the genome (Fig. 1b, Additional file 1: Figure S8), as are their closely related subfamilies RLG\_famc8.1 and RLG\_famc8.2 (Quinta). We identified new TE families that are also highly enriched in centromeres. The family RLG\_famc39 (Abia) is a relative of Cereba, although there is very little sequence DNA conservation between the two. However, at the protein level, Cereba is its closest homolog. Abia and Cereba have an extremely similar distribution (Fig. 1b, Additional file 1: Figures S8 and S9). Interestingly, on chromosome 6A Cereba is more abundant, while on 3B, Abia is more abundant, suggesting that the two TE families are competing for the centromeric niche. Abia seems to be a wheat-specific TE family, as it was not present in the recently published barley genome [25]. A recent study on the barley genome reported on a novel centromeric Gypsy family called Abiba [21]. We identified a homolog in wheat: RLG\_famc40 (Abiba), with two distinct subfamilies RLG\_famc40.1 and RLG\_famc40.2, corresponding to the putatively autonomous and non-autonomous variants. Abiba is enriched in central parts of chromosomes but with a broader spreading compared to Abia and Cereba (Additional file 1: Figures S10 and S11). At a higher resolution, we identified large tandem arrays of Cereba and Abia elements that correspond to the high *k*-mer frequencies observed at the centromeres (Fig. 2d), which might be the signature of functional centromeres (Additional file 1: Figure S12).



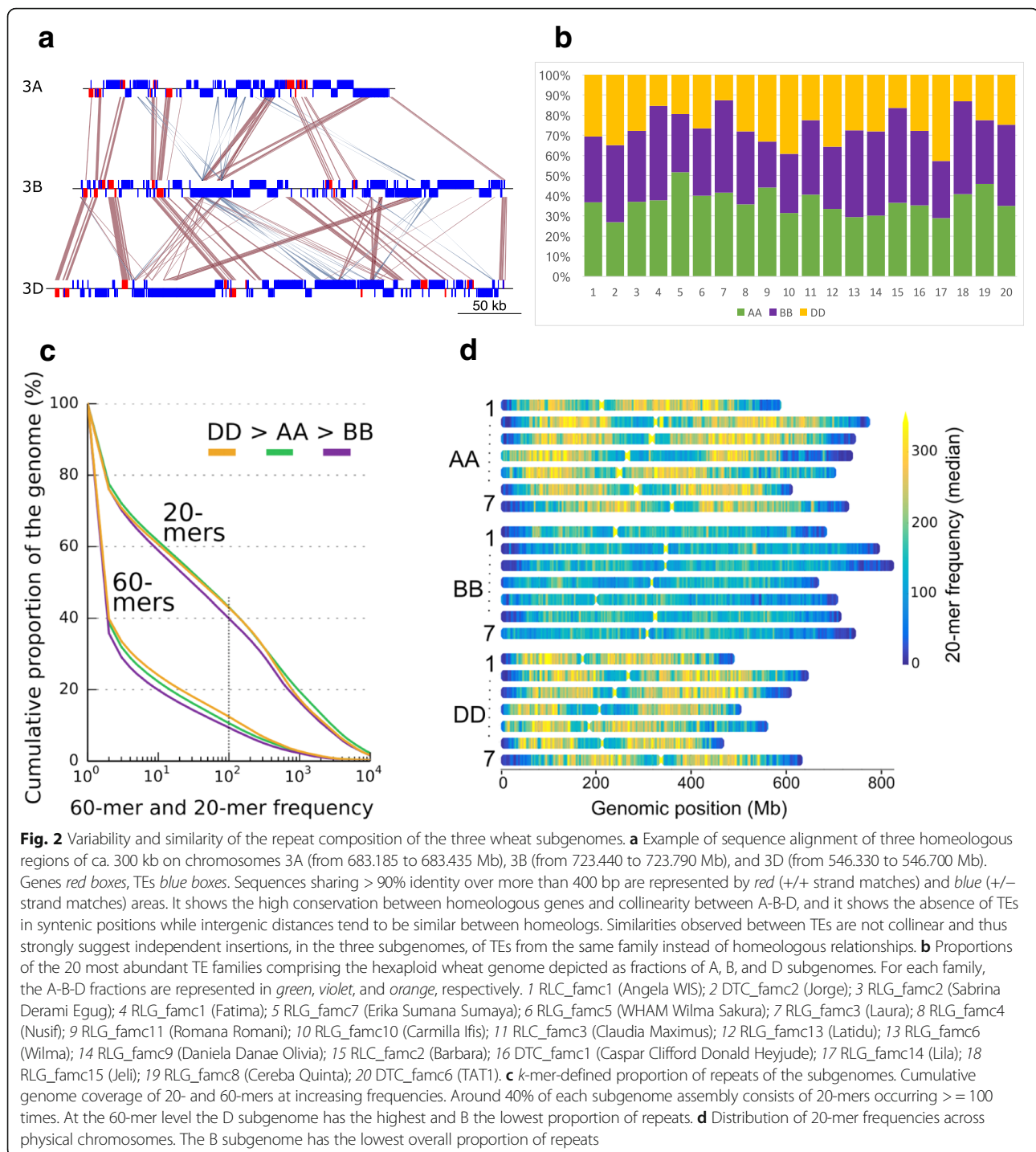
### Similarity and variability of the TE content between the A, B, and D subgenomes

A genome-wide comparative analysis of the 107,891 high-confidence genes predicted along the A, B, and D subgenomes (35,345, 35,643, and 34,212, respectively) was described in detail in [26]. It revealed that 74% of the genes are homeologs, with the vast majority being syntenic. Thus, gene-based comparisons of A-B-D highlighted a strong conservation and collinearity of the genes between the three genomes. However, outside the genes and their immediate surrounding regions, we found almost no sequence conservation in the TE portions of the intergenic regions (Fig. 2a). This is due to the “TE turnover” [27], which means that intergenic sequences (i.e., sequences that are not under selection pressure) evolve through rounds of TE insertions and deletions in a continuing process: DNA is produced by TE insertions into intergenic regions and

removed by unequal crossovers or deletions that occur during double-strand repair [28]. Previous studies showed that this process occurs at a pace implying that intergenic sequences are completely turned over within a few million years [27, 28]. Consequently, we found practically no conserved TEs (i.e., TEs that were inserted in the common ancestor of the A, B, and D genome donors). Thus, although the repetitive fraction in A, B, and D genomes is mostly composed of the same TE families (see below), their individual insertion sites and nesting patterns are completely different.

Analysis of the *k*-mer content of RefSeq\_v1.0 showed that 20-mers occurring 100× or more cover around 40% of the wheat genome sequence (Fig. 2c). For 60-mers, this value decreases to only 10%. This pattern was strongly similar between subgenomes, although a slight difference was observed: repeated *k*-mers covered a larger proportion





of the subgenome  $D > A > B$ . This lower proportion of repeats in the B subgenome is also obvious using a heat-map of 20-mer frequencies (Fig. 2d), showing that the B genome contains a smaller proportion of high copy number perfect repeats.

We then compared the A, B, and D subgenomes at the TE family level. We did not find any TE families (accounting > 10 kb) that are specific for a single subgenome or

completely absent in one subgenome (only two cases of subgenome-specific tandem repeats were found: XXX\_famc46/c47). More surprisingly, the abundance of most TE families is similar in the A, B, and D subgenomes. Indeed, among the 165 families which represent at least 1 Mb of DNA each, 125 (76%) are present in similar proportions in the three subgenomes; i.e., we found less than a twofold change of the proportion between subgenomes.

Figure 2b represents the proportions of the 20 most abundant families in the three subgenomes which account for 84% of the whole TE fraction. Their proportion is close to the relative sizes of the three subgenomes: 35%, 37%, 28% for A, B, D, respectively. This highlighted the fact that not only are the three subgenomes shaped by the same TE families, but also that these families are present in proportions that are conserved. Consistent with this, we identified only 11 TE families (7%) that show a strong difference (i.e., more than a threefold change in abundance) between two subgenomes, representing only 2% of the overall TE fraction.

Thus, despite the near-complete TE turnover that has occurred independently in the A-B-D diploid lineages (Fig. 2a), and although TEs have transposed and proliferated very little since polyploidization (0.5 Mya, see below), the TE families that currently shape the three subgenomes are the same, and more strikingly, their abundance remained very similar. We conclude that almost all families ancestrally present in the A-B-D common ancestor have been active at some point and their amplification has compensated their loss by deletion, thus suggesting a dynamic in which families are maintained at equilibrium in the genome for millions of years. This evolutionary scenario differs from the model where TEs evolve by massive bursts of a few families leading to rapid diversification [29]. For example, Piegu et al. showed that an amplification burst of a single retrotransposon family led to a near doubling of the genome size in *Oryza australiensis* [30]. In wheat, by contrast, many TE families contribute to the genome diversification, as suggested for plants with very large genomes (> 30 Gb) [31].

Strong differences in abundance between the A, B, and D genomes were observed at the subfamily level (Fig. 3). For example, the highly abundant RLC\_famc1 (Fatima) family has diverged into at least five subfamilies (1.1 to 1.5). Only RLC\_famc1.1 contains potentially functional reverse transcriptase (RT) and integrase (INT) genes, while RLC\_famc1.4 and RLC\_famc1.5 contain gag and protease open reading frames (ORFs). RLC\_famc1.2 and RLC\_famc1.3 appear to be non-autonomous, as they do not contain any intact ORFs. We suggest that RLC\_famc1.1 provides functional RT and INT proteins, while protease and GAG are provided by other subfamilies. Their contrasted abundance revealed that RLC\_famc1.4 and RLC\_famc1.5 proliferated specifically in the B and A lineages, respectively (Fig. 3a).

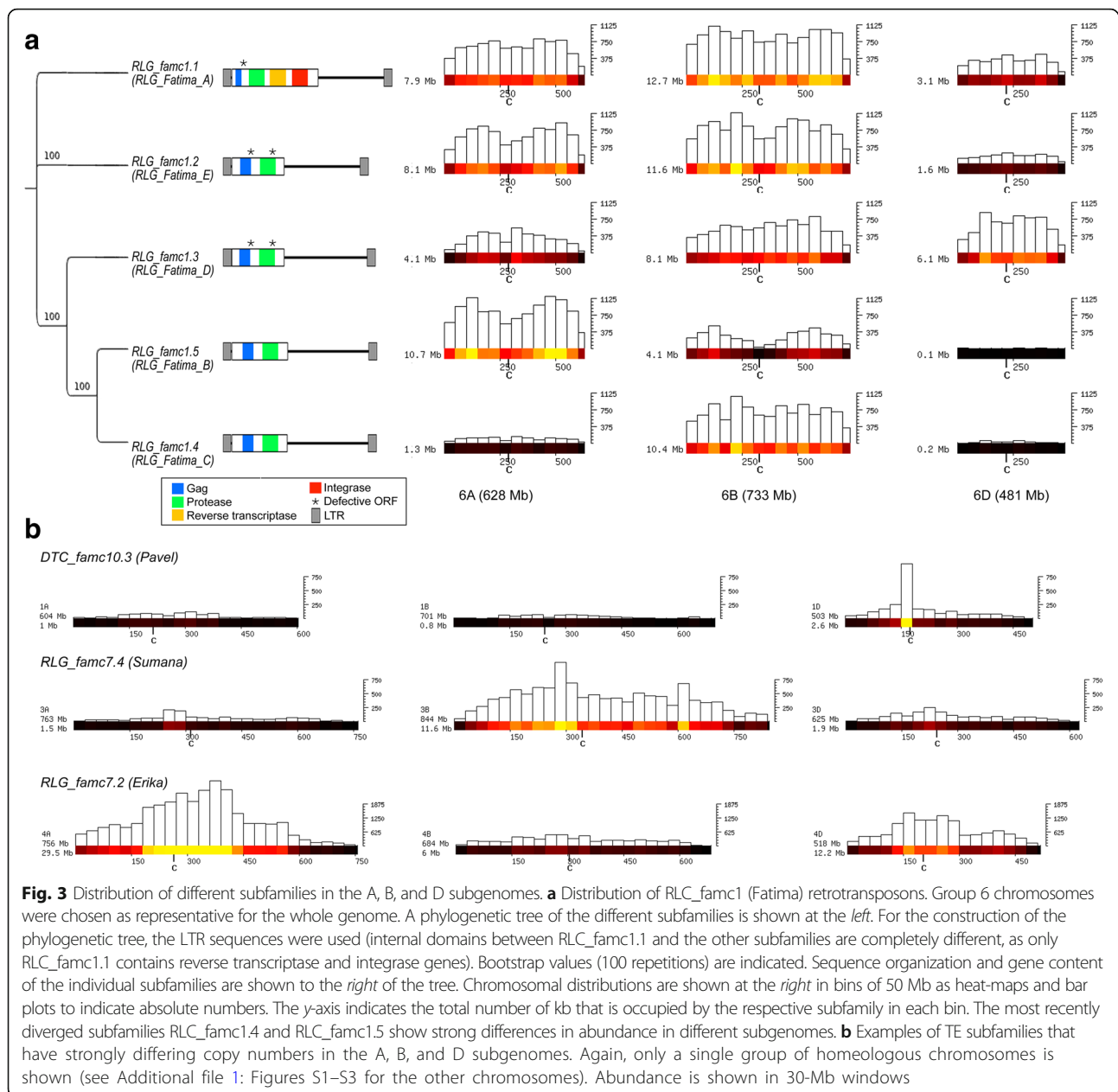
In total, we identified 18 different subfamilies (belonging to 11 different families) which show subgenome-specific over- or under-representation (Table 2). Here, we only considered TE families that contribute more than 0.1% to the total genome and are at least threefold over- or under-represented in one of the subgenomes. This illustrated that these 11 highly abundant families did not show a bias between A-B-D at the family level, but are

composed of several subfamilies that were differentially amplified in the three diploid lineages. The CACTA family DTC\_famc10.3 (Pavel) is much more abundant in the D subgenome than in the A and B subgenomes (Additional file 1: Figure S1). Interestingly, the Pavel subfamily also seems to have evolved a preference for inserting close to centromeres in the D subgenome, while this tendency is not obvious in the A and B subgenomes (Fig. 3b). Generally, subfamilies were enriched in a single genome (Table 2). In only four cases, a subfamily was depleted in one subgenome while abundant at similar levels in the other two. Three of these cases were found in the D subgenome. This is consistent with the smaller D subgenome size, and differences in highly abundant elements contribute to this difference.

#### Dynamics of LTR retrotransposons from the diploid ancestors to the hexaploid

The largest portion of plant genomes with size over 1 Gb consists of LTR-RTs. Intact full-length elements represent recently inserted copies, whereas old elements have experienced truncations, nested insertions, and mutations that finally lead to degenerated sequences until they become unrecognizable. Full-length LTR-RTs (fLTR-RTs) are bordered by two LTRs that are identical at the time of insertion and subsequently diverge by random mutations, a characteristic that is used to determine the age of transposition events [13]. In previous genome assemblies, terminal repeats tended to collapse, which resulted in very low numbers of correctly reconstructed fLTR-RTs (triangles in Additional file 1: Figure S13). We found 112,744 fLTR-RTs in RefSeq\_v1.0 (Additional file 1: Table S1, Figure S13), which was in line with the expectations and confirmed the linear relationship between fLTR-RTs and genome size within the Poaceae. This is two times higher than the number of fLTR-RTs assembled in TGAC\_v1 [32], while almost no fLTR-RTs were assembled in the 2014 gene-centric draft assembly [33].

We exploited this unique dataset to gain insights into the evolutionary history of hexaploid wheat from a transposon perspective. fLTR-RTs are evenly distributed among the subgenomes, with on average 8 elements per Mb (Additional file 1: Table S1). Among them, there were two times more Copia (RLC) than Gypsy (RLG) elements, although Gypsy elements account for 2.8× more DNA. This means that the proportion of young intact elements is higher for the Copia superfamily than for the Gypsy superfamily. Indeed, the median insertion ages for Copia, Gypsy, and RLX (unclassified LTR-RTs) are 0.95, 1.30, and 1.66 million years (Myr). RLXs lack a protein domain, preventing a straightforward classification into Gypsy or Copia. The missing domains can most likely be accounted for by their older age and, thus, their higher degree of



degeneration. RLX elements are probably unable to transpose on their own, but the occurrence of such very recently transposed elements suggests that they are non-autonomous, as described for the Fatima subfamilies (Fig. 3a). Between the A and B subgenomes, all fILTR-RT metrics are very similar, whereas the D subgenome stands out with younger insertions. In any case, age distributions of fILTR-RTs show that most of the identified full-length elements inserted after the divergence of the three subgenomes, thereby reflecting the genomic turnover that has removed practically all TEs that were present in the A-B-D ancestor (see above).

We analyzed the chromosomal distributions of the fILTR-RTs (Additional file 1: Figure S14). The whole set of elements is relatively evenly scattered along the chromosomes with high density spots in the distal gene-rich compartments. The most recent transpositions (i.e., copies with two identical LTRs) involved 457 elements: 257 Copia, 144 Gypsy, and 56 RLXs. They are homogeneously distributed along the chromosomes (Additional file 1: Figure S14B), confirming previous hypotheses stating that TEs insert at the same rate all along the chromosome but are deleted faster in the terminal regions, leading to gene-rich and TE-depleted chromosome extremities [17].



**Table 2** TE subfamilies that show differences in abundance between subgenomes

CLARITE name	TREP name <sup>a</sup>	A/B <sup>b</sup>	A/D <sup>c</sup>	B/D <sup>d</sup>	Comment	Genome <sup>e</sup>
DTC_famc14	DTC_Pavel	1.9	0.4	0.2	Enriched	D
DTC_famc8	DTC_TAT4	0.8	0.3	0.4	Enriched	D
RLC_famc1.7	RLC_Angela_A	3.9	1.0	0.2	Depleted	B
RLC_famc1.8	RLC_Angela_B	0.2	0.6	2.9	Enriched	B
RLC_famc3.2	RLC_Claudia	1.9	4.1	2.2	Enriched	A
RLG_famc1.2	RLG_Fatima_E	0.8	3.8	4.7	Enriched	A
RLG_famc1.4	RLG_Fatima_C	0.2	3.9	20.9	Enriched	B
RLG_famc1.5	RLG_Fatima_B	2.5	76.6	30.1	Enriched	A
RLG_famc10.2	RLG_Carmilla	1.5	9.6	6.6	Enriched	B
RLG_famc10.3	RLG_Ifis	0.9	0.2	0.3	Enriched	D
RLG_famc15	RLG_Jeli	1.0	3.3	3.4	Depleted	D
RLG_famc3.1	RLG_Laura_B	1.0	4.8	5.0	Depleted	D
RLG_famc3.4	RLG_Laura_A	1.0	5.3	5.3	Depleted	D
RLG_famc7.2	RLG_Erika	4.9	2.0	0.4	Enriched	A
RLG_famc7.3	RLG_Sumaya	2.9	4.1	1.4	Enriched	A
RLG_famc7.4	RLG_Sumana	0.2	0.9	4.1	Enriched	B
RLG_famc9.2	RLG_Daniela	0.6	7.1	11.0	Enriched	B
RLX_famc4	Unnamed	0.5	0.2	0.5	Enriched	D

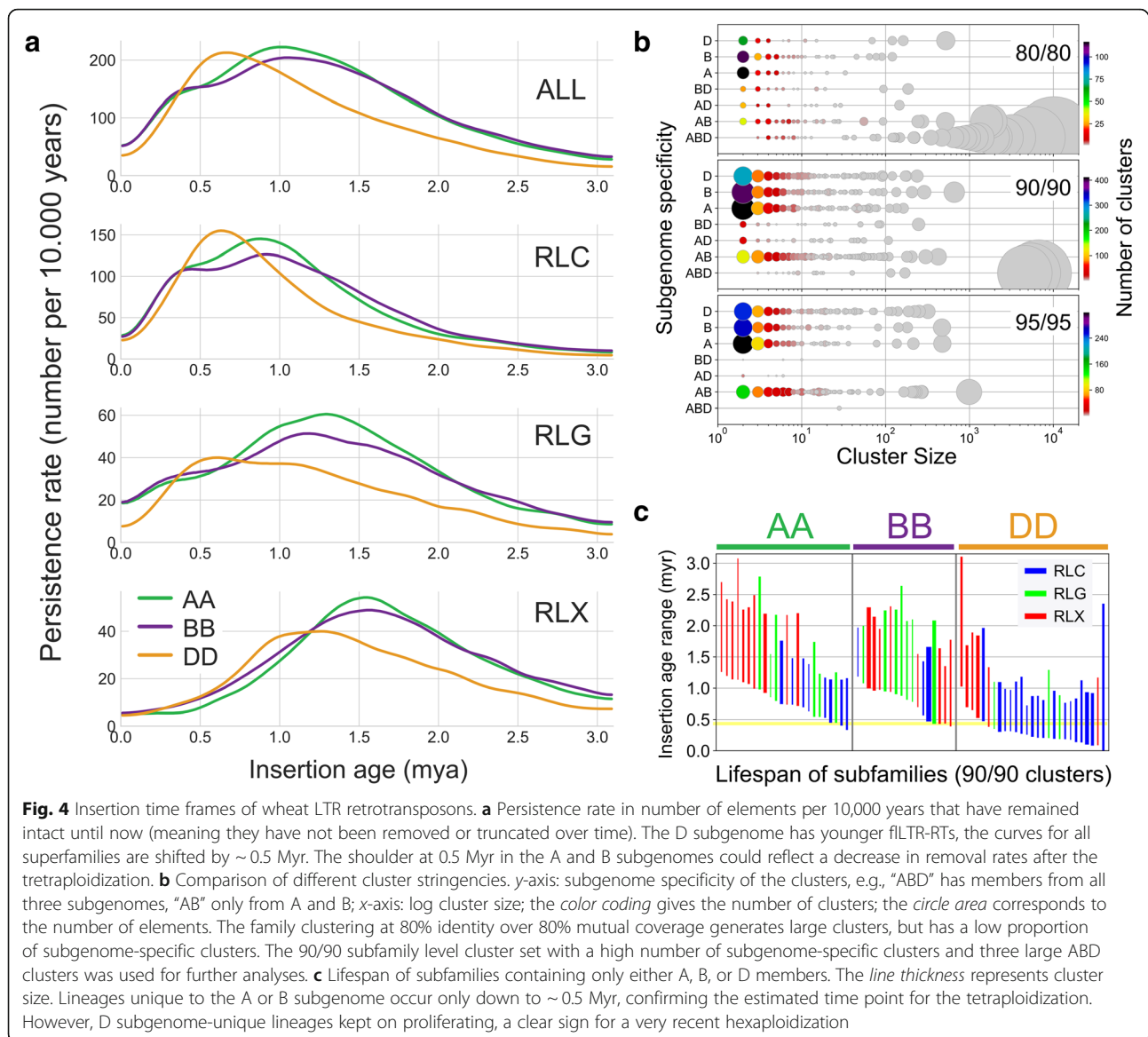
<sup>a</sup>Name under which the family was previously described and/or stored in the TREP database<sup>bcd</sup>Ratio of abundance between subgenomes<sup>e</sup>Subgenome in which abundance or distribution of the respective TE families differs

The current fLTR-RT content is the outcome of two opposing forces: insertion and removal. Therefore, we calculated a persistence rate, giving the number of elements per 10,000 years that have remained intact over time, for the 112,744 fLTR-RTs (Fig. 4a). It revealed broad peaks for each superfamily, with maxima ranging from 0.6 Mya (for Copia in the D subgenome) to 1.5 Mya (for RLX in the A and B subgenomes). The D subgenome contained on average younger fLTR-RTs compared to A and B, with a shift of activity by 0.5 Myr. Such peaks of age distributions are commonly interpreted in the literature as transposon amplification bursts. We find the “burst” analogy misleading, because the actual values are very low. For wheat, it represents a maximal rate of only 600 copies per 10,000 years. A more suiting analogy would be the formation of mountain ranges, where small net increases over very long time periods add up to very large systems. In the most recent time (< 10,000 years), after the hexaploidization event, we did not see any evidence in our data for the popular “genomic shock” hypothesis, postulating immediate drastic increases of transposon insertions [34–36]. For the A and B subgenomes, a shoulder in the persistence curves around 0.5 Mya (Fig. 4a), the time point of tetraploidization, was observed. We suggest that counter-selection of harmful TE insertions was relaxed in the tetraploid genome; i.e., the polyploid could

tolerate insertions which otherwise would have been removed by selection in a diploid.

To elucidate the TE amplification patterns that have occurred before and after polyploidization, we clustered the 112,744 fLTR-RTs based on their sequence identity. The family level was previously defined at 80% identity over 80% sequence coverage (80/80 clusters) [2]. We also clustered the fLTR-RTs using a more stringent cutoff of 90/90 and 95/95 to enable classification at the subfamily level (Fig. 4b). The 80/80 clusters were large and contained members of all three subgenomes. In contrast, the 90/90 and 95/95 clusters were smaller, and a higher proportion of them are specific to one subgenome. To trace the polyploidization events, we defined lifespans for each individual LTR-RT subfamily as the interval between the oldest and youngest insertion (Fig. 4c). Subfamilies specific to either the A or B subgenome amplified until about 0.4 Myr, which is consistent with the estimated time of the tetraploidization. Some of the D subgenome-specific subfamilies inserted more recently, again consistent with the very recent hexaploidization.

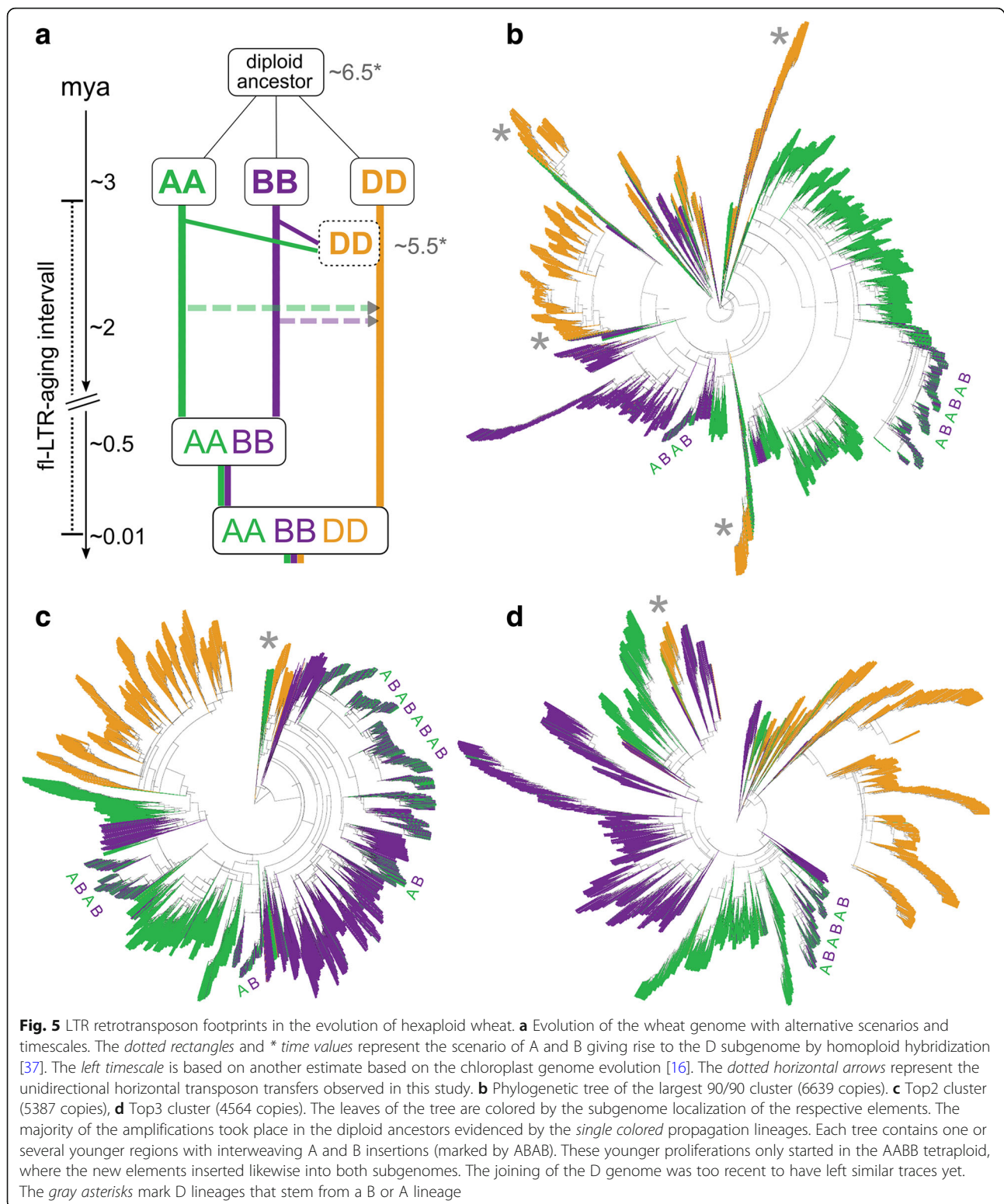
These results confirmed that the three subgenomes were shaped by common families present in the A-B-D common ancestor that have amplified independently in the diploid lineages. They evolved to give birth to different subfamilies that, generally, did not massively amplify



after polyploidization and, thus, are specific to one subgenome. To confirm this hypothesis, we explored the phylogenetic trees of the three largest 90/90 clusters color-coded by subgenome (Fig. 5 and Additional file 1: Figures S15–S17 for more details). The trees show older subgenome-specific TE lineages which have proliferated in the diploid ancestors (2–0.5 Mya). However, the youngest elements (<0.5 Mya) were found in clades interweaving elements of the A and B subgenomes, corresponding to amplifications in the tetraploid. Such cases involving the D subgenome were not observed, showing that flLTR-RTs from D have not yet transposed in large amounts across the subgenomes since the birth of hexaploid wheat 8000–10,000 years ago. We further noticed several incidences in the trees where D lineages were derived from older B or A lineages, but not the

reverse. This may be explained by the origin of the D subgenome through homoploid hybridization between A and B [37].

There are two proposed models of propagation of TEs: the “master copy” model and the “transposon” model [38]. The “master copy” model gives rise to highly unbalanced trees (i.e., with long successive row patterns) where one active copy is serially replaced by another, whereas the “transposon” model produces balanced trees where all branches duplicate with the same rate [39]. To better discern the tree topologies, we plotted trees with equal branch length and revealed that the three largest trees (comprising 15% of flLTR-RTs) are highly unbalanced (Additional file 1: Figure S18), while the smaller trees are either balanced or unbalanced (Additional file 1: Figure S19). Taken together, both types of tree topologies



exist in the proliferation of flLTR-RTs, but there is a bias towards unbalanced trees for younger elements, suggesting that TE proliferation followed the “master copy” model.

In summary, our findings give a timed TE atlas depicting detailed TE proliferation patterns of hexaploid wheat. They also show that polyploidization did not trigger bursts of TE activity. This dataset of well-defined transposon

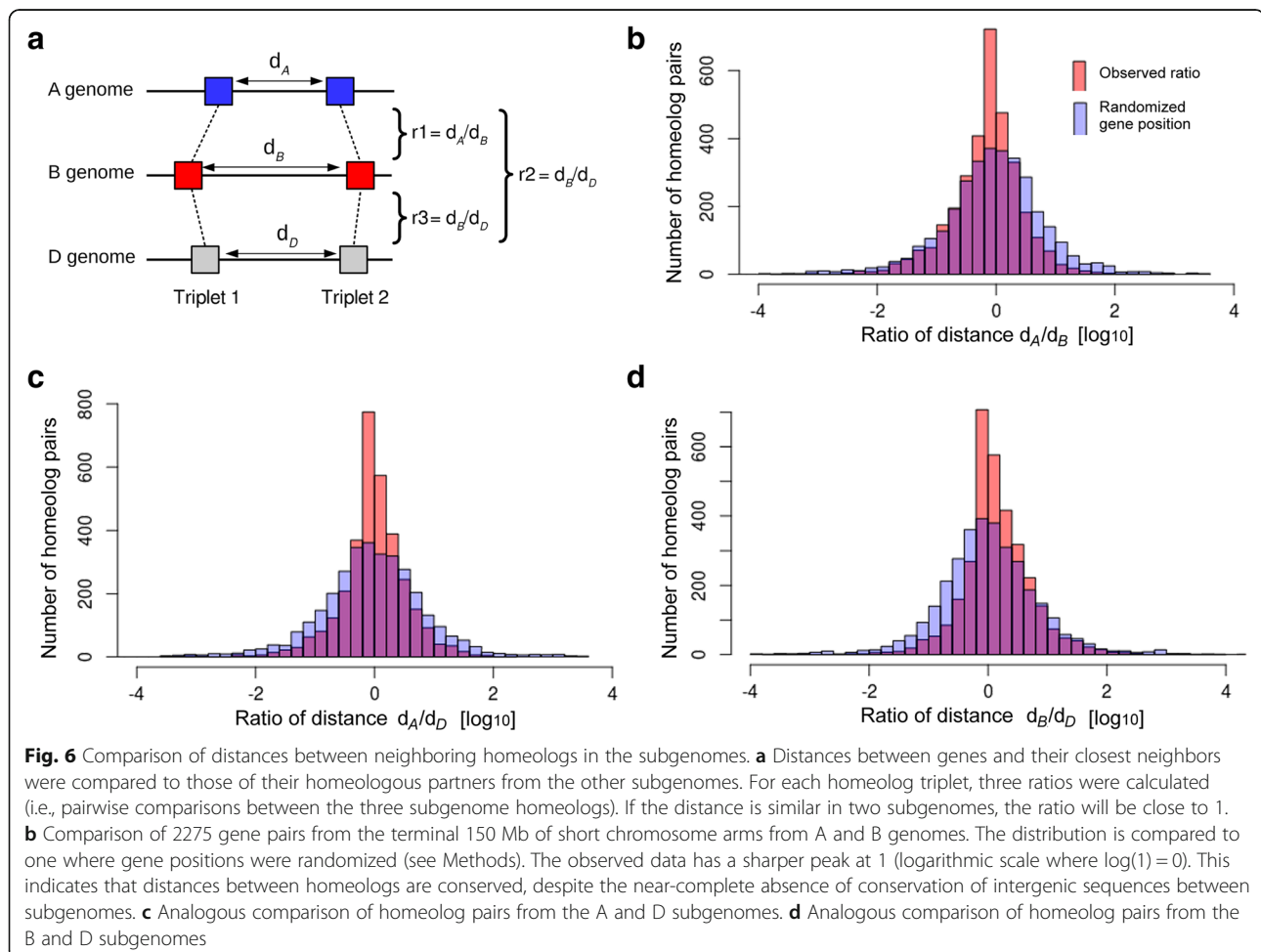
lineages now provides the basis to further explore the factors controlling transposon dynamics. Founder elements may help us obtain better insights into common patterns which could explain how and why amplification starts.

### A stable genome structure despite the near-complete TE turnover in the intergenic sequences

As described above, intergenic sequences show almost no conservation between homeologous loci. That means they contain practically no TEs that have inserted already in the common ancestor of the subgenomes. Instead, ancestral sequences were removed over time and replaced by TEs that have inserted more recently. Despite this near-complete turnover of the TE space (Fig. 2a), the gene order along the homeologous chromosomes is well conserved between the subgenomes and is even conserved with the related grass genomes (sharing a common ancestor 60 Mya [40]). Most interestingly and strikingly, not only gene order but also distances between neighboring homeologs tend to be conserved between subgenomes (Fig. 6). Indeed, we found that the ratio of distances between neighboring

homeologs has a strong peak at 1 (or 0 in log scale on Fig. 6), meaning that distances separating genes tend to be conserved between the three subgenomes despite the TE turnover. This effect is non-random, as ratio distribution curves are significantly flatter ( $p = 1.10^{-5}$ ) when gene positions along chromosomes are randomized. These findings suggest that distances between genes are likely under selection pressure.

We found this constrained distribution irrespective of the chromosome compartments, i.e., distal, interstitial, and proximal, exhibiting contrasted features at the structural (gene density) and functional (recombination rate, gene expression breadth) levels [25, 26]. However, constraints applied on intergenic distances seem relaxed (broader peak in Fig. 6) in proximal regions where the meiotic recombination rate is extremely low. At this point, we can only speculate about the possible impact of meiotic recombination as a driving force towards maintaining a stable chromosome organization. Previous studies have shown that recombination in highly repetitive genomes occurs mainly in or near genes [41]. We hypothesize that spacing of genes is



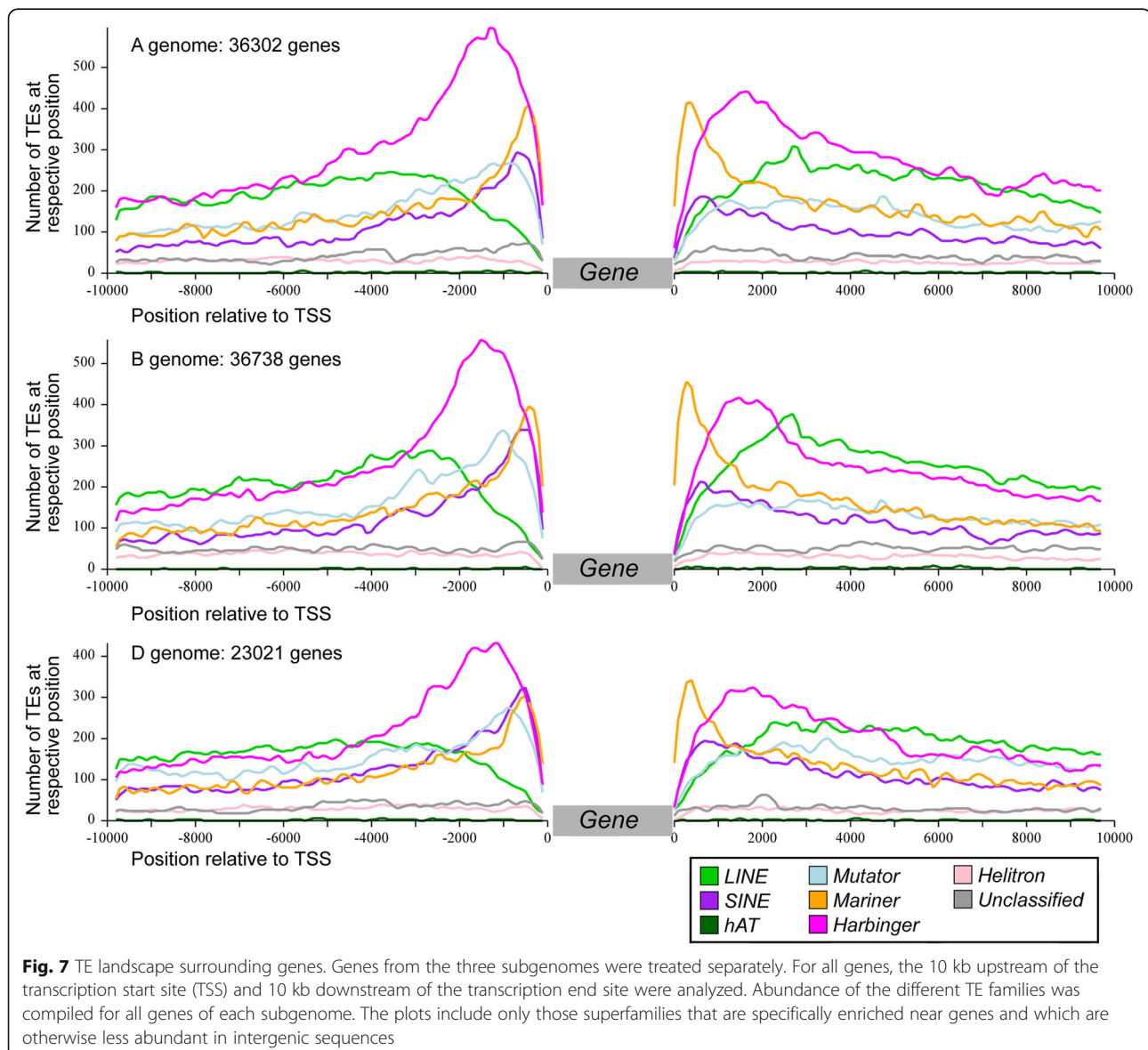


preserved for proper expression regulation or proper pairing during meiosis. Previous studies on introgressions of divergent haplotypes in large-genome grasses support this hypothesis. For instance, highly divergent haplotypes which still preserve the spacing of genes have been maintained in wheats of different ploidy levels at the wheat *Lr10* locus [42].

#### Enrichment of TE families in gene promoters is conserved between the A, B, and D subgenomes

The sequences flanking genes have a very distinct TE composition compared to the overall TE space. Indeed, while intergenic regions are dominated by large TEs such as LTR-RTs and CACTAs, sequences surrounding genes are enriched in small TEs that are usually just a few

hundred base pairs in size (Fig. 7). Immediately upstream and downstream of genes (within 2 kb), we identified mostly small non-autonomous DNA transposons of the Harbinger and Mariner superfamilies, referred to as Tourist and Stowaway miniature inverted-repeat transposable elements (MITEs), respectively [43], SINEs, and Mutators (Fig. 7). At the superfamily level, the A, B, and D subgenomes exhibit the same biased composition in gene surrounding regions (Additional file 1: Figure S20). We then computed, independently for each subgenome, the enrichment ratio of each TE family that was present in the promoter of protein-coding genes (2 kb upstream of the transcription start site (TSS)) compared to their overall proportion (in copy number, considering the 315 TE families with at least 500 copies). The majority (242, 77%)





showed a bias (i.e., at least a twofold difference in abundance) in gene promoters compared to their subgenome average, confirming that the direct physical environment of genes contrasts with the rest of the intergenic space. Considering a strong bias, i.e., at least a threefold over- or under-representation in promoters, we found 105 (33%) and 38 (12%) families, respectively, that met this threshold in at least one subgenome. While it was previously known that MITEs were enriched in promoters of genes, here we show that this bias is not restricted to MITEs but rather involves many other families. Again, although TEs that shaped the direct gene environment have inserted independently in the A, B, and D diploid lineages, their evolution converged to three subgenomes showing very similar TE composition. To go further, we showed that the tendency of TE families to be enriched in, or excluded from, promoters was extremely conserved between the A, B, and D subgenomes (Fig. 8), although TEs are not conserved between homeologous promoters (inserted after A-B-D divergence), except for a few cases of retained TEs (see below). In other words, when a family is over- or under-represented in the promoter regions of one subgenome, it is also true for the two other subgenomes. We did not find any family that was enriched in a gene promoter in one subgenome while under-represented in gene promoters of another subgenome.

Superfamily is generally but not always a good indicator of the enrichment of TEs in genic regions (Fig. 8). For instance, 83% (25/30) of the LINE families are over-represented in the promoter regions, while none of them is under-represented (considering a twofold change). We confirmed that class 2 DNA transposons (especially MITEs) are enriched in promoters, while Gypsy retrotransposons tend to be excluded from the close vicinity of genes. Indeed, among the 105 families strongly enriched in promoters (threefold change), 53% (56) are from class 2 and 21% (22) are LINES, and only 5% (5) are LTR-RTs. Contrary to Gypsy, Mutator, Mariner, and Harbinger, families belonging to CACTA and Copia superfamilies do not share a common enrichment pattern: some TE families can be either over- or under-represented in promoters (Fig. 8). This confirmed previous results about CACTAs annotated along the 3B chromosome [17], revealing that a part of the CACTA families is associated with genes while the other follows the distribution of Gypsy. Our results showed that this is also true for Copia.

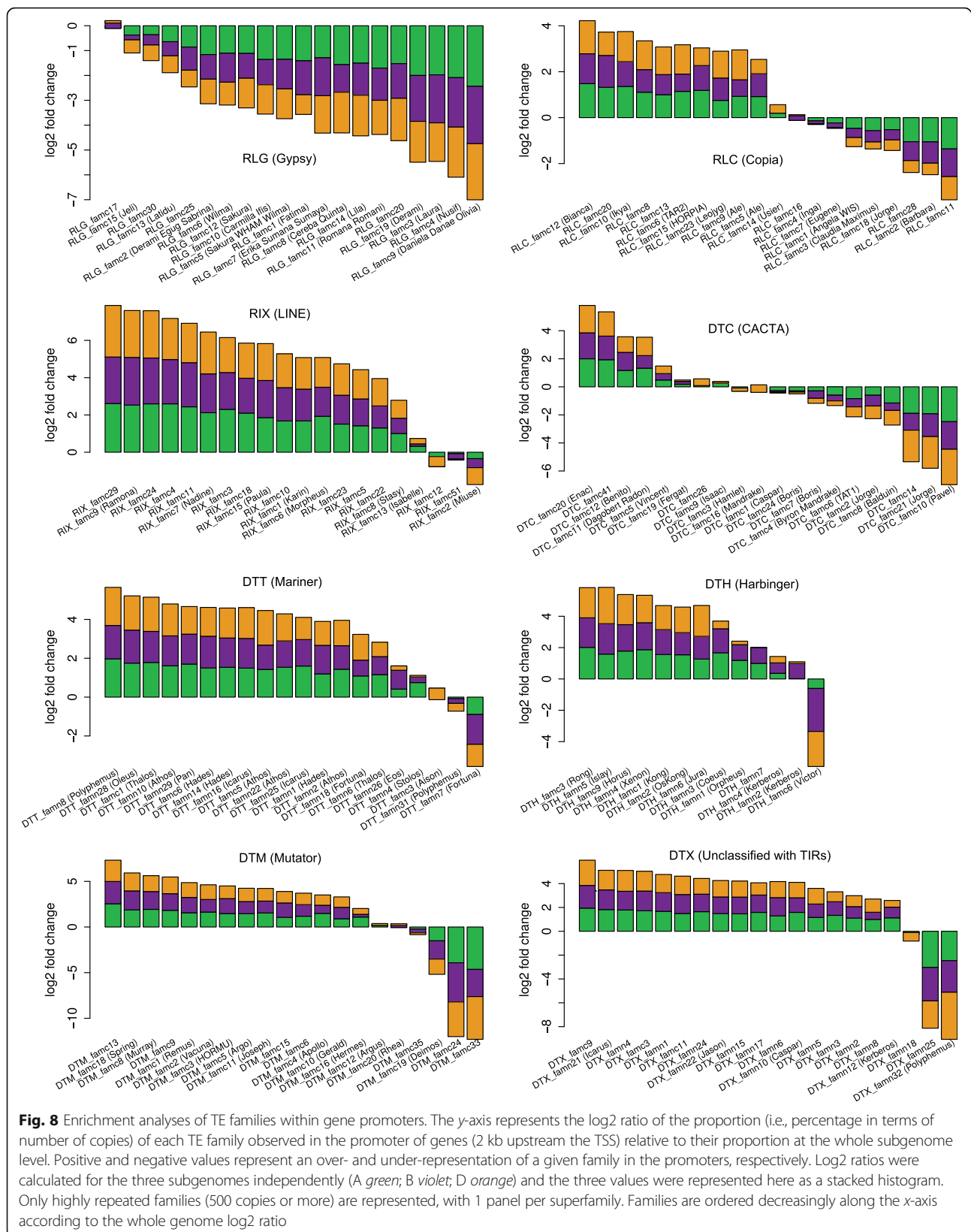
Thus, the TE turnover did not changed the highly organized genome structure. Given that not only proportions, but also enrichment patterns, remained similar for almost all TE families after A-B-D divergence, we suggest that TEs tend to be at the equilibrium in the genome, with amplification compensating their deletion (as described in [29]), and with families enriched around genes having remained the same.

### No strong association between gene expression and particular TE families in promoters

We investigated the influence of neighboring TEs on gene expression. Indeed, TEs are so abundant in the wheat genome, that genes are almost systematically flanked by a TE in the direct vicinity. The median distance between the gene TSS and the closest upstream TE is 1.52 kb, and the median distance between the transcription termination site (TTS) and the closest downstream TE is 1.55 kb, while the average gene length (between TSS and TTS) is 3.44 kb. The density as well as the diversity of TEs in the vicinity of genes allow us to speculate on potential relationships between TEs and gene expression regulation. We used the gene expression network built by [26] based on an exhaustive set of wheat RNA-seq data. Genes were clustered into 39 expression modules sharing a common expression profile across all samples. We also grouped unexpressed genes to study the potential influence of TEs on neighbor gene silencing. For each gene, the closest TE upstream was retrieved, and we investigated potential correlations through an enrichment analysis (each module was compared to the full gene set). Despite the close association between genes and TEs, no strong enrichment for a specific family was observed for any module or for the unexpressed genes.

We then studied the TE landscape upstream of wheat homeolog triplets, focusing on 19,393 triplets (58,179 genes) with a 1:1:1 orthologous relationship between A, B, and D subgenomes. For each triplet, we retrieved the closest TE flanking the TSS and investigated the level of conservation of flanking TEs between homeologs. For 75% of the triplets, the three flanking TEs belong to three different families, revealing that, even in the close vicinity of genes, TEs are in majority not conserved between homeologs due to rapid turnover. This suggests that most TEs present upstream of triplets were not selected for by the presence of common regulatory elements across homeologs. However, for 736 triplets (4%), the three homeologs are flanked by the same element, constituting a conserved noncoding sequence (CNS), suggesting that part of this element is involved in the regulation of gene expression. These TE-derived CNSs are on average 459 bp, which is three times smaller than the average size of gene-flanking TE fragments (on average 1355 bp), suggesting that only a portion of the ancestrally inserted TEs are under selection pressure. They represent a wide range (149 different families) of diverse elements belonging to all the different superfamilies.

The majority of homeolog triplets have relatively similar expression patterns [26, 44], contrary to what was found for older polyploid species like maize [45]. In synthetic polyploid wheat, it was shown that repression of D subgenome homeologs was related to silencing of neighbor TEs



[46]. Thus, we focused on triplets for which two copies are coexpressed while the third is silenced. However, enrichment analysis did not reveal any significant enrichment of specific TE families in promoters of the silenced homeologs. We also examined transcriptionally dynamic triplets across tissues [44]. Again, no TE enrichment in promoters was observed. These results suggest that recent changes in gene expression are not due to specific families recently inserted in the close vicinity of genes.

## Conclusions

The chromosome-scale assembly of the wheat genome provided an unprecedented genome-wide view of the organization and impact of TEs in such a complex genome. Since they diverged, the A, B, and D subgenomes have experienced a near-complete TE turnover, although polyploidization did not massively reactivate TEs. This turnover contrasted drastically with the high level of gene synteny. Apart from genes, there was no conservation of the TE space between homeologous loci. But surprisingly, TE families that have shaped the A, B, and D subgenomes are the same, and unexpectedly, their proportions and intrinsic properties (gene-prone or not) are quite similar despite their independent evolution in the diploid lineages. Thus, TE families are somehow at equilibrium in the genome since the A-B-D common ancestor. These novel insights contradict the previous model of evolution with amplification bursts followed by rapid silencing. Our results suggest a role of TEs at the structural level. TEs are not just “junk DNA”; our findings open new perspectives to elucidate their role in high-order chromatin arrangement, chromosome territories, and gene regulation.

## Methods

### TE modeling using CLARITE

The *Triticum aestivum* cv. Chinese Spring genome sequence was annotated as described in [26]. Briefly, two gene prediction pipelines were used (TriAnnot: developed at GDEC Institute [INRA-UCA Clermont-Ferrand] and the pipeline developed at Helmholtz Center Munich [PGSB]), and the two annotations were integrated (pipeline established at Earlham Institute [47]) to achieve a single high-quality gene set. TE modeling was achieved through a similarity search approach based on the ClariTeRep curated databank of repeated elements [48], developed specifically for the wheat genome, and with the CLARITE program that was developed to model TEs and reconstruct their nested structure [17]. ClariTeRep contains sequences present in TREP, i.e., a curated library of *Triticeae* TEs from all three subgenomes (originating from BACs sequenced during map-based cloning or survey sequencing projects) and TEs manually annotated in a previous pilot study of chromosome 3B [20]. For the annotation, we used the ClariTeRep naming system,

which assigns simple numbers to individual families and subfamilies; e.g., RLG\_famc1.1 and RLG\_famc1.2 are subfamilies of RLG\_famc1. Since many TE families have been previously named, we provided this previous name in parentheses.

### Detection and characterization of full-length LTR retrotransposons

Identification of fLTR-RTs was based on LTRharvest [49]. For RefSeq\_v1.0, LTRharvest reported 501,358 non-overlapping fLTR-RT candidates under the following parameter settings: “overlaps best -seed 30 -minlenltr 100 -maxlenltr 2000 -mindistltr 3000 -maxdistltr 25000 -similar 85 -mintsd 4 -maxtsd 20 -motif tgca -motifmis 1 -vic 60 -xdrop 5 -mat 2 -mis -2 -ins -3 -del -3”. All candidates were annotated for PfamA domains with hmmer3 [50] and stringently filtered for canonical elements by the following criteria: (1) presence of at least one typical retrotransposon domain (RT, RH, INT, GAG); (2) removal of mis-predictions based on inconsistent domains, e.g., RT-RH-INT-RT-RH; (3) Absence of gene-related Pfam domains; (4) strand consistency between domains and primer binding site; (5) tandem repeat content below 25%; (6) long terminal repeat size  $\leq 25\%$  of the element size; (7) N content  $< 5\%$ . This resulted in a final set of 112,744 high-quality fLTR-RTs. The Copia and Gypsy superfamilies were defined by their internal domain ordering: INT-RT-RH for RLC and RH-RT-INT for RLG [2]. When this was not possible, the prediction was classified as RLX. The 112,744 fLTR-RTs were clustered with vmatch dbcluster [51] at three different stringencies: 95/95 (95% identity over 95% mutual length coverage), 90/90, and 80/80, as follows: vmatch “-dbcluster 95 95 -identity 95 -exdrop 3 -seedlength 20 -d”, “-dbcluster 90 90 -identity 90 -exdrop 4 -seedlength 20 -d” and “-dbcluster 80 80 -identity 80 -exdrop 5 -seedlength 15 -d”. Subgenome specificity of clusters was defined by the following decision tree: (1) assignment of the respective subgenome if  $\geq 90\%$  of the members were located on this subgenome; (2) assignment to two subgenomes if members from one subgenome  $< 10\%$ , e.g., AB-specific if D members  $< 10\%$ ; (3) Assignment of the remaining clusters as ABD common. Muscle was used for multiple alignments of each cluster [52] in a fast mode (-maxiters 2 -diags1). To build phylogenetic trees, we used tree2 from the muscle output which was created in the second iteration with a Kimura distance matrix, and trees were visualized with ete3 toolkit [53]. The date of fLTR-RT insertions was based on the divergence between the 5' and 3' LTRs calculated with emboss distmat, applying the Kimura 2-parameter correction. The age was estimated using the formula: age = distance/(2\*mutation rate) with a mutation rate of  $1.3 \times 10^{-8}$  [13]. The lifespan of an

individual LTR-RT subfamily was defined as the 5th to 95th percentile interval between the oldest and youngest insertions. The densities for the chromosomal heat-maps were calculated using a sliding window of 4 Mb with a step of 0.8 Mb.

### Comparative analysis of distances separating neighbor genes between homeologous chromosomes

For the comparison of distances separating neighbor genes, homeologous triplets located in the three chromosomal compartments (distal, interstitial, and proximal; Additional file 1: Table S2) were treated separately. This was done because gene density is lower in interstitial and proximal regions, and because the latter show a lack of genetic recombination. Furthermore, we considered only triplets where all three homeologous genes are found on the homeologous chromosomes. Comparison of homeologous gene pairs from distal regions was done in two ways, both of which yielded virtually identical results. Distances were measured from one gene to the one that follows downstream. However, there were many small local inversions between the different subgenomes. Thus, if a gene on the B or D subgenome was oriented in the opposite direction compared to its homeologous copy in the A subgenome, it was assumed that that gene is part of a local inversion. Therefore, the distance to the preceding gene on the chromosome was calculated. The second approach was more stringent, based only on triplets for which all three homeologs are in the same orientation in the three subgenomes. The results obtained from the two approaches were extremely similar, and we presented only the results from the second, more stringent, approach. For the control dataset, we picked a number of random positions along the chromosomes that is equal to the number of homeologs for that chromosome group. Then, homeologous gene identifiers were assigned to these positions from top to bottom (to preserve the order of genes but randomize the distances between them). This was done once for all three chromosomal compartments. Histograms of the distributions of the distance ratios between homeologs were produced with *rstudio* ([rstudio.com](http://rstudio.com)). The significance of the differences between the largest group of actual and randomized gene positions (peak of the histogram) was established with a chi-square test.

### Analyses of TEs in the vicinity of genes and enrichment analyses

We developed a Perl script (*gffGetClosestTe.pl* [54]) to retrieve gene-flanking TEs from the feature coordinates in the GFF file. It was used to extract the closest TE on each side of every predicted gene (considering “gene” features that include untranslated regions). It was also used to extract all predicted TE copies entirely or partially present within 2 kb

upstream of the “gene” start position, i.e., the TSS. Enrichment analyses were then automated using R scripts.

### Enrichment of TE families in gene promoters (2 kb upstream)

Independently for the three subgenomes, we retrieved all TE copies present within 2 kb upstream of the TSSs of all gene models and calculated the percentage of the number of copies assigned to every family ( $\%famX^{promoter}$ ). We also calculated the percentage of the number of copies of each family at the whole subgenome level ( $\%famX^{whole\_subgenome}$ ). One enrichment log2 ratio was calculated for each A, B, and D subgenome using the formula  $\log_2(\%famX^{promoter} / \%famX^{whole\_subgenome})$ . Only families accounting for 500 copies or more in the whole genome were considered.

### TE families and expression modules

Here, we retrieved the closest TE present in 5′ of the TSS for all genes and calculated the percentage of each TE family for each expression module and the unexpressed genes (considered as a module), and compared them to the percentage observed for the whole gene set using the formula  $\log_2(\%famX^{genes\_module} / \%famX^{all\_genes})$ . The log2 ratio was calculated only for expression modules representing at least 1000 coexpressed genes, and we considered only log2 ratio values for families accounting for 500 copies or more. A similar approach was taken for the 10% stable, 80% middle, and 10% dynamic genes as defined by [44].

### Comparison of TE families in the promoter of homeologs

Here, we also retrieved the closest TE in 5′ of every gene and identified homeologous triplets for which the closest element in 5′ belongs to the same family for the three copies. For that, we developed a Perl script (*getTeHomeologs.pl* [54]) in order to integrate the information of homeologous genes and the data of the closest TE in 5′ of genes. Only “1–1–1” homeologs were considered.

### Additional files

**Additional file 1:** Tables S1 to S2, Figures S1 to S20. (PDF 2879 kb)

**Additional file 2:** Reviewer reports and Author's response to reviewers. (DOCX 30 kb)

### Abbreviations

CNS: Conserved non-coding sequence; fLTR-RT: Full-length long terminal repeat retrotransposon; INT: Integrase; LINE: Long interspersed nuclear element; LTR: Long terminal repeat; MITE: Miniature inverted-repeat transposable element; ORF: Open reading frame; RH: Ribonuclease H; RT: Retrotransposon; SINE: Short interspersed nuclear element; TE: Transposable element; TSS: Transcription start site; TTS: Transcription termination site

### Funding

The research leading to these results has received funding from the French Government managed by the Research National Agency (ANR) under the Investment for the Future programme (BreedWheat project ANR-10-BTBR-



03), from FranceAgriMer (2011–0971 and 2013–0544). This work was also supported by the UK Biotechnology and Biological Sciences Research Council (BBSRC) through grants (BB/P016855/1, BB/P013511/1, and BB/M014045/1). TW was funded by the University of Zurich. KFXM, HG, and MS would like to acknowledge funding from the German Federal Ministry of Food and Agriculture (2819103915) and the German Ministry of Education and Research (de.NBI 031A536).

#### Availability of data and materials

In terms of data availability, IWGSC RefSeq\_v1.0 assembly [55], gene and TE annotation [56], as well as related data are available on the IWGSC Data Repository hosted at URGI [57] and the source code of perl scripts developed for this study [54]. The raw sequencing data is in the Sequence Read Archive under accession number SRP114784 [58].

#### Review history

The reviewer reports with the author's response to the reviewers are available as Additional file 2.

#### Authors' contributions

TW, HG, and FC conceived and designed the study. FC coordinated the study. TW performed analyses of TE distribution along chromosomes, comparisons of intergenic distances between subgenomes, centromeric TE analyses, and also analyzed the TE landscape around genes. HG carried out the *k*-mer analyses and the fLTR-RT identification, dating, clustering, and phylogenetic analyses. FC carried out the A-B-D comparisons of TE families, enrichment analyses in gene promoters, comparisons between homeologs, and association with expression profiles. CU, RHRG, and PB built the RNA-seq-based gene expression network. TW, HG, and FC wrote the manuscript. RDO contributed text and figures for the manuscript. MS, CU, RHRG, PB, KFXM, RDO, and EP provided comments and corrections on the manuscript. All authors read and approved the final manuscript.

#### Ethics approval and consent to participate

Not applicable.

#### Consent for publication

Not applicable.

#### Competing interests

The authors declare that they have no competing interests.

#### Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

#### Author details

<sup>1</sup>Department of Plant and Microbial Biology, University of Zurich, Zurich, Switzerland. <sup>2</sup>PGSB Plant Genome and Systems Biology, Helmholtz Center Munich, German Research Center for Environmental Health, Neuherberg, Germany. <sup>3</sup>Department of Crop Genetics, John Innes Centre, Norwich Research Park, Colney, Norwich NR4 7UH, UK. <sup>4</sup>GDEC, INRA, UCA (Université Clermont Auvergne), Clermont-Ferrand, France. <sup>5</sup>IWGSC, 2841 NE Marywood Ct, Lee's Summit, MO 64086, USA. <sup>6</sup>School of Life Sciences, Technical University Munich, Munich, Germany.

Received: 25 January 2018 Accepted: 11 July 2018

Published online: 17 August 2018

#### References

- Feschotte C, Jiang N, Wessler SR. Plant transposable elements: where genetics meets genomics. *Nat Rev Genet*. 2002;3:329–41.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, Flavell A, Leroy P, Morgante M, Panaud O, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8:973–82.
- Hollister JD, Gaut BS. Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. *Genome Res*. 2009;19:1419–28.
- Feschotte C. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet*. 2008;9:397–405.
- Talbert PB, Masuelli R, Tyagi AP, Comai L, Henikoff S. Centromeric localization and adaptive evolution of an Arabidopsis histone H3 variant. *Plant Cell*. 2002;14:1053–66.
- International Barley Genome Sequencing Consortium, Mayer KF, Waugh R, Brown JW, Schulman A, Langridge P, Platzer M, Fincher GB, Muehlbauer GJ, Sato K, et al. A physical, genetic and functional sequence assembly of the barley genome. *Nature*. 2012;491:711–6.
- International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature*. 2005;436:793–800.
- Mayer KF, Martis M, Hedley PE, Simkova H, Liu H, Morris JA, Steuernagel B, Taudien S, Roessner S, Gundlach H, et al. Unlocking the barley genome by chromosomal and comparative genomics. *Plant Cell*. 2011;23:1249–63.
- Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al. The *Sorghum bicolor* genome and the diversification of grasses. *Nature*. 2009;457:551–6.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112–5.
- Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, Westerman RP, Sanmiguel PJ, Bennetzen JL. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet*. 2009;5:e1000732.
- Fu H, Dooner HK. Intraspecific violation of genetic colinearity and its implications in maize. *Proc Natl Acad Sci U S A*. 2002;99:9573–8.
- SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. *Nat Genet*. 1998;20:43–5.
- Swigonova Z, Bennetzen JL, Messing J. Structure and evolution of the *r/b* chromosomal regions in rice, maize and sorghum. *Genetics*. 2005;169:891–906.
- Makarevitch I, Waters AJ, West PT, Stitzer M, Hirsch CN, Ross-Ibarra J, Springer NM. Transposable elements contribute to activation of maize genes in response to abiotic stress. *PLoS Genet*. 2015;11:e1004915.
- Middleton CP, Senerchia N, Stein N, Akhunov ED, Keller B, Wicker T, Kilian B. Sequencing of chloroplast genomes from wheat, barley, rye and their relatives provides a detailed insight into the evolution of the Triticeae tribe. *PLoS One*. 2014;9:e85761.
- Daron J, Glover N, Pingault L, Theil S, Jamilloux V, Paux E, Barbe V, Mangenot S, Alberti A, Wincker P, et al. Organization and evolution of transposable elements along the bread wheat chromosome 3B. *Genome Biol*. 2014;15:546.
- Choulet F, Alberti A, Theil S, Glover N, Barbe V, Daron J, Pingault L, Sourdille P, Couloux A, Paux E, et al. Structural and functional partitioning of bread wheat chromosome 3B. *Science*. 2014;345:1249721.
- University of Zurich. TREP. 2016. <http://botserv2.uzh.ch/kelldata/trep-db/>.
- Choulet F, Wicker T, Rustenhiz C, Paux E, Salse J, Leroy P, Schlub S, Le Paslier MC, Magdelenat G, Gonthier C, et al. Megabase level sequencing reveals contrasted organization and evolution patterns of the wheat gene and transposable element spaces. *Plant Cell*. 2010;22:1686–701.
- Wicker T, Schulman A, Tanskanen J, Spannagl M, Twardziok S, Mascher M, Springer NM, Li Q, Waugh R, Li C, et al. The repetitive landscape of the 5,100 Mbp barley genome. *Mob DNA*. 2018; in press.
- Li B, Choulet F, Heng Y, Hao W, Paux E, Liu Z, Yue W, Jin W, Feuillet C, Zhang X. Wheat centromeric retrotransposons: the new ones take a major role in centromeric structure. *Plant J*. 2013;73:952–65.
- Presting GG, Malysheva L, Fuchs J, Schubert I. A Ty3/gypsy retrotransposon-like sequence localizes to the centromeric regions of cereal chromosomes. *Plant J*. 1998;16:721–8.
- Chatterjee AG, Leem YE, Kelly FD, Levin HL. The chromodomain of Tf1 integrase promotes binding to cDNA and mediates target site selection. *J Virol*. 2009;83:2675–85.
- Mascher M, Gundlach H, Himmelbach A, Beier S, Twardziok SO, Wicker T, Radchuk V, Dockter C, Hedley PE, Russell J, et al. A chromosome conformation capture ordered sequence of the barley genome. *Nature*. 2017;544:427–33.
- International Wheat Genome Sequencing Consortium. Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science*. 2018. <https://doi.org/10.1126/science.aar7191>.
- Vitte C, Panaud O. LTR retrotransposons and flowering plant genome size: emergence of the increase/decrease model. *Cytogenet Genome Res*. 2005;110:91–107.
- Buchmann JP, Matsumoto T, Stein N, Keller B, Wicker T. Inter-species sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. *Plant J*. 2012;71:550–63.



29. Bennetzen JL. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev*. 2005;15:621–7.
30. Piegut B, Guyot R, Picault N, Roulin A, Sanyal A, Kim H, Collura K, Brar DS, Jackson S, Wing RA, Panaud O. Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. *Genome Res*. 2006;16:1262–9.
31. Kelly LJ, Renny-Byfield S, Pellicer J, Macas J, Novak P, Neumann P, Lysak MA, Day PD, Berger M, Fay MF, et al. Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. *New Phytol*. 2015;208:596–607.
32. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, Borrill P, Kettleborough G, Heavens D, Chapman H, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Res*. 2017;27:885–96.
33. IWGSC. A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science*. 2014;345:1251788.
34. Feldman M, Levy AA. Allopolyploidy—a shaping force in the evolution of wheat genomes. *Cytogenet Genome Res*. 2005;109:250–8.
35. Yaakov B, Meyer K, Ben-David S, Kashkush K. Copy number variation of transposable elements in *Triticum-Aegilops* genus suggests evolutionary and revolutionary dynamics following allopolyploidization. *Plant Cell Rep*. 2013;32:1615–24.
36. Eilam T, Anikster Y, Millet E, Manisterski J, Feldman M. Nuclear DNA amount and genome downsizing in natural and synthetic allopolyploids of the genera *Aegilops* and *Triticum*. *Genome*. 2008;51:616–27.
37. Marcussen T, Sandve SR, Heier L, Spannagl M, Pfeifer M, International Wheat Genome Sequencing Consortium, Jakobsen KS, Wulff BB, Steuernagel B, Mayer KF, Olsen OA. Ancient hybridizations among the ancestral genomes of bread wheat. *Science*. 2014;345:1250092.
38. Brookfield JF, Johnson LJ. The evolution of mobile DNAs: when will transposons create phylogenies that look as if there is a master gene? *Genetics*. 2006;173:1115–23.
39. Le Rouzic A, Payen T, Hua-Van A. Reconstructing the evolutionary history of transposable elements. *Genome Biol Evol*. 2013;5:77–86.
40. International Brachypodium Initiative. Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*. 2010;463:763–8.
41. Darrier B, Rimbart H, Balfourier F, Pingault L, Josselin AA, Servin B, Navarro J, Choulet F, Paux E, Sourdille P. High-resolution mapping of crossover events in the hexaploid wheat genome suggests a universal recombination mechanism. *Genetics*. 2017;206:1373–88.
42. Isidore E, Scherrer B, Chalhoub B, Feuillet C, Keller B. Ancient haplotypes resulting from extensive molecular rearrangements in the wheat A genome have been maintained in species of three different ploidy levels. *Genome Res*. 2005;15:526–36.
43. Bureau TE, Wessler SR. *Stowaway* - a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. *Plant Cell*. 1994;6:907–16.
44. Ramírez-González R, Borrill P, Lang D, Harrington SA, Brinton J, Venturini L, Davey M, Jacobs J, Van Ex F, Pasha A, et al. The transcriptional landscape of polyploid wheat. *Science*. 2018. <https://doi.org/10.1126/science.aar6089>.
45. Pophaly SD, Tellier A. Population level purifying selection and gene expression shape subgenome evolution in maize. *Mol Biol Evol*. 2015;32:3226–35.
46. Li A, Liu D, Wu J, Zhao X, Hao M, Geng S, Yan J, Jiang X, Zhang L, Wu J, et al. mRNA and small RNA transcriptomes reveal insights into dynamic homoeolog regulation of allopolyploid heterosis in nascent hexaploid wheat. *Plant Cell*. 2014;26:1878–900.
47. Venturini L, Caim S, Kaithakottil G, Mapleson DL, Swarbreck D. Leveraging multiple transcriptome assembly methods for improved gene structure annotation. *bioRxiv*. 2017; <https://doi.org/10.1101/216994>.
48. CLARITE. [github.com/jdaron/CLARITE-TE/](https://github.com/jdaron/CLARITE-TE/).
49. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*. 2008;9:18.
50. Eddy SR. Accelerated profile HMM searches. *PLoS Comput Biol*. 2011;7:e1002195.
51. vmatch. [www.vmatch.de](http://www.vmatch.de).
52. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–7.
53. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol Biol Evol*. 2016;33:1635–8.
54. biotools. <https://github.com/fchoulet/biotools>.
55. IWGSC, RefSeq\_v1.0 Assembly. [https://urgi.versailles.inra.fr/download/iwgs/IWGSC\\_RefSeq\\_Assemblies/v1.0/](https://urgi.versailles.inra.fr/download/iwgs/IWGSC_RefSeq_Assemblies/v1.0/).
56. IWGSC, RefSeq\_v1.0 Annotations. [https://urgi.versailles.inra.fr/download/iwgs/IWGSC\\_RefSeq\\_Annotations/v1.0/](https://urgi.versailles.inra.fr/download/iwgs/IWGSC_RefSeq_Annotations/v1.0/).
57. URGI IWGSC Data Repository. <https://wheat-urgi.versailles.inra.fr/Seq-Repository>.
58. IWGSC, raw sequencing data. <https://www.ncbi.nlm.nih.gov/sra/?term=SRP114784>.

**Ready to submit your research? Choose BMC and benefit from:**

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

**At BMC, research is always in progress.**

Learn more [biomedcentral.com/submissions](https://biomedcentral.com/submissions)

