The repetitive landscape of the barley genome

Thomas Wicker, Heidrun Gundlach, Alan Schulman

Introduction

Transposable elements are small genetic units with the ability to move around in the genome, make copies of themselves, or both. They range in size from a few dozen bp to tens of kb. TEs are found in practically all known organisms, which suggests that they are of very ancient evolutionary origin. It is generally believed that they evolved from cellular enzymes that modify or synthesize DNA (Malik and Eickbush, 2001; Gladyshev and Arkhipova, 2011.). TEs can be divided into two main classes based on their mode of replication (Wicker et al., 2007). These two main classes can be further differentiated into at least 29 superfamilies, ancient lineages of which are found in most eukaryotes (Wicker et al., 2007).

Class I elements (retrotransposons) replicate via reverse transcription of an mRNA intermediate that is transcribed from a cellular DNA copy. Autonomous retrotransposons all encode a reverse transcriptase-RNAse (RT-RH), which produces the double-stranded DNA from the mRNA template. The long terminal repeat (LTR) retrotransposons, which are evolutionarily younger than the long-interspersed elements (LINE) retrotransposons (Malik and Eickbush, 2001), encode as well an integrase (IN), which integrates the dsDNA made by RT-RH back into the genome. The RT-RH and IN are generally expressed as a polyprotein, which is cleaved into functional subunits by the aspartic proteinase that is also part of the polyprotein (Schulman, 2013). The LTR retrotransposons can reach very high copy numbers because each replication cycle from a single mRNA transcript can produce a new copy in the genome, which can in turn generate new copies. Due to their large size (~9 kb) and replicative capacity, LTR retrotransposons are the most dominant elements that determine the size of most plant genomes (Paterson et al., 2009, Schnable et al., 2010, International Brachypodium Initiative, 2010; Mascher et al., 2017), although in mammals the LINE elements and not the LTR retrotransposons (Chalopin et al., 2015).

Class II elements (DNA transposons) have the ability to excise from, and insert into, the genome by a "cut and paste" mechanism involving a transposase enzyme encoded by autonomous forms of the elements (Kempken and Windhofer, 2001). A typical DNA transposon is flanked by terminal inverted repeats (TIRs), which act as recognition sites for the transposase. Five different superfamilies of DNA transposons have been described in most plant genomes: *Harbinger*, *Mariner*, *hAT*, *CACTA* and *Mutator* (Wicker et al., 2007). Occasionally, entire superfamilies go extinct. For example, the banana genome apparently has lost all *CACTA* elements (d'Hont et al., 2012). In addition to the typical TIR DNA transposons, plant genomes contain considerable amounts of *Helitrons*, which belong to a curious sub-class of DNA transposons that do not have transposase genes but instead encode helicase enzymes. It is therefore assumed that they replicate via a rolling-circle mechanism (Kapitonov and Jurka, 2007). *Helitrons* are very abundant in some plant genomes (e.g., maize; Yang and Bennetzen, 2009) and their actual contribution to plant genomes is probably under-reported because they are extremely diverse and difficult to identify.

Autonomous TEs are defined as transposable elements that possess all genes and regulatory sequences needed for them to replicate and/or move around in the genome. Often, these autonomous elements give rise to large populations of non-autonomous derivatives which have lost some or all their genes and which depend on enzymes encoded by autonomous elements for their replication. The most extreme cases are so-called miniature inverted-repeat transposable elements (MITEs; Bureau and Wessler, 1994a; Bureau and Wessler, 1994b) which range in size from roughly 80 to 500 bp. Most plant MITEs are derived from elements of the *Mariner* and *Harbinger* superfamilies. In grasses. MITEs can vastly outnumber their autonomous partners. Indeed, the *Brachypodium distachyon* genome contains over 20,000 *Mariner* MITEs but only a few dozen potentially autonomous elements. Non-autonomous elements from other superfamilies are usually longer. For example, the highly abundant *Helitron*, *Mothra* from rice, is over 1,200 bp long (Roffler et al., 2015). Similarly, all *Triticeae* species contain very large non-autonomous *CACTA* elements, which encode only partial genes or no genes at all (Wicker et al., 2003).

Whereas classification of TEs into superfamilies is relatively simple, it is at the family level where most TE diversity is found. TE families are usually defined as groups of TE sequences that can be aligned as DNA over most of their sequence (>80% sequence identity over >80% of the entire TE length; Wicker et al., 2007). This definition of TE family is somewhat controversial, but nevertheless, it has proven useful and practical. Most plant genomes contain hundreds of different TE families. In the relatively small *B. distachyon* genome of 275 Mbp, over 170 different TE families have been described (International Brachypodium Initiative, 2010). Curiously, the number of TE families is similar in the much larger sorghum and maize genomes (Paterson, 2009; Schnable, 2009). Thus, it is not the number of different TE families that defines the genome size, but the copy numbers within individual TE families.

The barley genome is among the largest plant genomes sequenced and well assembled so far (Mascher et al., 2017). Nevertheless, at 5,100 Mb, it is close to the average of the plant genome sizes estimated to date (Wicker et al., submitted). Nevertheless, the distribution of plant genome sizes has a mode (i.e. peak) at approximately 587 Mbp, with a long tail towards very large genomes. Thus, it appears that there is some selection for genome sizes in the range of 100-1000 Mb and an apparent "typical" size of approximately 700 Mbp. The smallest plant genome sequenced so far is that of the carnivorous *Genlisea aurea,* which has a size of only 63 Mb (Leushkin et al., 2013). Interestingly, there seems to be no clear upper limit for genome sizes; many plants tolerate very large genomes with no phenotypic effect. The largest plant genome described so far is that of the lily *Fritillaria assyriaca*, which has a size of 120,000 Mbp (Leitch et al., 2007; Kelly et al., 2015). All angiosperms have very similar numbers of genes in their basic (monoploid) chromosome set; 32,000 genes of 3.5 kb each comprise together only about 112 Mbp of DNA. Hence, genome size is determined almost exclusively by the amount of TE-derived sequences. Barley represents plants with genomes that are much larger than both the mode of the genome size distribution for angiosperms and much larger than well-studied genomes until now, but close to the average of genome sizes that have been estimated. Thus, it can show us what to expect when even larger plant genomes in the future.

The repetitive fraction of the barley genome is dominated by a small number of high-copy TE families

Early on, it became obvious that the barley genome contains a few TE families that are present in extremely high copy numbers (Vicient et al., 1999; Middleton et al., 2012). The completion of the barley genome sequence (Mascher et al., 2017) revealed that ten *Gypsy*, three *Copia*, and two *CACTA* families together comprise over 50% of the whole genome (Figure 1). How many copies each of these families have in the genome is difficult to say, because many copies are fragmented by deletions or nested insertions of other TEs, or reduced to solo LTRs through intra-element recombination. Copy numbers of individual TE families can be estimated by dividing the total number of annotated base pairs by the length of the reference sequence for the respective TE**.** Using this approach, it was estimated that the 10 most abundant TE families together represent approximately 230,000 individual copies (Table 1, Wicker et al., submitted). The rest of the repetitive landscape is comprised of at least 350 TE families with moderate or low copy numbers (Mascher et al., 2017; Wicker et al., submitted). As described above for sorghum and maize, here too in barley the number of families is similar that found in smaller plant genomes despite the large difference in size among these genomes. Indeed, the relatively small *B. distachyon* genome (275 Mbp) went through a very detailed repeat annotation, leading to the identification of over 170 different TE families (International Brachypodium Initiative, 2010). By comparison, barley has less than twice as many TE families, although the barley genome is almost 20 times larger than the *B. distachyon* genome. Thus, the factor that determines genome size is the copy numbers of the most abundant families.

Almost 81% of the barley genome was classified as derived from TEs (Mascher et al., 2017). Considering that gene space contributes only 2-3% to the barley genome, approximately ~16% remains un-annotated so-called “dark sequence”. This proportion of un-annotated sequence is comparable to that in other genomes. In maize, approximately 12% remained un-annotated (Schnable et al., 2009), while in *B. distachyon*, un-annotated sequences comprise approximately 25% of the genome (International Brachypodium Initiative, 2010). It is assumed that the un-annotated portions of these genomes contain additional, yet uncharacterized, TE families (Wicker et al., submitted). These could be highly degenerate TEs, or exotic TE types that have very low copy numbers and thus escape detection. Future efforts will be needed to further characterize the un-annotated fractions in various genomes, but it is safe to say that the actual complexity of the repetitive fraction of plant genomes has likely been under-estimated.

BARE1 – the most abundant TE family in the barley genome

As previously described (Vicient et al., 1999; Chang et al., 2008; Middleton et al., 2012), the *Copia* family *RLC_BARE1* is the most abundant in terms of copy numbers (>72,000) as well as absolute contribution to the genome (>14%; Figure 1; Table 1), Together with other *Copia* RTNs, it is preferentially localized in the gene-rich distal regions of chromosomes (Mascher et al., 2017). *BARE1* is among the best characterized TE families in plants. Autonomous copies of *BARE1* contain a canonical *Copia* coding domain between the two LTRs that encodes, in the direction of transcription, the capsid protein Gag, integrase (INT), aspartic proteinase (AP), and the reverse transcriptase-RNAse H complex (RT-RH, Manninen and Schulman 1993, Suoniemi et al., 1996). *BARE1* is not only actively transcribed, but also translated, and forms virus-like particles (VLPs) (Jääskeläinen et al., 2013, Jääskeläinen et al., 1999). *BARE1* produces two groups of transcripts, one that can replicate via reverse transcription, which is not capped, polyadenylated, or translated (Chang et al., 2013). The other set is capped, polyadenylated, and translated, but not replicated. The second set of transcripts is also differentially spliced in response to stress to make more Gag protein for formation of VLPs. *BARE1* is not only actively transcribed, but also translated, and forms virus-like particles (VLPs) (Jääskeläinen et al., 2013, Jääskeläinen et al., 1999).

Non-autonomous retrotransposons, lacking one or more functional coding domain, are commonly encountered and perhaps are the dominant form in plant genomes (Sabot and Schulman, 2007). The inability of many of non-autonomous elements to carry out a full retrotransposon life cycle of transcription, translation, reverse transcription, packaging and integration (Figure 2) may be complemented by other, autonomous retrotransposons with which they share signals needed for these steps. The *BARE1* element has a non-autonomous form, *BARE2* (Tanskanen et al., 2007). *BARE-2* elements cannot synthesize their own Gag due to a deletion, which is conserved among *BARE2* elements, of the initiating ATG in the *gag* ORF. Nevertheless, *BARE2* contains the key *BARE1* signals for replication, including the PBS (Primer Binding Site) for reverse transcription, the DIS (DImerization Signal) for association of the two RNAs to be packaged, and the PSI (Packaging SIgnal) for packaging into VLPs. Indeed, *BARE2* economizes by not synthesizing the sub-genomic *gag* RNA but does transcribe the replication-competent RNAs (Chang et al., 2013). These are packaged into *BARE1* VLPs. The success of the *BARE2* strategy is indicated by it outnumbering *BARE1* by about 2:1 in the genomes of cultivated and wild barley, respectively *Hordeum vulgare* and *H. spontaneum.*

The ability of new RTN insertions to be inherited and drive genome size growth critically depends on where in the plant replication occurs. Immunolocalization with anti-Gag antibodies and *in situ* hybridization have shown that *BARE* protein and transcripts strongly vary from tissue to tissue (Jääskeläinen et al. 2013). Gag is strongly localized to provascular tissues and to companion cells in mature vascular tissues. Gag and *BARE* RNA appears in the developing floral spike, following transition to flowering. The localization of Gag in the floral meristems suggests that newly replicated copies there can be passed to the next generation. The visualized expression patterns are consistent with those expected from the response elements that have been identified in the *BARE* promoter.

The barley genome contains large populations of non-autonomous retrotransposons

Beyond *BARE2* described above, three of the five most abundant TE families seem to be non-autonomous (*RLG_Sabrina*, *RLG_WHAM*, and *RLG_Surya*), because they have none or only fragments of the genes that are typically found in autonomous elements (Figure 3). Thus, it is assumed that they rely on enzymes encoded by other TEs for their proliferation. The *Gypsy* family *RLG_Surya*, is possibly cross-mobilized by the much less abundant *RLG_Sukkula* family. Indications for this are a similar chromosomal distribution (Figure 3a) and strong sequence homology in the *RLG_Surya* and *RLG_Sukkula* LTRs. LTRs contain regulatory regions and serve as start points for replication. Such cross-mobilization has been described previously for *BARE2* elements (see above).

For the *Gypsy* families *RLG_Sabrina* and *RLG_WHAM*, no putative autonomous elements have been identified so far. Both *RLG_Sabrina* and *RLG_WHAM* can be subdivided into different subfamilies, some of them contain no coding sequences at all, i.e., the LTRs flank an internal domain of a few kb, which has no coding capacity. These are reminiscent of the widely distributed LARD elements, which are of full length (~9kb), but lack any coding capacity (Kalendar et al., 2004). Additionally, *RLG_Sabrina* and *RLG_WHAM* have subfamilies that contain a gene that probably encodes a Gag-like protein and a partial reverse transcriptase, similar to *Morgane* elements in wheat and its near relatives (Sabot et al., 2006). Sequence similarity of these partial proteins suggests that their autonomous master elements are homologs of the *Athila* retrotransposon from Arabidopsis (*Athila* clade, Figure 3a). Moreover, the *Copia* family *RLC_Giselle* likely depends upon closely related autonomous *RLC_Inga* family elements for transposition, because *RLC_Giselle* does not have *rt* and *int* genes whereas *RLC_Inga* does (Figure 3b). These observations indicate that non-autonomous retrotransposons mobilized by a relatively small number of autonomous elements contribute substantially to barley genome size.

In addition to the large *Gypsy*, *Copia*, and *CACTA* elements, which can range in size from roughly 2kb to over 30 kb, the barley genome also contains approximately 54,000 small non-autonomous DNA transposons of the *Mariner* and *Harbinger* superfamily (i.e., MITEs; Table 1). Most dominant is the *Mariner* superfamily, which is represented by at least 36 families. The 10 most abundant

Mariner families are all small non-autonomous elements ranging in size from 81 bp (*DTT_Athos*) to 274 bp (*DTT_Stolos* and *DTT_Pluto*; Table 1). Such small *Mariner* elements are also referred to as *Stowaway* MITEs (Bureau and Wessler, 1994b). The most abundant *Mariner* family, *DTT_Thalos*, is present in more than 17,000 copies. Interestingly, only about 150 potentially functional, autonomous *Mariner* elements have been identified in the barley genome (Wicker et al., submitted). Thus, a vast number of non-autonomous DNA transposons apparently rely on a very small number of functional master elements for their potential mobilization. The situation is similar for non-autonomous *Harbinger* transposons, although these elements are about four times less abundant (Table 1). Despite their enormously high copy numbers, the contribution of *Harbinger* and *Mariner* to barley genome size is negligible because of their shortness. Interestingly, these non-autonomous TEs are present in copy numbers similar to that in smaller genomes. Both rice and *B. distachyon* contain roughly 25,000 MITEs, whereas the barley genome contains approximately 54,000—only about twice as many—despite an over tenfold larger genome. We assume that this is related to MITEs being enriched near genes (Bureau and Wessler, 1994a; Bureau and Wessler, 1994b; Buchmann et al., 2012; Roffler et al., 2015; Wicker et al., 2016) and gene number being very similar in the monoploid set of chromosomes in all plant genomes.

Individual TE lineages occupy distinct chromosomal "niches"

Gypsy and *Copia* LTR retrotransposons are found throughout the genome, resulting in a more or less even distribution of coding sequences for reverse transcriptase and integrase along the chromosomes. However, at the individual family level, distributions vary strongly. For example, the *Copia* element *RLC_BARE1* is enriched in the distal regions of chromosome arms, as is the closely related but far less abundant *RLC_HORPIA2* (Figure 3b). In contrast, *RLC_Lara* and *RLC_Maximus* show a clear preference for proximal (peri-centromeric) chromosomal regions (Figure 3b). Retrotransposon families of the *Gyspy* superfamily occupy complementary genomic niches: the interstitial regions of chromosome arms are dominated by families from the *Athila* clade (*RLG_Sabrina*, *RLG_WHAM*, and

RLG_Derami, Figure 3a), while *RLG_Surya* and *RLG_Sukkula* are enriched in the proximal and distal regions. Generally, closely related families tend to have similar distribution patterns. Indeed, there is a good congruence between the phylogenetic tree of LTR retrotransposons and their chromosomal location (Figure 3). An interesting exception is the *RLG_Abiba* family, which is highly enriched in peri-centromeric regions, while its closest relative *RLG_Romina* shows a virtually inverse chromosomal distribution (Figure 3b).

DNA (Class II) transposon families also show distinct individual distribution patterns. The superfamily of the *CACTA* transposons is highly abundant and represented by at least 20 families in the barley genome. In total, they contribute at least 5% to the genome as a whole (Mascher et al., 2017). Among them, the proximal (centromeric and pericentromeric) regions are preferably occupied by the high-copy *CACTA* family *DTC_Balduin* (Figure 4). In contrast, families of the *Caspar* clade are strongly enriched in distal regions. Indeed, over 75% of *DTC_Caspar* elements are located in the terminal 20% of chromosome arms, the strongest niche enrichment we found for any TE group (Figure 4). For less abundant Class II superfamilies, such as *Mutator, Mariner,* or *Harbinger,* we observed the familiar pattern of enrichment in distal regions that was found in other plant genomes (Paterson et al., 2009; International Brachypodium Initiative, 2010; Han et al., 2013).

It is important to mention that compartmentalization into chromosomal niches was only observed for the distribution of large transposable elements such as LTR retrotransposons and *CACTA* elements. The very extensive populations of short non-autonomous elements (i.e. MITEs) tend to cluster near genes (Bureau and Wessler, 1994a; Bureau and Wessler, 1994b; Paterson et al., 2009; International Brachypodium Initiative, 2010; Han et al., 2013; see below), making their overall distribution largely congruent with that of genes.

The space surrounding genes is a distinct genomic compartment

In addition to large-scale gradients in the distribution of TE families, the barley gene space represents its own unique genomic compartment with its own TE "environment". Genes tend to be enriched in the distal chromosomal regions in of barley, with gene density forming an exponential gradient from centromeres to telomeres (Mascher et al., 2017). In addition to this gradient along chromosomes, genes are distributed non-randomly. They are found mostly in clusters of two to seven genes, (here, genes that are separated by less than 20 kb were defined as belonging to the same cluster). Individual clusters usually are usually separated by long stretches (hundreds of kb) that are comprised exclusively of TE sequences.

InterestinglyNotably, the TE landscape close to genes differs strongly from that of intergenic regions (here, we arbitrarily define "intergenic regions" as stretches of at least 200 kb that do not contain genes). This particular gene space environment is strictly local, including the gene and a few kilobases upstream and downstream, and largely independent of the gene's particular location along the chromosome. Of particular interest are insertions of LTR retrotransposons and *CACTA*s that are very near genes, because these generally large elements have the potential to influence the function of the gene. Retrotransposon composition changes drastically near genes: starting approximately 10 kb upstream and downstream of genes, the frequency of LTR retrotransposons (i.e. *Gypsy* and *Copia* elements) and *CACTA* elements drops sharply (Figure 5a).

As mentioned above, close to genes, we find mostly small, non-autonomous DNA transposons. More than a third (36%) of *Mariner* and 25.7% of *Harbinger* transposons are found within 5 kb of genes, a highly significant enrichment. Within 10 kb, this enrichment increases to almost 50% of *Mariner* and over 40% of *Harbinger* elements (Figure 5b). Interestingly, different types of elements also occupy different niches in the proximity of genes: *Mariner* transposons preferably reside immediately upstream and downstream of the coding regions of genes (Figure 5c), whereas *Harbinger* transposons are found further away. The observed distribution of different TE types around genes may reflect selective pressures that allow the smallest elements (*Mariners*) to be tolerated closest to genes. Most interestingly, *Helitrons* (particularly the high-copy families *DHH_Walter* and *DHH_Xobar*),as well as elements of the *Harbinger* superfamily, have a clear preference for promoter regions and are less abundant in downstream regions (Figure 5a). This asymmetric distribution suggests that their presence in promoters may be advantageous. However, it could also be that DNA transposons preferably insert near genes because chromatin is most accessible in transcriptionally active regions. Curiously, also LINEs show an asymmetric distribution around genes, and are found preferentially in downstream regions (Figure 5b). One explanation for the higher frequency of *LINEs* downstream of genes is that insertion of these relatively large elements in promoters might be deleterious, while they are tolerated better in downstream regions.

What molecular mechanisms drive genomic TE niche specificity?

It is still unclear how TE families “find” their chromosomal niches. Niche specificity could be driven by a preference of the respective transposase or integrase enzymes to bind to specific sequence motifs. Analysis of the insertion sites of several high-copy TEs, including *RLC_BARE1* and *RLG_Sabrina*, as well as of multiple families of *Mariner*, *Harbinger*, and *Helitron* elements, revealed pronounced differences in target site preference (Figure 6). Class II elements target very specific motifs: *Mariner* elements prefer A/T-rich targets with having the consensus [T/A][T/A]nnT-Ann[T/A][T/A], where the dash represents the insertion site (Figure 6a);, whereas *Harbinger* transposons prefer a short TAA motif (Figure 6b). AT-rich motifs (e.g. TATA boxes) are enriched in gene promoters in the barley genome (Table 2). This target preference could in part explain their preference for promoter sequences, or it could be the result of selection for promoters as their preferred site of insertion. However, it is also possible, as mentioned above, that these elements might simply target open chromatin (i.e., transcriptionally active) regions during transposition and establish themselves close to genes because their small size does not disrupt promoter function.

Interestingly, *Helitrons* have a preference for an asymmetric target, requiring an AAA triplet starting 8 bp downstream of an A-T insertion site (Figure 6c). Previous studies reported the preference of

Helitrons for a 5’-AT-3’ insertion site (Wicker et al., 2007) and for generally A/T rich sequences (Yang and Bennetzen, 2009). However, preference for an asymmetric target has, to our knowledge, not been reported previously for any type of TE. The asymmetric sequence composition of the target site suggests that the helicase/recombinase protein of *Helitron*s binds the target DNA at the insertion site as well as one rotational period away in the DNA double-helix (i.e. 10 bp).

In contrast to DNA transposons, LTR retrotransposons do not show obvious target site preferences: the high-copy LTR retrotransposon *RLC_BARE1* has a weak preference for G/C 7-8 bp away from the insertion site, while *RLG_Sabrina* has a slight preference for GGG motif 3-4 bp upstream of the insertion site and a CC motif 4 bp downstream. These findings are consistent with previous studies that showed very little target site preference for LTR elements (Abe et al., 2014). Nevertheless, different LTR retrotransposon families show very distinct chromosomal distributions. This suggests that their integrase enzymes target epigenetic patterns, such as histone modifications, rather than DNA sequence motifs. Previous studies reported that *RLG_Cereba* retrotransposons are particularly enriched in peri-centromeric regionss (Hudakova et al., 2001), as are its homologs (the CRM elements) in maize, rice, and *Brachypodium B. distachyon* (Schnable et al., 2009, International Brachypodium Initiative, 2010). However, for barley we could not confirm such enrichment (Figure 3a). Instead, we found that the *Abiba* family has taken over the proximal (peri-centromeric) ”niche” in barley. We speculate that its unique preference for centromeric regions may be due to the product encoded by an ORF that is not found in any other retrotransposon family (Figure 3a). This protein might have novel properties that enable *Abiba* elements to specifically target centromeric regions, potentially similar to the chromodomains in the integrase proteins of CRM elements that likely target centromere-specific histone modifications (Neumann et al., 2011).

Conclusions

The 5,100 Mb barley genome provides insight on the repetitive landscape of plant genomes that are large by current standards of analysis, but near the average size for angiosperms. The most striking characteristic of the barley TE landscape is that it is strongly compartmentalized. As we have described in this chapter, there are several hints as to what could drive this chromosomal niche specificity. The TEs surrounding genes are of particular interest because insertions in or near genes can alter the expression and function of genes (Hirsch and Springer, 2016). However, it is still not clear whether TEs play an active role in driving barley genome evolution by providing long-term fitness advantages, or whether they are merely present because they are selfish elements that have evolved successful strategies to amplify within the genome without causing deleterious effects. These two poles, respectively selectionist and neutralist, need not apply, of course, equally to all TEs in the genome, but rather represent a framework for future research.

References

Abe, H., Gemmell, NJ. (2014) Abundance, arrangement, and function of sequence motifs in the chicken promoters. BMC Genomics **15:**900.

Buchmann, J.P., Matsumoto, T., Stein, N., Keller, B., Wicker, T. (2012) Interspecies sequence comparison of Brachypodium reveals how transposon activity corrodes genome colinearity. Plant J. 488:213-7.

Bureau, T., and Wessler, S.R. (1994a) Mobile inverted-repeat elements of the Tourist family are associated with the genes of many cereal grasses. Proc. Natl. Acad. Sci. USA **9:** 907-916.

Bureau, T., and Wessler, S.R. (1994b) Stowaway: a new family of inverted repeat elements associated with the genes of both monocotyledonous and dicotyledonous plants. Proc. Natl. Acad. Sci. USA **9:** 1411-1115.

Chalopin, D., Naville, M., Plard, F., Galiana, D., Volff, J.N. (2015) Comparative Analysis of Transposable Elements Highlights Mobilome Diversity and Evolution in Vertebrates. Genome

Biol Evol 7: 567-580.

Chang, W., and Schulman, A.H. (2008) BARE retrotransposons produce multiple groups of rarely polyadenylated transcripts from two differentially regulated promoters. Plant J. 56:40-50.

Chang, W., et al. (2013) *BARE* Retrotransposons are translated and replicated via distinct RNA pools. PLoS One 8: e72270.

Gladyshev, E.A., and Arkhipova, I.R. (2011) A widespread class of reverse transcriptase-related cellular genes. Proc Natl Acad Sci USA 108: 20311-20316.

Han, Y., Qin, S., and Wessler, S.R. (2013) Comparison of class 2 transposable elements at superfamily resolution reveals conserved and distinct features in cereal grass genomes. BMC Genomics. **14:**71.

Hirsch, C.D., Springer, N.M. (2016) Transposable element influences on gene expression in plants. Biochim. Biophys. Acta. **S1874-9399:**30100-30106.

d'Hont A, Denoeud F, Aury JM, Baurens FC, Carreel F, Garsmeur O, Noel B, Bocs S, Droc G, Rouard M, et. al. (2012) The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature.* 488: 213-217.

Hudakova, S., Michalek, W., Presting, GG., ten Hoopen, R., dos Santos, K., Jasencakova, Z., and Schubert, I. (2001) Sequence organization of barley centromeres. Nucleic Acids Res. **29:**5029-35.

International Brachypodium Initiative. (2010) Genome sequencing and analysis of the model grass Brachypodiumdistachyon. Nature 463:763–768.

Jääskeläinen, M. et al. (1999) Retrotransposon *BARE*-1: Expression of encoded proteins and formation of virus-like particles in barley cells. Plant J 20: 413-422.

Jääskeläinen, M. et al. (2013) Retrotransposon *BARE* displays strong tissue-specific differences in expression. New Phytol 200: 1000-1008.

Kalendar, R., Vicient, C.M., Peleg, O., Anamthawat-Jonsson, K., Bolshoy, A., Schulman, A.H. (2004) LARD retroelements: Novel, non-autonomous components of barley and related genomes. Genetics 166:1437-1450.

Kapitonov, V.V., and Jurka, J. (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet. **23:**521-529.

Kelly, L.J., Renny-Byfield, S., Pellicer, J., Macas, J., Novák, P., Neumann, P., et al. (2015) Analysis of the giant genomes of *Fritillaria* (Liliaceae) indicates that a lack of DNA removal characterizes extreme expansions in genome size. New Phytol. 208:596-607.

Kempken, F., and Windhofer, F. (2001) The hAT family: a versatile transposon group common to plants, fungi, animals, and man. Chromosoma 110: 1-9.

Leushkin, E.V., Sutormin, R.A., Nabieva, E.R., Penin, A.A., Kondrashov, A.S., Logacheva, M.D. (2013) The miniature genome of a carnivorous plant Genlisea aurea contains a low number of genes and short non-coding sequences. BMC Genomics. **14:**476.

Leitch, I.J., Beaulieu, J.M., Cheung, K., Hanson, L., Lysak, M.A., Fay, M.F. (2007) Punctuated genome size evolution in Liliaceae. J Evol Biol. **20:**2296–308.

Malik, H.S., and Eickbush, T.H. (2001) Phylogenetic analysis of ribonuclease H domains suggests a late, chimeric origin of LTR retrotransposable elements and retroviruses. Genome Res 11: 1187-1197.

Manninen, I., and Schulman, A.H. (1993). *BARE*-1, a *copia*-like retroelement in barley (*Hordeum vulgare* L.). Plant Mol Biol 22: 829-846.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S.O., Wicker, T., et al. (2017) A chromosome conformation capture ordered sequence of the barley genome. Nature 544: 427-433.

Middleton, C.P., Stein, N., Keller, B., Kilian., B., Wicker T. (2012) Comparative analysis of genome composition in Triticeae reveals strong variation in transposable element dynamics and nucleotide diversity. Plant J. 73: 347-56.

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S., Wicker, T., Radchuk, V., et al. (2007) A chromosome conformation capture ordered sequence of the barley genome.

Neumann, P., Navrátilová, A., Koblížková, A., Kejnovský, E., Hřibová, E., Hobza, R., Widmer, A., Doležel, J., Macas, J. (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA. 2:4.

Paterson, A.H., Bowers, J.E., Bruggmann, R., Dubchak, I., Grimwood J., Gundlach H., et al. (2009) The *Sorghum bicolor* genome and the diversification of grasses. Nature **457:**551-6.

Roffler, S., and Wicker, T. (2015). Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. Mob DNA. **6:**8.

Sabot, F., and Schulman, A.H. (2006) Parasitism and the retrotransposon life cycle in plants: A hitchhiker's guide to the genome. Heredity 97: 381-388.

Sabot, F., Sourdille, P., Chantret, N., Bernard, M. (2006) *Morgane*, a new LTR retrotransposon group, and its subfamilies in wheats. Genetica 128:439-447

Schnable, P.S., Ware, D., Fulton, R.S., Stein, J.C., Wei, F. et al. (2009) The B73 maize genome: complexity., diversity., and dynamics. Science. 326:1112-1115.

Schulman A.H. (2013) Retrotransposon replication in plants. Curr Opin Virol 3: 604-614.

Suoniemi, A. et al. (1996). The *BARE*-1 retrotransposon is transcribed in barley from an LTR promoter active in transient assays. Plant Mol Biol 31: 295-306.

Tanskanen, J.A. et al. (2007) Life without GAG: The *BARE*-2 retrotransposon as a parasite´s parasite. Gene 390: 166–174.

Vicient, C.M., Suoniemi, A., Anamthawat-Jonsson, K., Tanskanen, J., Beharav, A., Nevo, E., and Schulman, A.H. (1999) Retrotransposon BARE-1 and Its role in genome evolution in the genus Hordeum. Plant Cell 11:1769–1784.

Wicker, T., Guyot, R., Yahiaoui, N., and Keller, B. (2003) *CACTA* Transposons in Triticeae. A Diverse Family of High-Copy Repetitive Elements. Plant Physiol. **132:**52-63.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, JL., Capy, P., Chalhoub, B., et al. (2007) A unified classification system for eukaryotic transposable elements. Nat Rev Genet. **8:** 973-82.

Wicker, T., Yu, Y., Haberer, G., Mayer, KFX., Marri, P.R., Rounsley, S., et al. (2016) DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. Nat Commun. **7:**12790.

Wicker, T., Schulman, A.H., Tanskanen, J., Spannagl, M., Twardziok, S., Mascher, M., Springer, N.M., Li, Q., Waugh, R., Li, C., Zhang, G., Stein, N., Mayer, KFX., Gundlach, H (2017) The repetitive landscape of the 5,100 Mbp barley genome, Mob. DNA, under revision.

Yang, L., and Bennetzen, J.L. (2009) Distribution, diversity, evolution, and survival of Helitrons in the maize genome. Proc. Natl. Acad. Sci. USA. **106:**19922-19927.

Yu, J., Hu, S., Wang, J., Wong, G.K., Li, S., Liu, B., et al. (2002) A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 296:79-92.

Table 1
Copy number estimates of the most abundant Class 1 and Class 2 element families in the barley genome.

TE family	Total kb[a]	Length[b]	Copy number[b]
RLC_BARE1	623,043	8,630	72,195
RLG_Sabrina	407,047	8,030	50,691
RLG_BAGY2	240,798	8,630	27,902
RLG_WHAM	167,138	9,450	17,687
RLG_Surya	163,300	14,470	11,285
RLC_Maximus	110,928	14,400	7,703
RLG_BAGY1	102,843	14,400	7,142
DTC_Balduin	70,688	11,740	6,021
RLG_Haight	57,185	13,080	4,372
DTC_Caspar	54,465	11,568	4,708
Total	**1,997,435**		**209,707**
DTT_Thalos	2,865	163	17,574
DTT_Pan	716	123	5,822
DTT_Athos	394	81	4,868
DTT_Icarus	555	117	4,747
DTT_Hades	392	108	3,627
DTT_SAF	177	85	2,087
DTT_Eos	506	326	1,552
DTT_Oleus	231	150	1,540
DTT_Pluto	328	274	1,197
DTT_Stolos	205	274	749
Total	**6,369**		**43,763**
DTH_Thorne	716	273	2,624
DTH_Kerberos	594	285	2,086
DTH_Xumet	591	376	1,571
DTH_Rong	1,218	1,227	993
DTT_Marimom	2,024	2,129	951
DTH_Orpheus	183	272	674
DTH_Xenon	203	312	650
DTH_Xian	650	1,161	560
DTH_Kong	489	2,119	231
DTH_Tibone	187	1,037	180
DTH_Zong	278	2,396	116
Total	**7,133**		**10,634**

[a]Total kb annotated as respective family-specific
[b]Length of the reference TE that was used for annotation
[c]Copy number estimate based on total kb occupied by the TE family and length of it consensus sequence.

Table 2

Sequence motifs that are enriched in promoter compared to intergenic sequences (i.e. 10 kb upstream of genes). Copy numbers were compiled for 2000 gene loci comparing the 1kb upstream of the transcription start site with the segment starting at 10 kb and ending at 9kb upstream of it.

Motif	Promoter[a]	Intergenic[b]	Enrichment
AAAAAA	5554	2492	2.2
CCGCCG	1333	585	2.2
CGCCGC	1380	619	2.2
GCCGCC	1378	666	2
TTTTTT	4874	2427	2
GCGGCG	1165	611	1.9
TAAAAA	2197	1219	1.8
TTAAAA	1549	850	1.8
CGGCGG	1056	617	1.7
TTTTAA	1573	920	1.7
TTTTTA	2110	1193	1.7
AAAAAG	1791	1112	1.6
TAAAAT	1481	875	1.6
TATAAA	1258	784	1.6
TTTAAA	1567	922	1.6

[a] Number of motifs found in the 1kb upstream of the transcription start site.

[b] Number of motifs found in the window 9-10 kb upstream of the transcription start site.

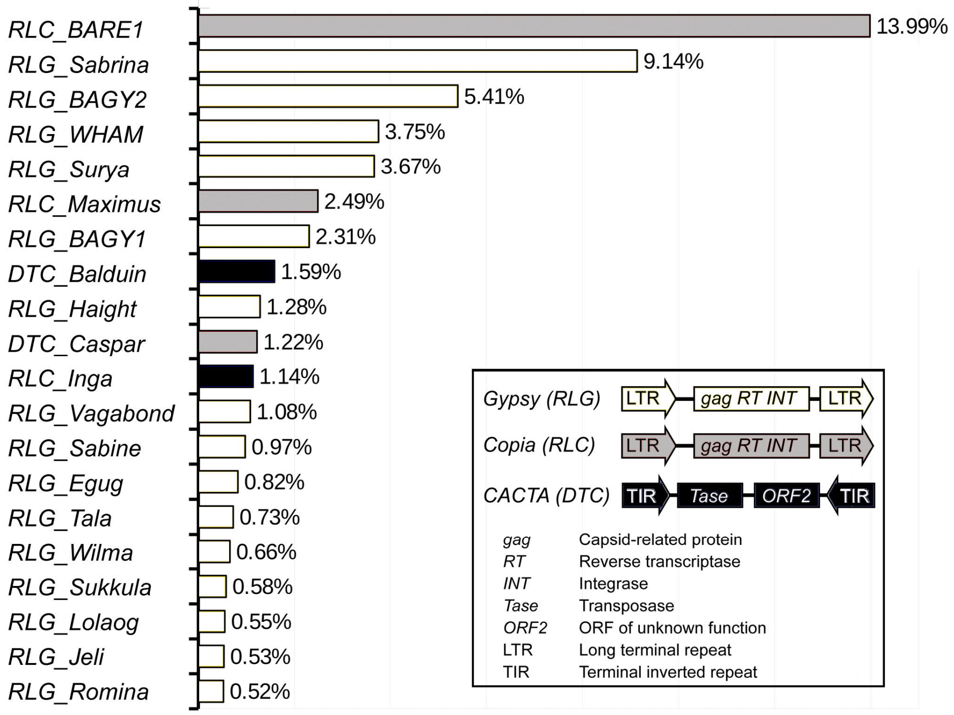

Figure 1. Contribution to total genome sequence of the top 20 TE families in the barley genome. Most broadly represented is the Gypsy superfamily to which 15 of the top 20 TE families belong (name prefix "*RLG_*"). The *Copia* superfamily is represented by three families (prefix "*RLC_*"). The only Class II superfamily represented in the top 20 is the *CACTA* superfamily (prefic "*DTC_*"). The inset shows the schematic sequence organization of these three superfamilies.

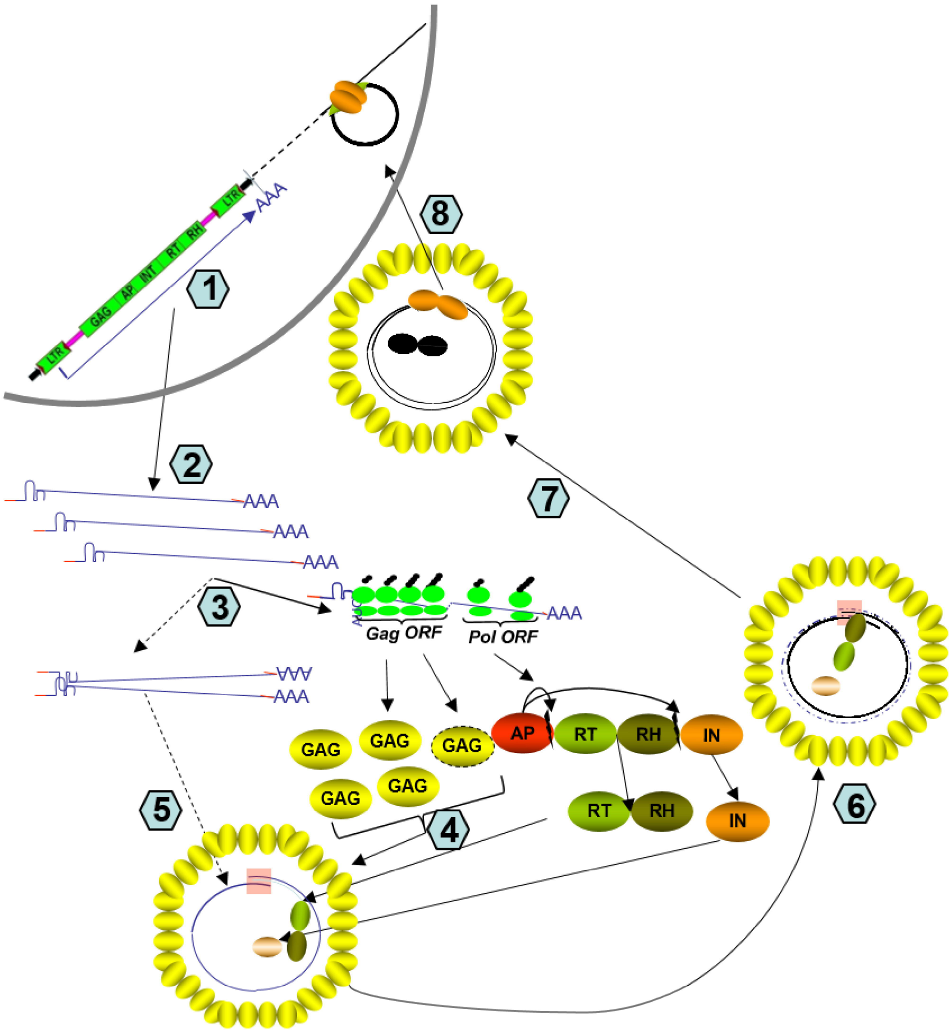

Figure 2. Retrotransposon *BARE1* life cycle. The steps of *BARE* replication are schematically depicted: 1: transcription; 2: nuclear export; 3: alternatively translation or packaging; 4: formation of virus like particles (VLPs); 5: packaging; 6: reverse transcription; 7: nuclear localization; 8: integration. The gery curve represents the nuclear envelope. GAG: capsid protein; AP: aspartic proteinase; RT: reverse transcriptase; RH: RNaseH; IN: integrase.

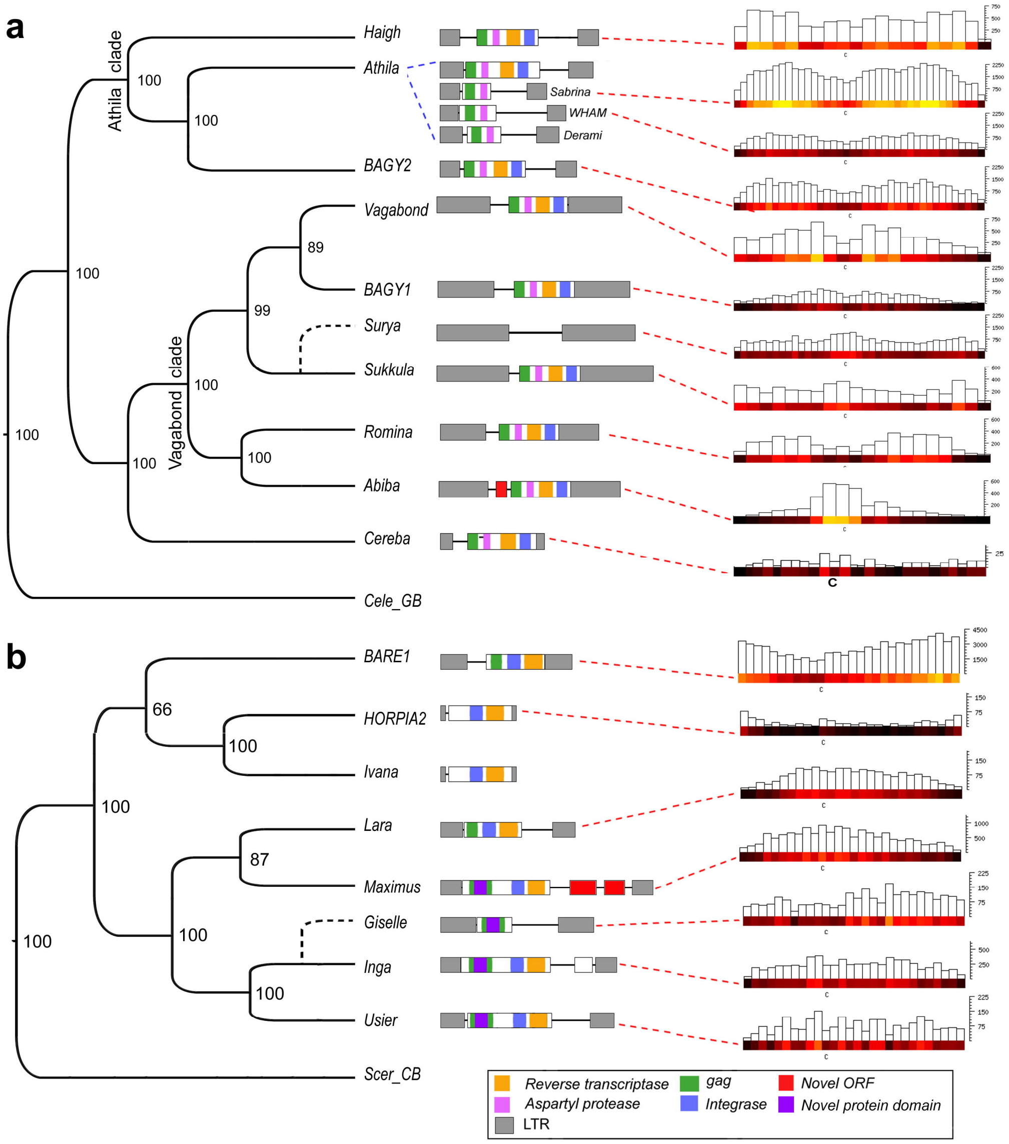

Figure 3. Distribution of *Gypsy* and *Copia* retrotransposons along barley chromosomes. Phylogenetic relationships of a selection of abundant families are shown at the left. Retrotransposon structure and gene content is shown in the center. Chromosomal distributions are shown at the right in bins of 20 to 40 Mb (depending on the copy number) as heat maps and bar plots to indicate absolute numbers. The y-axis indicates the total number of kb that is occupied by the TE family in each bin (Note that scales differ between families). Retrotransposon families with different evolutionary histories show different chromosomal distribution patterns. **a** Distribution of *Gypsy* elements on chromosome 2. **b** Distribution of *Copia* elements on chromosome 1.

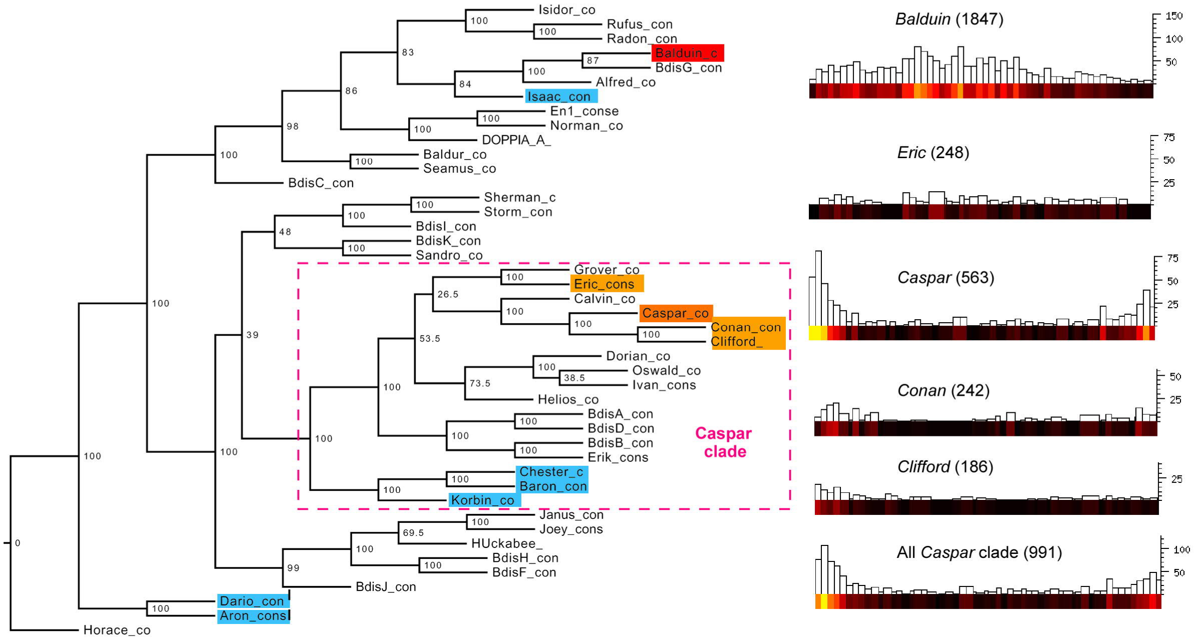

Figure 4. Phylogeny and distribution of *CACTA* elements in the barley genome. For the phylogenetic tree, 44 predicted *CACTA* transposase proteins deposited at TREP were used. *CACTA* sequences come from *Brachypodium distachyon*, Sorghum, rice, *Arabidopsis thaliana* and the Triticeae. High-copy elements from the Triticeae are highlighted in orange and red, while Triticeae low-copy families are highlighted in blue. Chromosomal distributions (shown is chromosome 1H as the representative for all barley chromosomes) are shown at the right in windows of 10 Mbp. Total copy number on 1H are given in parentheses next to the family name. *DTC_Balduin* dominates in centromeric regions, while elements of the *DTC_Caspar* clade occupy centromeres. It is not certain whether the preference for different chromosomal regions is

evolutionarily conserved, as similar analyses have not been done yet in other grasses. However, it is clear that *DTC_Caspar* and *DTC_Balduin* represent ancient lineages that were present already in the common ancestor of the grasses (Buchmann et al., 2014).

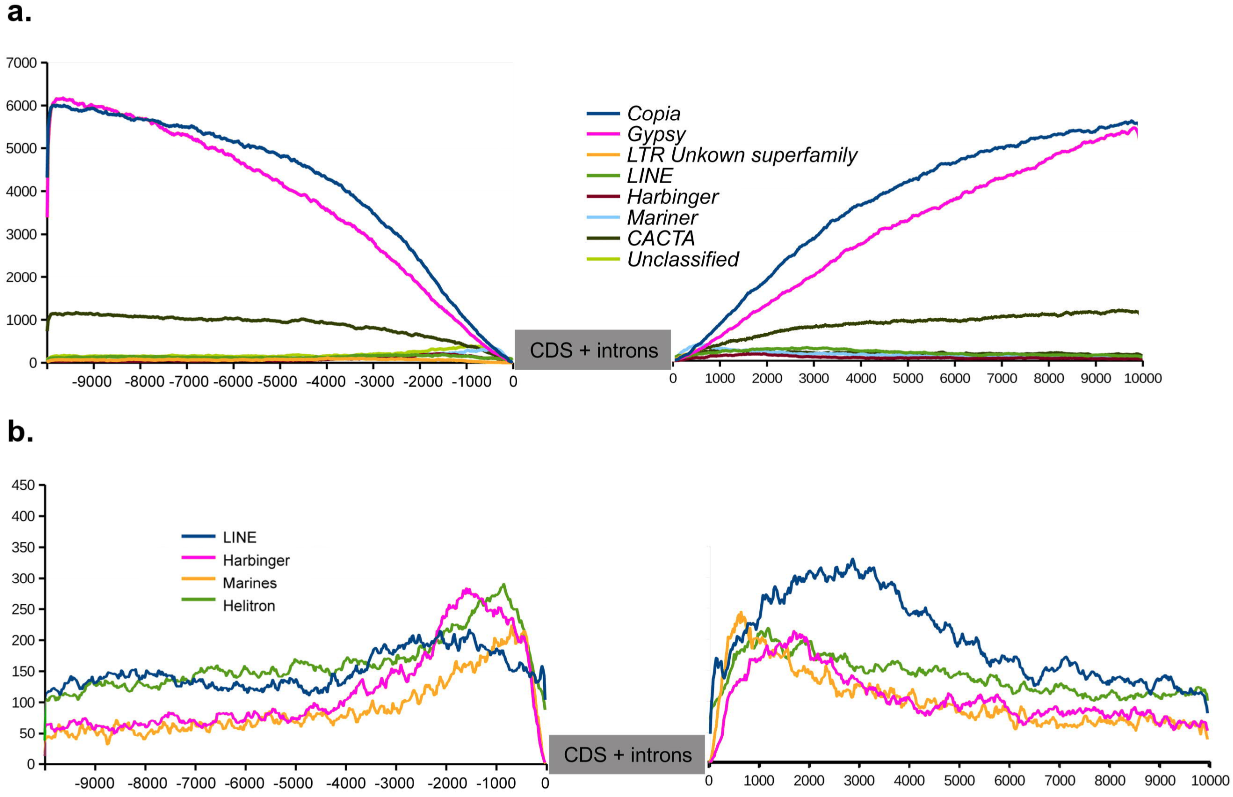

Figure 5. TE composition upstream and downstream of genes. The CDS of 28,316 high-confidence genes were used as anchor points. TE composition was determined 10 kb upstream and downstream of the gene. The x-axis indicates the position relative to the gene while the y-axis indicates how many genes had a TE of the respective superfamily at the particular position up- or downstream. Close to genes, Class 2 and LINE elements dominate. *Helitrons* and Harbinger elements have a clear preference for promoter regions while *LINE* elements are found poreferentially downstream of genes. **a.** Distribution of all major TE superfamilies around genes. With increasing distance from genes, *Gypsy* (RLG) and *Copia* (RLC) elements completely dominate genomic sequences, reflecting the overall composition of the barley genome. **b.** Zoom-in of the graph from (**a.**), displaying only the TE superfamilies that are enriched near genes. The y-axis was adjusted for better visibility of the less abundant superfamilies and the most abundant ones (*Gypsy*, *Copia* and *CACTA*) were omitted.

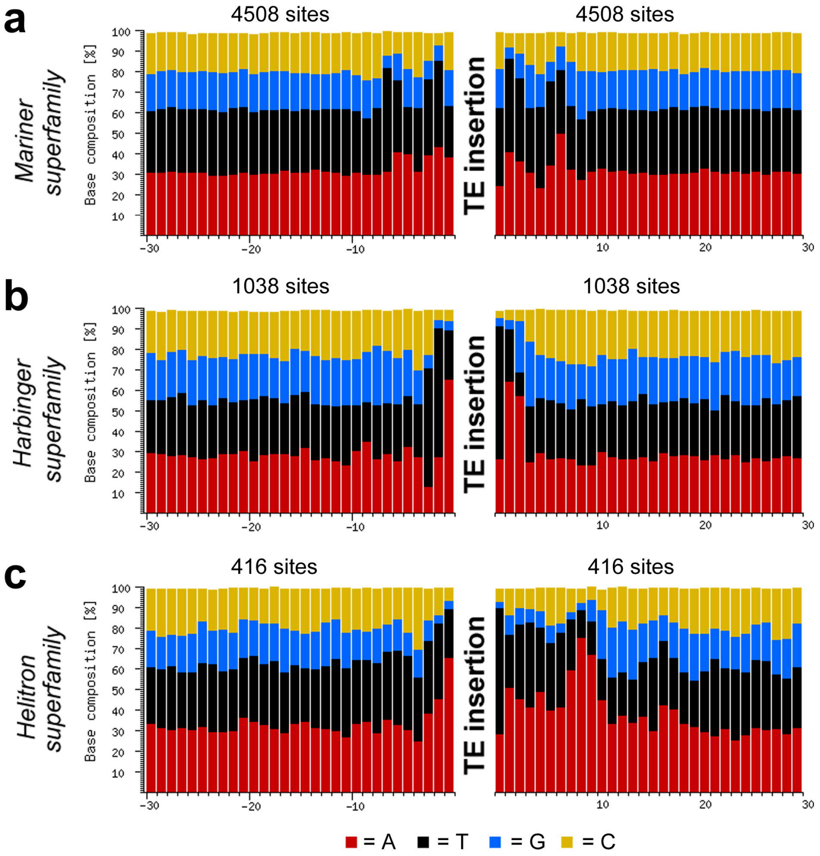

Figure 6. Target site preferences of high-copy Class 2 transposons from barley. For the plots, the 30 bp flanking complete (i.e., not truncated) elements on both sides were collected. Then the different nucleotides at each position were counted across across all insertion sites of a given TE type. The x-axis is the bp position relative to the TE insertion site, while the y-axis shows the relative nucleotide composition for each position. **a** *Mariner* elements have a strong preference for A/T dinucleotides 2 and 5 bp away from the insertion site, while (**b**) *Harbinger* elements almost invariably prefer 3 bp A/T -rich motifs. **c** Notably, *Helitrons* have a preference for an asymmetric target, strongly preferring an AAA motif 8 bp downstream of the insertion site.