

Chapter 9

LTR Retrotransposon Dynamics and Specificity in *Setaria italica*

Jeffrey L. Bennetzen, Minkyu Park, Hao Wang, and Hongye Zhou

Abstract The distributions of different LTR retrotransposon families and structures were analyzed across the ~400 Mb assembly for the ~500 Mb genome of *Setaria italica*. The results indicated different genomic distributions for all five of the highly abundant LTR retrotransposon families that were investigated. Unequal recombination and illegitimate recombination appeared to be more active in LTR retrotransposon removal in the gene-rich regions towards the ends of all chromosomes. In striking contrast to this result, LTR retrotransposon ages did not differ dramatically across the assembled genome, suggesting that LTR retrotransposon removal rates are not dramatically influenced by genomic location. These two, largely incompatible, observations indicate that the dynamics of LTR retrotransposon activation, insertion, and removal all need a great deal of additional investigation, including highly detailed intraspecies analyses and interspecies comparisons.

Keywords *Setaria italica* • Foxtail millet • LTR retrotransposons • Unequal recombination • Illegitimate recombination • Insertion times

9.1 Introduction

Transposable elements (TEs) are highly abundant in the nuclear genomes of all higher plants, usually constituting the majority of DNA in any species with a genome size exceeding 800 Mb (Bennetzen and Wang 2014). There is continued debate regarding the possible roles of these TEs. Although pure Darwinian theory indicates that selfish sequences like TEs would be obligated to come into existence via the process of natural selection for superior transmission (Doolittle and Sapienza 1980; Orgel and Crick 1980), there are many cases where these TEs have been co-opted by their hosts for a novel function, particularly in gene regulation and sometimes in the creation of new genes (Naito et al. 2009, reviewed in Feschotte 2008

J.L. Bennetzen (✉) • M. Park • H. Wang • H. Zhou
Department of Genetics, University of Georgia, Athens, GA 30677, USA
e-mail: maize@uga.edu

and in Bennetzen and Wang 2014). Overall, however, the biology and distribution of TEs is consistent with the properties of a selfish DNA (Daniels et al. 1990; Baucom et al. 2009).

In many genomes, including most or all flowering plant genomes (Grandbastien and Casacuberta 2012; Bennetzen and Wang 2014), TEs are the major source of genome rearrangement, either by chromosome breakage, by gene acquisition/mobilization or as sites for ectopic/unequal homologous recombination. In plant species with a strong record of recent TE activity, like maize, more than half of the nuclear genome can be structurally rearranged in as little as 1–2 million years, by a combination of macro and micro events (Wang and Dooner 2006; Wang and Bennetzen 2012).

Despite their ubiquity and importance in plants, there is surprisingly little information regarding the genomic properties or specificities of TEs. TEs within many angiosperm genomes have been detected, although pure whole genome shotgun (WGS) sequences are often masked for repeats, thence leading to a genome assembly that is deficient in most or all of the highly abundant TEs (e.g., Al-Dous et al. 2011). Once the TEs in a sequenced genome are found, the standard next step is for TEs to be broadly categorized by structural or homology criteria (Wicker et al. 2007). However, these searches often involve only discovery of highly repeated elements (which often make up a tiny minority of the TE families in a genome) or elements with homology to TEs that have already been described in other species. Even with the current wealth of deeply sequenced plant genomes (>95), these two criteria still can miss >50% of the LTR retrotransposon families that are present in a newly sequenced genome (H. Wang and J. Bennetzen, unpub. obs.). Applying sensitive structural criteria in genome analysis and annotation has led to particularly comprehensive TE discovery results (Schnable et al. 2009; Bennetzen et al. 2012; Hellsten et al. 2013).

Once discovered, the TEs in any “fully sequenced” genome can be mapped across chromosomes. This analysis has been undertaken on many occasions and has yielded the routine observation that different classes of TEs show very different abundances and very different genomic distributions. The LTR retrotransposons are routinely the most abundant TEs in plants, and the most routine LTR retrotransposon pattern has been that elements of the *Gypsy* superfamily mostly accumulate in heterochromatin (particularly pericentromeric heterochromatin) and that elements of the *Copia* superfamily show less of a bias towards the pericentromeric regions. More detailed analysis in maize, at the LTR retrotransposon family level, indicates that the likelihood of an LTR retrotransposon family’s insertion into euchromatin is inversely proportional to its copy number, regardless of the superfamily designation (Baucom et al. 2009). The DNA TEs, like the classic “controlling element” *Ac/Ds* studied by McClintock (McClintock 1956), tend to exist in lower copy numbers and show preferential association with genes. Of all TE families, the CACTA TEs (for instance, *Spm/dspm* of maize) seem to show the most “random” distribution, but all evidence to date for eukaryotes and prokaryotes indicates that TEs are far from random in their insertion or accumulation specificities (Schnable et al. 2009; Bennetzen et al. 2012).

The very complete sequence of the *Setaria italica* genome (Bennetzen et al. 2012) provides a particularly useful resource for TE characterization. In the first analyses, using the full spectrum of search criteria (repetitiveness, homologies to

known elements, and structural properties), TEs were found to constitute at least 40% of the genome. Most of the TE genome space (>60%) comprises LTR retrotransposons, including 98, 107, and 361 families of *Gypsy*, *Copia*, and “superfamily-unknown” LTR retrotransposons. As seen with many previously studied plant genomes, the *Gypsy* LTR retrotransposons were found to be enriched in the pericentromeric heterochromatin (Bennetzen et al. 2012). However, this general *Gypsy* observation might be driven by the properties of a few very abundant families, so a more detailed family-by-family analysis is warranted. This chapter provides that analysis and also includes investigations of the processes, rate, and genomic specificities of the loss of LTR retrotransposon sequences from the *S. italica* genome.

9.2 Results

9.2.1 *The Distributions of Different LTR Retrotransposon Families*

LTR retrotransposon families within a species are designated as distinct by an “80% homology rule” for their LTR nucleotide sequences (Wicker et al. 2007). This threshold was not purely arbitrary in its choice because the homology for LTR retrotransposons that show close internal relatedness is usually quite high (>90%) while those that lack close internal relatedness show little LTR homology (<50%). LTRs were chosen as the defining sequence source because they are more commonly intact (exclusively in solo LTRs, but also in complete/intact elements and fragments) than are the internal coding sequences. Figure 9.1 shows the distribution of the five most abundant families of LTR retrotransposons from the *Gypsy* and *Copia* superfamilies for two scaffolds in *S. italica*. As is obvious, the distributions of these families are quite different. Because all LTR retrotransposons appear to be removed from genomes by the same processes (Devos et al. 2002, reviewed in Bennetzen and Wang 2014), the differences between families are most likely to be caused by differences in insertion preferences.

9.2.1.1 LTR Retrotransposon Removal

LTR retrotransposons have been defined by our lab as intact (having both LTRs, and the appropriate target site duplications (TSDs)), fragmented (missing at least all or part of one LTR and usually missing some internal sequences), or solo LTRs (an LTR with no internal sequences, but with a TSD indicating that the solo LTR originated from unequal recombination between the two LTRs). Transposition of an LTR retrotransposon requires that it be intact, with two LTRs that are of the appropriate structure and orientation. An intact LTR retrotransposon does not need to have all of its internal coding sequences functional because some of the transposition functions

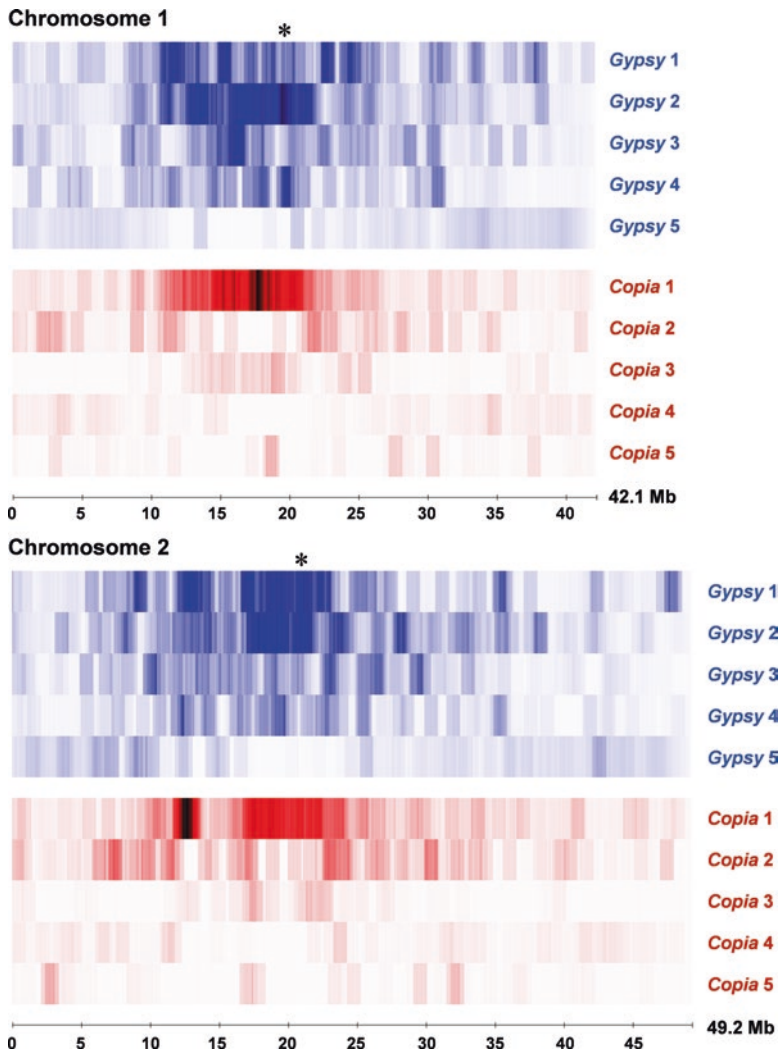


Fig. 9.1 Distributions of the ten most abundant LTR retrotransposon families across two *S. italica* chromosomes. The heat map was derived from a sliding window analysis of 1 Mb each, with 10 kb steps. Higher pigment density indicates higher LTR retrotransposon density, as determined by intact LTR retrotransposon amount (kb) in that window. The * indicates the approximate position of the centromere (Bennetzen et al. 2012). This same analysis was performed for all nine *S. italica* chromosomes, with similar results, but only two are shown here due to space considerations. It should be noted that these are actually scaffold depictions rather than full chromosomes, with numerous sequence gaps, especially in the centromeric regions

can be provided in trans (Jin and Bennetzen 1989), but it does need to have a primer binding site (PBS) and a polypurine tract (PPT), two short internal sequences needed for synthesis of the two strands of the integration intermediate. Because only an intact LTR retrotransposon can make an appropriate copy for transposition, then all solo LTRs and fragmented elements are an indication of LTR retrotransposon decay and removal (Devos et al. 2002). Figure 9.2 shows the distributions of solo LTRs, fragmented LTR retrotransposons, and intact LTR retrotransposons across two *S. italica* scaffolds.

As noted previously in rice (Ma and Bennetzen 2006), solo LTRs are in relatively low abundance in the low recombination regions around the centromere (*). This is an expected outcome of the low level of homologous recombination in these regions, and thus underscores the very rapid rate at which solo LTRs can be generated in euchromatic regions by unequal recombination. The fragmented LTR retrotransposons, primarily derived from various deletion processes (Kirik et al. 2000; Devos et al. 2002; Wicker et al. 2010), are also found primarily where LTR retrotransposons of all levels of intactness are most abundant (Fig. 9.2). Surprisingly, the fragmented LTR retrotransposons also show a higher ratio to intact LTR retrotransposons in the distal regions of chromosome arms. This indicates that the deletion mecha-

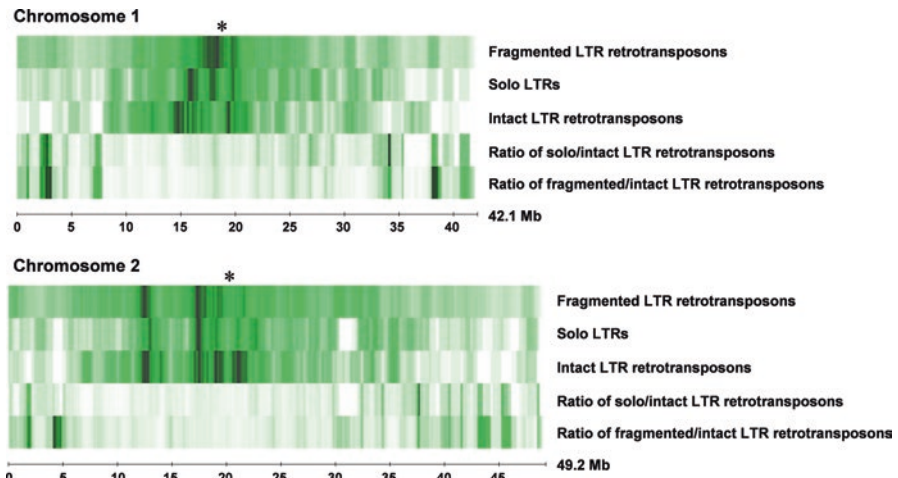


Fig. 9.2 Distributions and ratios of solo LTRs, fragmented LTR retrotransposons, and intact LTR retrotransposons across two *S. italica* chromosomes. The heat map was derived from a sliding window analysis of 1 Mb, with 100 kb steps. Higher pigment density indicates higher element density, as measured in kb in that window. In order to smooth the ratio curves, two kb of intact LTR retrotransposon was added to each window so that the denominator in the ratio was never zero. Hence, for this and several other reasons, the actual values of these ratios are not important, but the difference in the ratios should be noted across the chromosomal length. “Fragmented” LTR retrotransposons are defined as those that lack at least part of one LTR (and, often, additional internal sequences), while “intact” LTR retrotransposons are defined as those with two appropriate LTRs (in correct orientation and full size). The * indicates the approximate position of the centromere. This same analysis was performed for all nine *S. italica* chromosomes, with similar results, but only two are shown here due to space considerations

nisms involved in their fragmentation are not evenly active across the genome. This is an expected result for unequal recombination (which is known to preferentially occur in euchromatin) but not necessarily for fragmentation caused by illegitimate recombination (where possible differences in rates across the genome have not yet been thoroughly investigated).

9.2.2 Insertion Times of Detected LTR Retrotransposons

At the time of insertion, the mechanism of LTR retrotransposon replication indicates that the two LTRs in a single element will almost always be identical in sequence. Hence, divergence of the two LTRs in an element can be used to date the time that has elapsed since insertion (SanMiguel et al. 1998). This molecular clock has not been calibrated nearly so well as the molecular clock for “neutral” third codon positions in genes and is also likely to be somewhat variable between chromosome locations and species. Still, it provides an estimate of insertion date that is likely to be fairly consistent within a single genome for LTR retrotransposons with similar insertion biases. Unexpectedly, the results depicted in Fig. 9.3 indicate relatively similar average ages of LTR retrotransposons across the *Setaria* genome, suggesting that removal of TEs is not much more rapid from euchromatin than it is from heterochromatin.

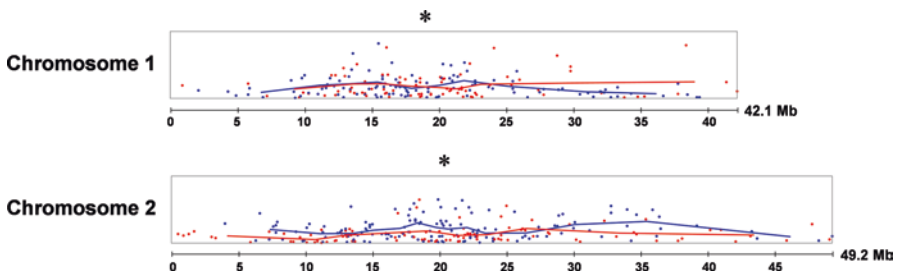


Fig. 9.3 Approximate insertion dates of all intact LTR retrotransposons across two *S. italica* chromosomes. *Individual dots* indicate individual elements and their insertion sites. The *curves* indicate average ages for the LTR retrotransposons of either the *Gypsy* (red) or *Copia* (blue) superfamilies in each 10-element window. The * indicates the approximate position of the centromere. This same analysis was performed for all nine *S. italica* chromosomes, with similar results, but only two are shown here due to space considerations. The X axis indicates position along the chromosome, and the Y axis indicates the degree of sequence divergence between two LTRs in the same element, which is an indication of the time that has expired since the TE inserted. Larger values mean more ancient insertion dates

9.3 Discussion

Although LTR retrotransposons make up the majority of most plant genomes, their distributions, origins, specificities, and transposition histories are only crudely understood in any plant system. More research is needed to look across a broad (and informatively selected) phylogenetic spectrum of organisms to see how TEs are transmitted and behave. In plants, where the median genome size is >4 Gb, the cost of de novo investigation of this process across hundreds of taxa remains prohibitive if one requires a fully sequenced and annotated genome. However, sample sequencing followed by repeat annotation of genomes can allow a cost-effective approach for this analysis (Devos et al. 2005; Macas et al. 2007). For instance, using only a few thousand Sanger sequences, we discovered that the doubling of the *Zea luxurians* genome size in the last 1–3 million years was caused by the amplification of numerous (but not all) LTR retrotransposon families in that genomic lineage (Estep et al. 2013).

Even when a genome is fully sequenced, the TEs are often given only a cursory investigation. This was certainly true in our earlier sequencing and annotation of the *S. italica* genome (Bennetzen et al. 2012). Even the much more comprehensive analysis of the TEs in the maize genome (Schnable et al. 2009; Baucom et al. 2009; Yang and Bennetzen 2009) left numerous questions regarding ancestry, activity, specificity, and fate unanswered. Questions investigated in detail in maize (but mostly not in any other genome sequence description in plants) were the number of LTR retrotransposon families (406), the frequency of gene fragment acquisition (high for *Helitrons* but also seen for MITEs, LTR retrotransposons, and other TE types), and TE distributions across the chromosomes. In *S. italica*, relatively little LTR retrotransposon data analysis was presented although >500 families were identified and mapped across the genome (Bennetzen et al. 2012 and unpub. data).

In our current analysis, we show that different LTR retrotransposon families have very different genomic distributions, as noted previously in maize (Baucom et al. 2009). It has been proposed (Bennetzen 2000, 2005) that this apparent differential insertion specificity is caused by recognition of different chromatin states, as is known to be the case for LTR retrotransposons in yeast (Kirchner et al. 1995; Zou and Voytas 1997). As future studies investigate chromatin structure in *Setaria* and other plants in more detail, it will be interesting to see which associations hold up between domains in integrase (the TE enzyme involved in opening up the host DNA for LTR retrotransposon insertion) and specific chromatin compositions and/or configurations. For instance, LTR retrotransposons that encode a chromodomain are more likely to be found in heterochromatic regions (Gao et al. 2008), suggesting one initial level of very general specificity. However, as there are likely to be thousands of actual “types” of heterochromatin or euchromatin, sporting different protein modifications and compositions (Bennetzen 2000), it is likely that integrases will have sufficient diversity to find unique target sites for most or all LTR retrotransposon families.

Although the general story that LTR retrotransposons in plants preferentially accumulate in pericentromeric regions has been adopted as a general concept, it is not a particularly accurate representation of reality. As seen previously in maize

(Baucom et al. 2009), this pericentromeric bias is quite common for the most abundant LTR retrotransposons, but has dramatic exceptions (see, for instance, the second most abundant *Copia* family in *Setaria*, *Copia 2*, which strongly avoids pericentromeric heterochromatin). Exceptions are also found to the direct relationship between copy number of the LTR retrotransposon and its preferential accumulation in heterochromatin although that role does hold true on average. It will be interesting, in future studies, to see what insertion niche is found by high-copy-number LTR retrotransposon families that do not find a safe haven in heterochromatin, as seen for *Gypsy 2*, *Copia 2*, and *Copia 4* in this study. We predict that these elements will not be found inserted into genes, but in class(es) of small heterochromatic blocks that are interspersed with genes, or perhaps in gene regulatory regions. If the latter, then these would be an excellent set of TEs to investigate for their ability to bring new genetic diversity to gene regulation (Kidwell and Lisch 1997; Feschotte 2008).

Our two most surprising results were (1) the lack of any dramatic difference in average LTR retrotransposon age between pericentromeric and euchromatic regions of *S. italica* chromosomes and (2) very uneven ratios of fragmented LTR retrotransposons to intact retrotransposons across the chromosomes. LTR retrotransposons were seen to average a somewhat more ancient time of insertion in pericentromeric heterochromatin, but a stronger effect was expected because both natural selection (to remove TEs that cause mutations) and random ectopic recombination are expected to decrease LTR content especially rapidly in genic regions. The relatively low frequency of solo LTRs in pericentromeric heterochromatin that we observed agrees with this prediction. Perhaps our results are caused by the fact that the LTR retrotransposons at the ends of the chromosomes are so rare that a useful ratio could not be determined or by the fact that not all of the pericentromeric DNA was assembled for *S. italica*, due to its highly repetitive DNA content. If values for these most dissimilar regions had been plotted, then one expects that a more impressive differential would have been observed. The second issue remains even more mysterious. One model suggests that the major mode of DNA removal from plants involves small deletion caused by illegitimate recombination, primarily as an outcome of inaccurate double strand break repair (DSBR) (reviewed in Bennetzen 2007). It is not at all clear why DSBR would be less common or less accurate in heterochromatic regions, but this is certainly implied by our data. More comprehensive analyses are needed investigating specific sequence change types and rates across plant chromosomes.

As with every other TE study conducted in plants, one is inundated with enormous numbers of possible interspecies and intraspecies investigations. Such studies can focus on TE effects on genome structure, on genome function and/or on gene evolution. Our study has provided a small part of this analysis, finding general similarity and some interesting differences with comparable studies in maize, rice, and other angiosperm genomes. We look forward to future studies that will investigate additional properties of these dynamic genome components.

Acknowledgements These analyses and the writing of this manuscript were supported by funding from the Giles Professorship to J. L. B. We thank Aye Htun for assistance with manuscript production.

References

- Al-Dous EK, George B, Al-Mahmoud ME, Al-Jaber MY, Wang H, Salameh YM, et al. *De novo* genome sequencing and comparative genomics of date palm (*Phoenix dactylifera*). *Nat Biotechnol.* 2011;29:521–7.
- Baucom RS, Estill JC, Chaparro C, Upshaw N, Jogi A, Deragon JM, et al. Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. *PLoS Genet.* 2009;5:e1000732.
- Bennetzen JL. The many hues of plant heterochromatin. *Genome Biol.* 2000;1:Reviews107.
- Bennetzen JL. Transposable elements, gene creation and genome rearrangement in flowering plants. *Curr Opin Genet Dev.* 2005;15:621–7.
- Bennetzen JL. Patterns in grass genome evolution. *Curr Opin Plant Biol.* 2007;10:176–81.
- Bennetzen JL, Wang H. The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annu Rev Plant Biol.* 2014;65:505–30.
- Bennetzen JL, Schmutz J, Wang H, Percifield R, Hawkins J, Pontaroli AC, et al. Reference genome sequence of the model plant *Setaria*. *Nat Biotechnol.* 2012;30:555–61.
- Daniels SB, Peterson KR, Strausbaugh LD, Kidwell MG, Chovnick A. Evidence for horizontal transmission of the P-transposable elements between *Drosophila* species. *Genetics.* 1990;124:339–55.
- Devos KM, Brown JKM, Bennetzen JL. Genome size reduction through illegitimate recombination counteracts genome expansion in Arabidopsis. *Genome Res.* 2002;12:1075–9.
- Devos KM, Ma J, Pontaroli AC, Pratt LH, Bennetzen JL. Analysis and mapping of randomly chosen bacterial artificial chromosome clones from hexaploid bread wheat. *Proc Natl Acad Sci U S A.* 2005;102:19243–8.
- Doolittle WF, Sapienza C. Selfish genes, the phenotype paradigm and genome evolution. *Nature.* 1980;284:601–3.
- Estep MC, DeBarry JD, Bennetzen JL. The dynamics of LTR retrotransposon accumulation across 25 million years of panicoid grass evolution. *Heredity.* 2013;110:194–204.
- Feschotte C. Opinion—transposable elements and the evolution of regulatory networks. *Nat Rev Genet.* 2008;9:397–405.
- Gao X, Hou Y, Ebina H, Levin HL, Voytas DF. Chromodomains direct integration of retrotransposons to heterochromatin. *Genome Res.* 2008;18:359–69.
- Grandbastien M-A, Casacuberta JM. Plant transposable elements impact on genome structure and function. Berlin: Springer; 2012. SpringerLink (Online service). doi:10.1007/978-3-642-31842-9.
- Hellsten U, Wright KM, Jenkins J, Shu S, Yuan Y, Wessler SR, et al. Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing. *Proc Natl Acad Sci U S A.* 2013;110:19478–82.
- Jin YK, Bennetzen JL. Structure and coding properties of Bs1, a maize retrovirus-like transposon. *Proc Natl Acad Sci U S A.* 1989;86:6235–9.
- Kidwell MG, Lisch D. Transposable elements as sources of variation in animals and plants. *Proc Natl Acad Sci U S A.* 1997;94:7704–11.
- Kirchner J, Connolly CM, Sandmeyer SB. Requirement of RNA polymerase III transcription factors for *in vitro* position-specific integration of a retroviruslike element. *Science.* 1995;267:1488–91.
- Kirik A, Salomon S, Puchta H. Species-specific double-strand break repair and genome evolution in plants. *EMBO J.* 2000;19:5562–6.
- Ma J, Bennetzen JL. Recombination, rearrangement, reshuffling, and divergence in a centromeric region of rice. *Proc Natl Acad Sci U S A.* 2006;103:383–8.
- Macas J, Neumann P, Navratilova A. Repetitive DNA in the pea (*Pisum sativum* L.) genome: comprehensive characterization using 454 sequencing and comparison to soybean and *Medicago truncatula*. *BMC Genomics.* 2007;8:427.
- McClintock B. Controlling elements and the gene. *Cold Spring Harb Symp Quant Biol.* 1956;21:197–216.

- Naito K, Zhang F, Tsukiyama T, Saito H, Hancock CN, Richardson AO, et al. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature*. 2009;461:1130–4.
- Orgel LE, Crick FH. Selfish DNA: the ultimate parasite. *Nature*. 1980;284:604–7.
- SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL. The paleontology of intergene retrotransposons of maize. *Nat Genet*. 1998;20:43–5.
- Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, et al. The B73 maize genome: complexity, diversity, and dynamics. *Science*. 2009;326:1112–5.
- Wang H, Bennetzen JL. Centromere retention and loss during the descent of maize from a tetraploid ancestor. *Proc Natl Acad Sci U S A*. 2012;109(51):21004–9.
- Wang QH, Dooner HK. Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. *Proc Natl Acad Sci U S A*. 2006;103:17644–9.
- Wicker T, Sabot F, Hua-Van A, Bennetzen JL, Capy P, Chalhoub B, et al. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet*. 2007;8:973–82.
- Wicker T, Buchmann JP, Keller B. Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res*. 2010;20:1229–37.
- Yang L, Bennetzen JL. Distribution, diversity, evolution, and survival of *Helitrons* in the maize genome. *Proc Natl Acad Sci U S A*. 2009;106:19922–7.
- Zou S, Voytas DF. Silent chromatin determines target preference of the *Saccharomyces* retrotransposon Ty5. *Proc Natl Acad Sci U S A*. 1997;94:7412–6.