Chapter 3
Using Nextgen Sequencing to Investigate Genome Size Variation and Transposable Element Content

Concepcion Muñoz-Diez, Clémentine Vitte, Jeffrey Ross-Ibarra, Brandon S. Gaut, and Maud I. Tenaillon

Abstract Transposable element (TE) content explains a large part of Eukaryotic genome size variation. TE content is determined by transposition, removal and host responses, but the efficiency of these forces is ultimately governed by genetic drift and natural selection. Contribution of TE families to genome size variation has been recently quantified using next generation sequencing (NGS) in two species pairs: *Zea mays* ssp. *mays* and *Zea luxurians*, *Arabidopsis lyrata* and *A. thaliana*. In both interspecific comparisons, genome-wide differences in TE content rather than the proliferation of a small subset of TE families was observed. We discuss three nonexclusive hypotheses to explain this pattern: selection for genome shrinkage, differential efficiency of epigenetic control, and a purely stochastic process of genome size evolution. Additional genome-wide assessments are needed to assess the extent to which selection shapes TE genomic content. To facilitate such studies, we discuss the use of NGS in "orphan" species.

Keywords Repetitive DNA • Selection • Genome shrinkage • Effective population size • Epigenetic control • Maize • Arabidopsis

C. Muñoz-Diez • B.S. Gaut
Department of Ecology and Evolutionary Biology, UC Irvine, 321 Steinhaus Hall, Irvine, CA 92617, USA

C. Vitte • M.I. Tenaillon (✉)
CNRS, UMR 0320 / UMR 8120 Génétique Végétale, INRA/CNRS/Univ Paris-Sud/ AgroParisTech, Ferme du Moulon, F-91190 Gif-sur-Yvette, France
e-mail: tenaillon@moulon.inra.fr

J. Ross-Ibarra
The Department of Plant Sciences and The Genome Center and Center for Population Biology, UC Davis, 262 Robbins Hall, Davis, CA 95616-5294, USA

M.-A. Grandbastien and J.M. Casacuberta (eds.), *Plant Transposable Elements*,
Topics in Current Genetics 24, DOI 10.1007/978-3-642-31842-9_3,

3.1 Introduction

Eukaryotes vary widely in genome size both within and among species. Genome sizes were first compared among species based on flow cytometry; subsequently CoT analyses revealed that most genome size variation is attributable to repetitive DNA. However, it is only with the development of DNA sequencing that we have been able to determine both the basis of this variation and to identify the mechanisms underlying it. In plants, for example, the comparison of large orthologous regions through BAC sequencing has lead to two important observations: first, the intergenic fraction of genomes is primarily comprised of transposable elements (TEs) and second, much of the genomic variation observed between species is due to the rapid turnover of TE sequences in intergenic regions (Ramakrishna et al. 2002; Ma and Bennetzen 2004; Wang and Dooner 2006).

Further analyses based on complete genome sequences has enabled precise quantification of the TE fraction for several taxa, revealing that the genomic fraction of TEs is positively correlated with genome size [Fig. 3.1, see Gaut and Ross-Ibarra (2008) for a review]. Moreover, analysis of full genomes has allowed characterization of the molecular bases of sequence turnover in intergenic regions: TE proliferation and elimination of TE sequences through homologous recombination and illegitimate recombination (Devos et al. 2002; reviewed in Vitte and Panaud 2003). Comparison of the extent and timing of the counteracting forces of proliferation and removal have revealed that large genomes harbor at least a few highly repetitive TE families in their genome, suggesting that some of the differences observed may be due to the capacity of some TEs to escape epigenetic control by the host genome (Vitte and Bennetzen 2006).

Genome size may therefore be determined by (1) the genome's intrinsic capacity to suppress TE activity by epigenetic mechanisms, and (2) the ability of TEs to escape this suppression system. In recent years, this idea has been strengthened by characterization of the molecular bases underlying this suppression system: the transcriptional and posttranscriptional silencing of TE sequences through pathways involving small interfering RNAs (siRNAs) (Lisch 2009). This characterization has revealed that siRNAs serve as molecular guides for silencing protein complexes to target TE sequences. Their presence is, therefore, an indicator of the deployment of a genomic defence mechanism toward silencing TEs and is correlated with the DNA methylation status of targeted sequences (Lister et al. 2008; Schmitz et al. 2011).

Beyond its structural impact on the genomic landscape, variation in TE content and genome size may have an evolutionary significance (Biemont 2008). For example, genome size correlates with rates of plant development, because smaller genomes presumably facilitate faster cell division and therefore a higher growth rate. In addition, a few studies have reported within-species correlations between genome size and ecological variables such as altitude, latitude, and temperature (see Knight et al. 2005 and references therein) and between genome size and phenotypes such as flowering time, flower size, leaf size, and photosynthetic rate (for a review, Knight et al. 2005; Meagher and Vassiliadis 2005). Species with smaller genomes

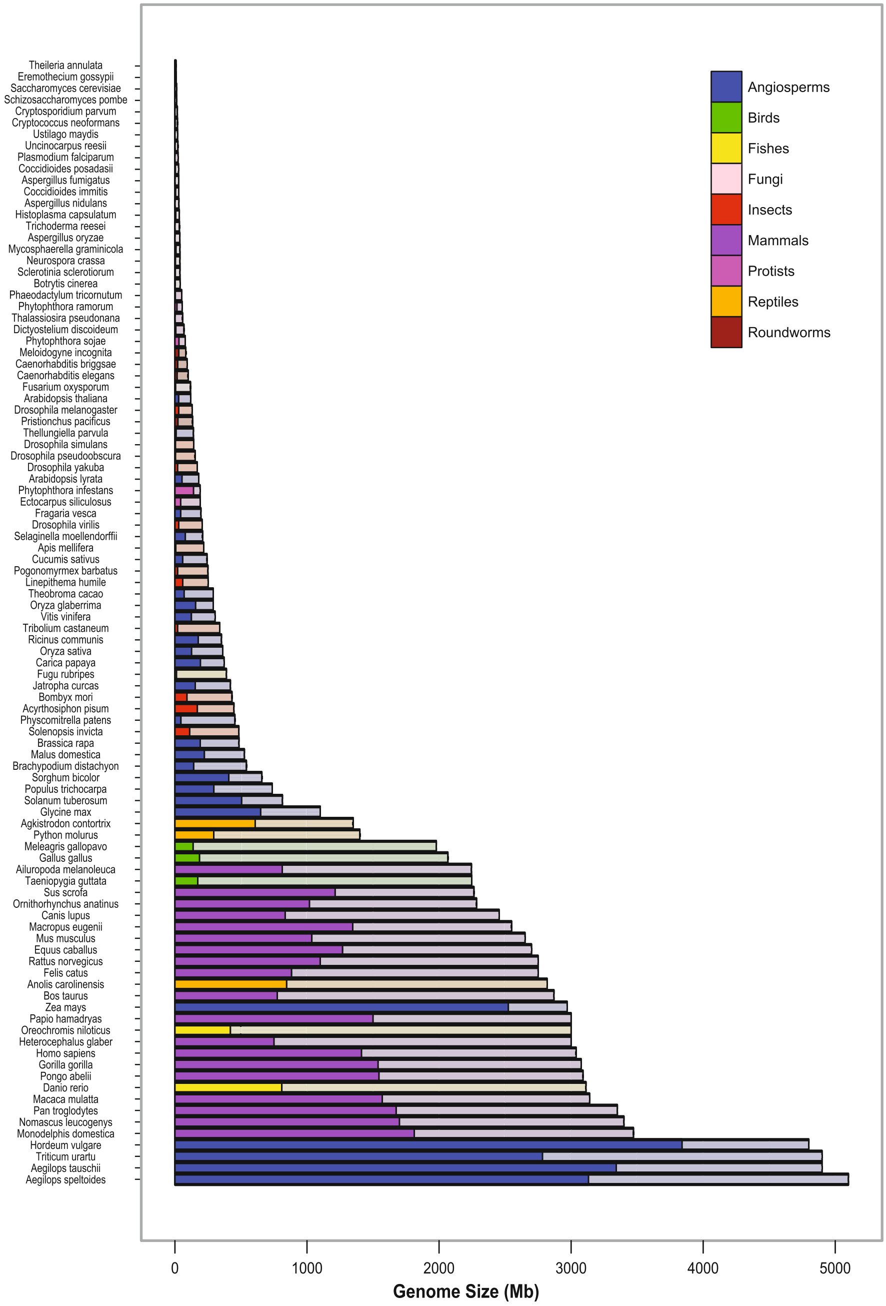


Fig. 3.1 Genome size (GS) and transposable element (TE) content of 98 eukaryote species, whose genomes have been sequenced. The total length of the *bars* indicates GS while the *darker portion* indicates TE content

also have enhanced colonization potential, due to an increase in seed mass, growth related traits, and decrease in generation time (Bennett et al. 1998; Grotkopp et al. 2004) that may altogether translate into a greater invasiveness (Lavergne et al. 2010).

While these examples appear to offer convincing evidence of the pervasiveness of the action of natural selection on genome size variation between and within-species, alternative hypotheses have been proposed. For example, a purely mechanistic model in which genome size evolves stochastically at a proportional rate can account for the skewed distributions of eukaryotic genome size (Oliver et al. 2007), but this model fails to provide a compelling reason for correlates between ecological factors and genome size. More recently, Whitney et al. (2010) have reported a lack of relationship between effective population size and genome size in angiosperms. Because the efficacy of selection is expected to scale with population size, the lack of relationship may indicate that selection has had little impact on broad-scale genome size evolution.

In summary, it is now well established that a balance between transposition, TE sequence removal, and host response determines a genome's TE content. These mechanisms are, in turn, affected by population processes, such as genetic drift and natural selection that ultimately determine the fate of TE insertions in plant genomes (Tenaillon et al. 2010). However, the extent to which selection shapes the TE genomic content is still debated. This debate would benefit greatly from genome-wide assessments that integrate across species and population levels—i.e., comparisons of genomes from various environments and taxa. Next Generation Sequencing (NGS) technologies provide such data, allowing exploration of the repetitive fraction of genomes.

Thus far, NGS has been employed largely for resequencing targeted regions in eukaryotic species with reference genomes on which NGS reads can be aligned (Li et al. 2010b; Xu et al. 2010) or for de novo assembly of prokaryotic or "simple" eukaryotic genomes with a restricted repetitive fraction (Galagan et al. 2005; Aury et al. 2008; Tenaillon et al. 2012). While de novo assembly of NGS data from more complex genomes such as the Giant panda (Li et al. 2010a), the human and the mouse (Gnerre et al. 2011), and *Arabidopsis thaliana* (Cao et al. 2011; Schneeberger et al. 2011) has been achieved, de novo approaches are still technically limited. Therefore, most NGS projects have been confined to describing sequence variants in the unique (single-copy) genomic fraction. However, NGS data can also be used to explore the components of repetitive DNA, such as TEs and satellite repeats, as well as their contribution to genome size variation within and among species.

In this chapter we will use the genus *Zea* as an example to illustrate how this can be achieved. Furthermore, we will take advantage of the recent publication of the *A. lyrata* genome (Hu et al. 2011) to establish a comparison between *A. thaliana*/*A. lyrata* on one hand and *Z. mays*/*Z. luxurians* on the other hand, and we will use these examples to discuss the factors that have contributed to genome size difference between closely related species. Finally, we will also provide some guidelines to determine TE content from NGS data in non-model species.

3.2 Exploring the Repetitive Fraction Within and Among Species Using NGS: An Example from the Genus *Zea*

3.2.1 Genome Size Variation in the Genus Zea

The genus *Zea* is traditionally divided into two sections (Fig. 3.2): *Luxuriantes* and *Zea*. The former encompasses several species, including the annual diploids *Z. luxurians and Z. diploperennis*. Section *Zea* includes a single diploid annual species (*Zea mays*), which consists of the cultivated maize (*Z. mays* ssp. *mays*) and its closest wild relatives (ssp. *parviglumis* and ssp. *mexicana*). The divergence between *Zea luxurians* and *Zea mays* is estimated to have occurred ~140,000 years ago (Hanson et al. 1996; Ross-Ibarra et al. 2009).

The genus encompasses extensive variation in genome size both within and between species. For example, within *Zea mays* genome size varies 30 % among cultivated accessions (i.e., landraces and inbred lines) and up to 32 % and 10 % in ssp. *mexicana* and ssp. *parviglumis*, respectively (Fig. 3.2 and included references). Between species, the average genome size of the diploid *Z. luxurians*, $2C = 9.07$ pg, is nearly 30 % larger than that of the average *Zea mays* ssp. *mays* genome (Fig. 3.2 and included references).

Differences in genome size may have multiple, potentially nonexclusive sources including whole genome duplication (polyploidy), segmental duplications, an increase of repetitive DNA (i.e., satellite sequences or TEs), or differential loss of TEs associated with recombination (Petrov et al. 2000). While *Z. luxurians* and *Z. mays* are both ancient polyploids (Gaut et al. 2000), extensive chromosomal rearrangements associated with the loss of some homeologs have resulted in the diploidization of *Zea* species, with $2n = 10$ chromosomes (Table 3.1). Therefore, variation between and within-species may arise from differences in the retention and the rate of production of segmental duplications as well as differential proliferation/elimination of repeated DNA.

In *Zea*, most repetitive DNA consists of interspersed TEs and heterochromatic blocks (knobs) which harbor 180- and 360-bp tandem repeats interspersed with retrotransposons (Peacock et al. 1981; Ananiev et al. 1998). Knob content varies among individuals of *Z. mays*, and knobs may be more abundant in *Z. luxurians* than in *Z. mays* (Tito et al. 1991; Gonzalez and Poggio 2011). Fully 85 % of the maize reference genome sequence consists of TEs, but the 20 most common TE families comprise ~70 % of the total (Baucom et al. 2009). These 20 families are all LTR retrotransposons (RNA elements). Amplification of LTR retrotransposons in the maize genome has been particularly dramatic in the last few million years, leading to a doubling of genome size (San Miguel and Bennetzen 1998; Brunner et al. 2005). Investigation of variation in TE copy number between *Z. luxurians* and *Z. mays* for six retrotransposon families using dotblots revealed little evidence of variation between species (Meyers et al. 2001), suggesting that these TEs may not have played a major role in genome size differentiation.

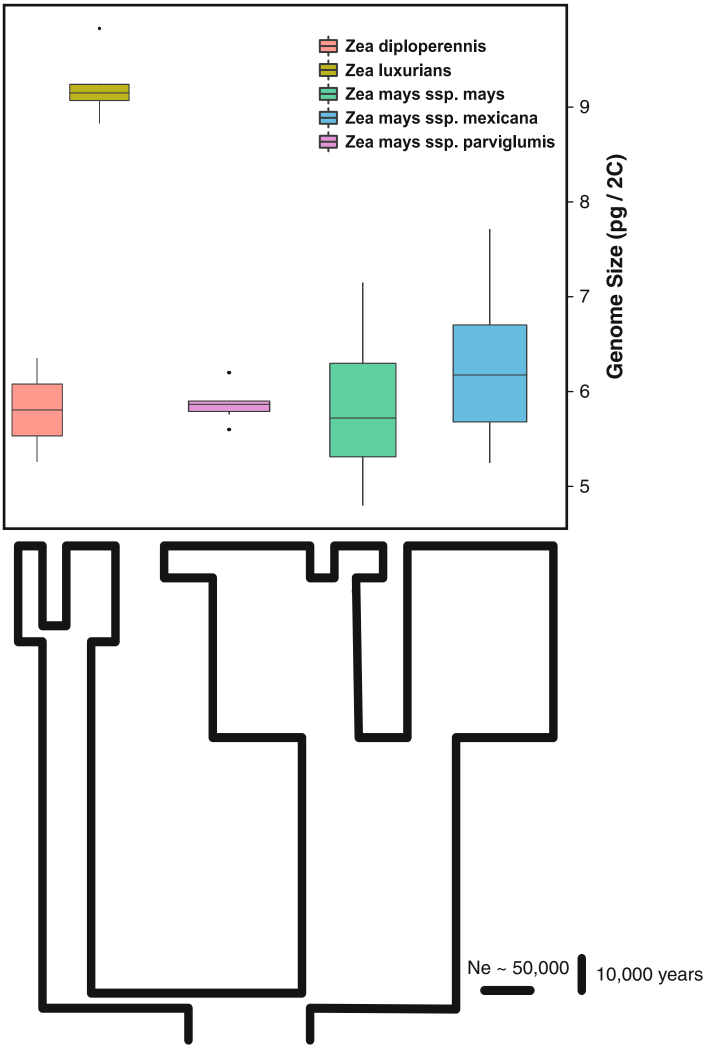


Fig. 3.2 Dendrogram and box plots showing demographic history and genome size variation in *Zea*. The branch width and length of the dendrogram are proportional to population size (N_e) and time, respectively, with scale bars shown (Ross-Ibarra et al. 2009). Divergence between *Z. mays* ssp. *parviglumis* and ssp. *mays*, and between *Z. mays* and *Z. luxurians*, was estimated to be 9,000 years (Piperno et al. 2009) and 140,000 years, respectively (Hanson et al. 1996; Ross-Ibarra et al. 2009). The *boxes* indicate the first quartile (*lower line*), the second quartile or median (*central line*), and the third quartile (*upper line*). Additionally the *whiskers* represent the standard deviation with the *dots* as the outliers. Genome size data were obtained from Laurie and Bennett (1985), Rayburn et al. (1985), Rayburn and Auger (1990), Tito et al. (1991), Guillin et al. (1992), Rayburn et al. (1993), Poggio et al. (1998), and Tenaillon et al. (2011) for a total of 2, 5, 8, 10, and 80 measures in *Z. diploperennis*, *Z. luxurians*, *Z. mays* ssp. *parviglumis*, *Z. mays* ssp. *mexicana* and Z *mays* ssp, *mays*, respectively

Table 3.1 Comparison of life-history traits, population parameters, and genomic content of *Arabidopsis* and *Zea* species

	A. lyrata	*A. thaliana*	*Z. mays*	*Z. luxurians*
Divergence time (Myr)	10[a]		0.140[b,c]	
Effective population size (N_e)	75,000[d]	250,000–300,000[e]	600,000[f]	50,000[b]
Mating system	Outcrosser	Selfer	Outcrosser +Recent inbreeding	Outcrosser
Genome size (Mb/C)	207[a]	125[a]	2,914[g]	4,435[g]
Chromosome number	$2n = 2x = 16$	$2n = 2x = 10$	$2n = 2x = 20$	$2n = 2x = 20$
Genes	32,670[a]	27,025[a]	39,656[h]	NA
TE content (% genome)	29.7[a]	23.7[a]	85[i]	NA
Ratio gene/TE	0.96[a]	1.78[a]	0.18[g]	0.18[g]

[a] Hu et al. (2011)
[b] Ross-Ibarra et al. (2009)
[c] Hanson et al. (1996)
[d] Ross-Ibarra et al. (2008). N_e value was calculated as the average among five subdivided populations
[e] Cao et al. (2011)
[f] Gossmann et al. (2010)
[g] Tenaillon et al. (2011)
[h] http://www.maizesequence.org
[i] Schnable et al. (2009)
NA not available

3.2.2 *Assessing the Contribution of TE Families to Genome Size Variation Between Maize and* **Z. luxurians** *Using NGS*

Recently, Tenaillon et al. (2011) performed a detailed analysis of TE content in one maize and one *Z. luxurians* genome using NGS. The approach was bolstered by the availability of a maize Filtered Gene Set (FGS) consisting of >32,000 high-quality annotated genes and a maize database of 1,526 exemplar (consensus) sequences representing distinct TE families and subfamilies (Baucom et al. 2009; Schnable et al. 2009). The method consisted of three discrete steps (Fig. 3.3). The first was creating a unique TE database (UTE) from the curated maize exemplar TE database (Baucom et al. 2009). The purpose of the UTE was to represent each of the 1,526 TE families of maize by their unique sequence signatures in order to minimize NGS reads that map ambiguously to more than one TE exemplar. In order to do so, each element of the exemplar TE database was cut into 104 bp fragments that were mapped against the exemplar TE database using the short read assembler SSAHA2 version 0.1 (Ning et al. 2001) with 80 % identity. Mapping results were used to determine the per base pair coverage of all 1,526 elements by the other elements contained in the exemplar TE database. This procedure allowed identification of portions of TEs not overlapping other elements in the exemplar database and to

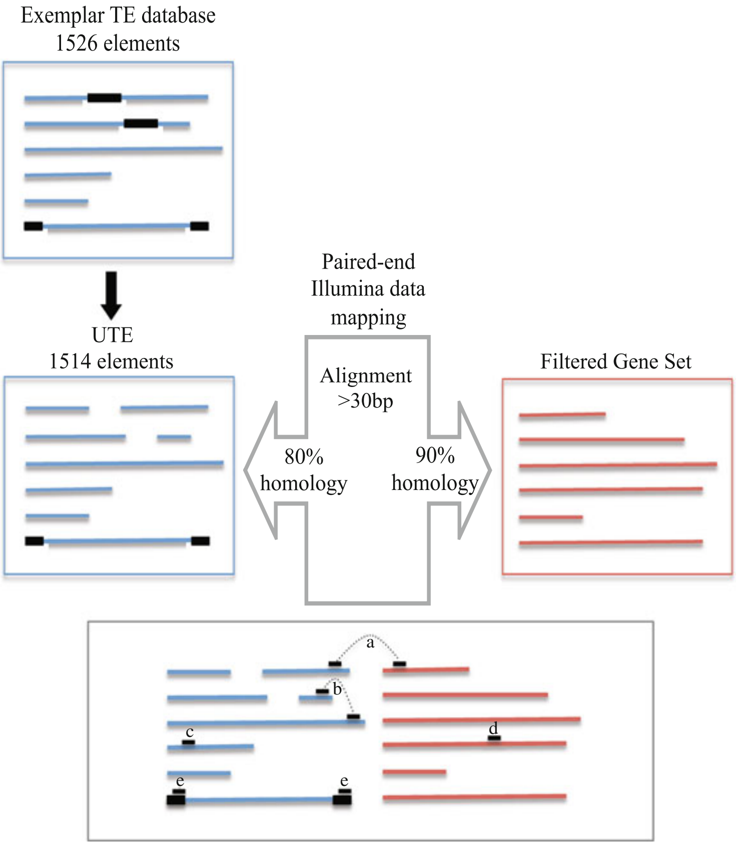


Fig. 3.3 Flowchart of the strategy used to analyze the TE content of maize and *Z. luxurians* genome using NGS data. The original exemplar TE database, represented here by six elements (*blue bars*), is filtered against the repeated regions among elements (*black boxes*). The resulting UTE contains the unique portion of each element, sometimes cut into several segments. Paired-end Illumina data are mapped against the UTE and the FGS, represented here by six genes (*red bars*). TE-gene pairs (**a**) and TE-nested pairs (**b**) are used to infer the proportions of TEs inserted into genes versus TEs inserted into other TEs. Read mapping against TEs (**c**) and genes (**d**) are used respectively to count the number of hits against a given element and estimate the coverage of the Illumina data. Note that because the UTE was not filtered against repeated regions within element (*black boxes*), two hits against a single element are counted only once

restrict the UTE to the sequences found in only a single TE in the exemplar database. Ultimately, the UTE consisted of 83 % of the original exemplar database, with 1,514 elements represented for read mapping (Tenaillon et al. 2011).

The second step was to generate high-throughput paired-end Illumina sequencing of the B73 maize inbred line and the *Z. luxurians* accession PI441933 (hereafter, *luxurians*). The paired-end libraries produced for each sample (B73 and *luxurians*) were each sequenced on a single lane of a flow cell with an Illumina Genome

Analyzer II, generating ~19 million paired-end reads of 84 and 104 bp in length. Tenaillon et al. (2011) also determined the genome size of the two accessions sequenced by flow cytometry: 5.96 pg/2 C for B73 and 9.07 pg/2 C for *luxurians*.

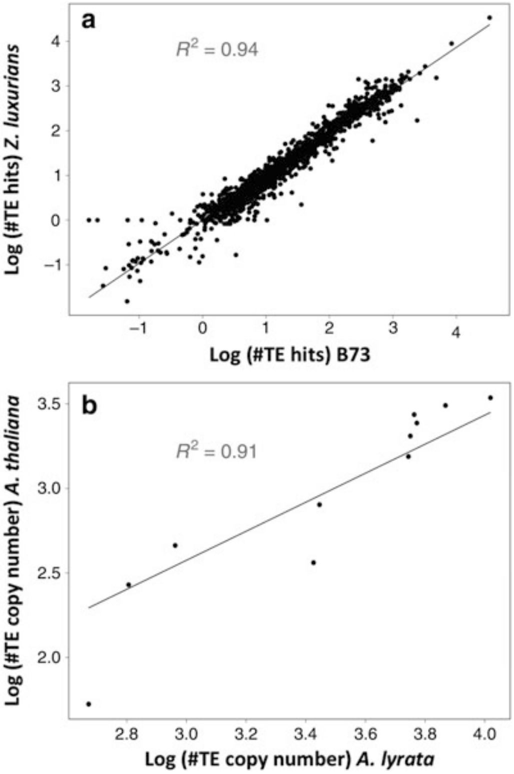
The third step was mapping the sequencing reads to the B73 reference genome, the UTE and the FGS, the latter providing an internal control for coverage. Using SSAHA2 version 0.1 (Ning et al. 2001) reads were mapped against the 1,514 elements of the UTE with 80 % identity, considering alignment length $\geq$30 bp. Reads aligning to a TE under these criteria were counted as single hit to the TE. One obvious caveat of the UTE is that the method as implemented is only as good as the annotated TE set, i.e., reads can only be mapped to annotated TEs. Median values of the distribution of per bp coverage from mapping of B73 and *luxurians* against each gene in the FGS were used to determine the genomic coverage of the Illumina data. In addition, by combining information about mapping against the UTE and FGS, it was possible to differentiate TEs inserted into other TEs (i.e., the two paired-ends mapped to two different TEs), from TEs inserted near genes (i.e., one paired-end mapped to a TE and the other to a gene).

Using the UTE and FGS from the maize reference genome, Tenaillon et al. (2011) were able to map 76.4 % and 75.5 % of reads to B73 and *luxurians*, respectively. They also verified reliability of their method via comparison between the Illumina data for B73 and the reference B73 genome. They observed >fivefold more TE-nested pairs than TE-gene pairs in both B73 and *luxurians*, indicating that TEs insert much more often in other TEs than genes. Assuming that gene content was similar between species, Tenaillon et al. (2011) found that at least 70 % of the 50 % genome size difference between maize and *Z. luxurians* (as determined by flow cytometry) was due to variability in TE copy number.

But the difference in genome size may have multiple origins. For example, it is possible that the *luxurians* genome encompasses genes and TEs that are absent from the B73 maize genome. These differences may occur as a consequence of differential genomic loss since species divergence. However, that similar proportions of reads were observed to map both to the UTE and FGS in both B73 and *luxurians* gives little support to this hypothesis, i.e., we would expect to observe significantly less mapping if TEs or genes present in *luxurians* were absent from B73. Alternatively, *luxurians* may exhibit a higher rate of retention of duplicated segments. If these duplicated segments offer a fair representation of the genome, encompassing both unique and repetitive DNA, one would expect to conserve similar proportions of gene to TEs and also TE families between species. Consistently, the proportion of mapped reads against FGS and UTE was similar in B73 (15.4:84.6) and in *luxurians* (14.8:85.2) and the number of hits to TE families was highly correlated between B73 and *Z. luxurians* (Fig. 3.4a, $r = 0.94$).

These observations are consistent with both TEs and genes being involved in genome size difference. They also reveal that differences in TE content between species are not due to the proliferation of a handful of TE families, as has been observed in other genera (Hawkins et al. 2006; Piegu et al. 2006), but rather due to a shift toward higher copy numbers in *Z. luxurians* for several hundred different TE families. Note, however, that *Gossypium* (Hawkins et al. 2006) and *Oryza* (Piegu

Fig. 3.4 Relative contribution of TE families to the genomes of two species pairs, the maize inbred line B73 and one accession of *Z. luxurians* (**a**), and the genomic sequences of *Arabidopsis thaliana* and *A. lyrata* (**b**). In (**a**), TE content was measured in 1,509 TE families as the number of Reads per Kilobase per Million mapped reads (RPKM) against the B73 Unique Transposable Element database (UTE). Values are shown on a log scale; the data are from Tenaillon et al. (2011). In (**b**), TE copy number was estimated from the annotation of the genomic sequence of *A. thaliana* and *A. lyrata* (Hollister et al. 2011)



et al. 2006) species divergence is much more ancient (on the order of a few million years) than in *Zea*, which may contribute to the difference between the observed patterns. For species with older divergence time, recurrent TE horizontal transfers between species are more likely to cause bursts of TE proliferation in the recipient species (Diao et al. 2006). This scenario seems less likely in *Zea*, not because there is no gene transfer among species but rather because there are likely no unique TEs among these recently diverged species that may easily escape the host suppression system.

3.3 Evolution of TE Profiles Through Evolutionary Times: A Comparison Between *Zea* and *Arabidopsis*

To date, the population dynamics of plant TEs have been studied primarily in the *Arabidopsis* species, *A. thaliana* and *A. lyrata*, which have relatively small genomes and for which reference genomes are available (Hu et al. 2011; the Arabidopsis Genome Initiative 2000). The two species diverged about 10 million years ago and exhibit several features that make their comparison especially interesting (Table 3.1). First, *A. lyrata* is a self-incompatible perennial while *A. thaliana* is a self-compatible annual species. Second, *A. lyrata* has $2n = 16$ chromosomes and its genome is larger than 200 Mb, whereas *A. thaliana* has $2n = 10$ chromosomes and one of the smallest angiosperm genomes at about 125 Mb. Third, Hu et al. (2011) have determined that more than 50 % of the *A. lyrata* genome appears to be missing from the *A. thaliana* reference genome but only about 25 % of the *A. thaliana* genome is absent from *A. lyrata*. Overall, *A. thaliana* exhibits a much higher ratio of genes to TEs than *A. lyrata*, and much of the genome size difference between these two species is likely caused by (1) reduced transposable element activity, (2) more efficient transposable element elimination in *A. thaliana*, especially near genes, or (3) systematic shortening of nontransposable element intergenic sequences and introns in *A. thaliana* (Fawcett et al. 2011; Hu et al. 2011).

Interestingly, Hollister et al. (2011) found a similar trend to the one observed in the *Zea* comparison (Tenaillon et al. 2011), which is that the relative contribution of TE families is well conserved between species (Fig. 3.4b, $r = 0.91$). Hence, in both interspecific comparisons, there are genome-wide differences in TE content rather than the proliferation of a small subset of TE families (as documented in *Gossypium* and *Oryza*). Two nonexclusive processes may help to explain this observation. First, there could be ongoing positive selection for genome shrinkage in both systems through the loss of TEs and genes. Supporting this idea, fewer insertions than deletions were found in a population of 95 individuals of *A. thaliana* among both segregating polymorphisms and fixed differences, with deletions longer on average than insertions (Hu et al. 2011). Moreover, a higher intron loss rate in *A. thaliana* than *A. lyrata* has been reported recently, reinforcing the hypothesis of selection for genome shrinkage (Fawcett et al. 2011). Additionally, simple calculations (Chevin and Hospital 2008) suggest that, in a species with a large effective population size similar to *Zea mays* (Fig. 3.2), even weakly beneficial mutations (TE deletions in this case) could increase to high frequency in timescales similar to the divergence between *luxurians* and *Zea mays* (Ross-Ibarra et al. 2009). If selection was driving this pattern, we would expect it to be more efficient in the species characterized by a greater effective population. While *A. thaliana* and *Z. mays* are thought to have higher effective population sizes than *A. lyrata* and *Z. luxurians* (Table 3.1), consistent with the observed differences in genome size, at least some estimates find weaker selection in *A. thaliana* than its congener (Wright et al. 2001; Lockton and Gaut 2010).

A second explanation is that closely related species may differ in aspects that control TE proliferation, such as the efficiency of epigenetic modification via pathways that include small interfering RNAs (siRNAs). Epigenetic mechanisms act by suppressing the expression of TEs (transcriptional silencing) or by cleaving TE mRNA (posttranscriptional silencing) (Slotkin et al. 2005; Matzke et al. 2009). Both pathways achieve site-specificity by homology between siRNA and their target sequences (Almeida and Allshire 2005). In plants, DICER-LIKE RNase enzymes produce 21–24-bp siRNA that guides ARGONAUTE and other downstream proteins to complementary DNA sequences, thereby promoting and maintaining DNA and histone methylation (Zhang 2008; Teixeira and Colot 2009). Hence, silenced TE sequences are generally characterized by identity with siRNAs and dense, even DNA methylation (Lippman et al. 2004; Zilberman and Henikoff 2007; Lister et al. 2008).

Differences in the efficiency of TE silencing by siRNAs has been investigated in *A. thaliana* and *A. lyrata* (Hollister et al. 2011). Sequences of siRNAs generated by NGS have been mapped to the two reference genomes and mapped siRNAs have been used as a proxy for TE methylation. Consistent with the hypothesis of differences in epigenetic control between the two species, the expression level of siRNAs was higher in *A. thaliana* by ~1.7-fold on average than in *A. lyrata*. The two species also exhibited a substantial difference in the ratio of uniquely- to multiply-mapping siRNAs. In fact a much higher proportion of TEs lacked uniquely mapping siRNA reads in *A. lyrata* (25 %) than in *A. thaliana* (10 %). Interestingly, Hollister et al. (2011) have shown that TEs targeted by uniquely mapping siRNAs are silenced more efficiently in both species. Altogether, lower TE expression levels, higher siRNA expression levels, and a higher ratio of unique/multiply-mapping siRNA signal more efficient silencing in *A. thaliana*, which correlates with its genomic characteristics: smaller genome and lower TE copy number. These phenomena should be evaluated in other pairs of closely related species with contrasting genome sizes, but reference genomes are still lacking in plant species to apply this approach.

Finally, it is also possible that genome size evolution is subject to a purely stochastic process in which the rate of genome size evolution (mean and variance) simply depends on current genome size, i.e., proportional evolution. Oliver et al. (2007) have supported this model by demonstrating the existence of a positive correlation between the rate of evolution and the average genome size in 20 eukaryotic taxonomic groups. The analysis of 68 eukaryotic sequenced genomes has revealed that the variation (as measured by standard deviation) of both the repetitive, i.e., masked, and unique, i.e., non-masked fraction, were proportional to the average repeat and unique fraction within a clade, suggesting that genome expansion is dominated by stochastic processes (Li et al. 2011). However, while genome size variation between closely related species such as described may be affected by drift, drift alone is difficult to reconcile with the observed ecological correlates of genome size.

3.4 Using NGS to Estimate TE Content and Diversity in Non-model Species

The examples presented above highlight how the availability of a reference genome and an exemplar TE database helps decipher the molecular origins of differences in TE content among species, by remapping short reads of DNA, RNA, or siRNAs.

But most species still lack a reference genome and are not closely related to a model species with a reference genome. When such a reference genome is not available, NGS can nonetheless serve to get a better understanding of TE content and diversity within a genome.

For species where BAC sequences are available, NGS can provide important help to refine TE annotation. Even though collections of TEs are now available for a vast number of species, these sequences may be too distant to the TEs of the species of interest. As a result, NGS reads from the focal species may match only to the most conserved regions of TEs from well-annotated species. For this reason, direct annotations of the focal species using computer tools such as Repeatmasker (Bedell et al. 2000) can lead to erroneous annotations, where TEs appear fragmented although they are not. The use of computer tools that look for specific structural features can provide de novo annotations in the focal genome. However, this approach is limited to TE families that harbor recognizable structural features (e.g., the terminal repeats of LTR retrotransposons) and to recent TE insertions that still harbor these features, leaving many TE copies unresolved.

This is where NGS may provide substantial help: TEs, which are repeated, are likely to show increased coverage as compared to unique sequences. Hence, mapping of NGS reads to a BAC sequence will delimit regions of high coverage (likely to be repeated) and regions with low coverage (likely low-copy). This, along with the annotation of conserved TE regions using TE databases from other species, may allow precise mapping of element boundaries. Of course, the detection of TE boundaries will be enhanced as sequencing coverage increases, but even low coverage may greatly facilitate annotation. NGS may thus be greatly valuable for TE annotation, which is the first step toward building a reference exemplar TE database for a given species. The quality and representation of the database will, however, depend on the number of BACs sequenced and whether they represent most or only a limited subset of TE families.

For species where no BAC sequences are available, NGS can still be used to generate consensus copies of the most abundant elements (exemplar TEs). For the same reasons presented above, highly repetitive elements will be represented by a large number of sequencing reads, which can then be used to reconstruct de novo consensus sequences of specific TE families. Such methodology has been implemented in the AAARF software (DeBarry et al. 2008) and has been successfully used on 454 reads. Adaptation of such tools to work on Illumina paired-end and mate-pair reads will likely provide improvements for TE detection. Note, however, that this approach will likely provide exemplar TE database of limited quality since the element builds may correspond to chimeric elements rather than a consensus sequence of several individual copies. For example, it may prove difficult to differentiate autonomous elements from their nonautonomous partners, because both may be merged in a single exemplar element. Nevertheless, such a database will be useful to determine a first approximation of TE content and diversity in the genomes of non-model species.

3.5 Conclusion

NGS technologies have enabled the generation of a vast amount of data. For complex genomes such as those of plants, their utilization has so far been limited to the analysis of the non-repetitive fraction of genomes, thus ignoring what is often the majority of the data. In this chapter, we illustrated how these data could be utilized to investigate the evolutionary processes driving variation in TE content, and hence genome size, among closely related species. The approach developed by Tenaillon et al. (2011) could, for species with a reference genome, be directly applied at the population level to assess the forces that determine TE content and the abundance of other heterochromatic repeats, as well as how repeat abundance relates to genome size variation. Coupled with NGS of siRNAs and mRNAs, such an approach may also provide substantial insights into the dynamics of TE methylation, its impact on gene expression (Hollister and Gaut 2009; Hollister et al. 2011), and more generally on the efficiency of the host response to TE invasion.

Application of this approach to species with no reference genome is more challenging. As a first step, we propose here to build exemplar TE databases using NGS to improve TE annotation from BAC sequences or for de novo TE assembly. Of course, these data will not provide a picture as complete as the one provided by a reference genome. In particular, it will not allow analysis of individual TE insertions, therefore hampering investigation of the distribution pattern of copies (e.g., between genic and nongenic regions) or the analysis of TE regulation by siRNAs. It nonetheless offers a first estimate of the most abundant elements and can be applied to many "orphan" species, thus providing a horizontal view of TE diversity among populations and species.

References

Almeida R, Allshire R (2005) RNA silencing and genome regulation. Trends Cell Biol 15:251–258

Ananiev EV, Phillips RL, Rines HW (1998) A knob-associated repeat in maize capable of forming fold-back DNA segments: Are chromosome knobs megatransposons? Proc Natl Acad Sci USA 95:10785–10790

Aury JM, Cruaud C, Barbe V, Rogier O, Mangenot S, Samson G, Poulain J, Anthouard V, Scarpelli C, Artiguenave F, Wincker P (2008) High quality draft sequences for prokaryotic genomes using a mix of new sequencing technologies. BMC Genomics 9:603

Baucom R, Estill J, Chaparro C, Upshaw N, Jogi A, Deragon J, Westerman R, Sanmiguel P, Bennetzen J (2009) Exceptional diversity, non-random distribution, and rapid evolution of retroelements in the B73 maize genome. PLoS Genet 5:e1000732

Bedell JA, Korf I, Gish W (2000) MaskerAid: a performance enhancement to RepeatMasker. Bioinformatics 16:1040–1041

Bennett MD, Leitch IJ, Hanson L (1998) DNA amounts in two samples of angiosperm weeds. Ann Bot 82:121–134

Biemont C (2008) Genome size evolution: within-species variation in genome size. Heredity 101:297–298

Brunner S, Fengler K, Morgante M, Tingey S, Rafalski A (2005) Evolution of DNA sequence nonhomologies among maize inbreds. Plant Cell 17:343–360

Cao J, Schneeberger K, Ossowski S, Gunther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Muller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D (2011) Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet 43:956–960

Chevin LM, Hospital F (2008) Selective sweep at a quantitative trait locus in the presence of background genetic variation. Genetics 180:1645–1660

DeBarry JD, Liu R, Bennetzen JL (2008) Discovery and assembly of repeat family pseudomolecules from sparse genomic sequence data using the Assisted Automated Assembler of Repeat Families (AAARF) algorithm. BMC Bioinformatics 9:235

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

Diao XM, Freeling M, Lisch D (2006) Horizontal transfer of a plant transposon. PLoS Biol 4:119–128

Fawcett JA, Rouzé P, Van de Peer Y (2011) Higher intron loss rate in *Arabidopsis thaliana* than *A. lyrata* is consistent with stronger selection for a smaller genome. Mol Biol Evol 29:849–859

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, Selker EU, Archer DB, Penalva MA, Oakley BR, Momany M, Tanaka T, Kumagai T, Asai K, Machida M, Nierman WC, Denning DW, Caddick M, Hynes M, Paoletti M, Fischer R, Miller B, Dyer P, Sachs MS, Osmani SA, Birren BW (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 438:1105–1115

Gaut B, Ross-Ibarra J (2008) Selection on major components of angiosperm genomes. Science 320:484–486

Gaut B, Le Thierry d'Ennequin M, Peek A, Sawkins M (2000) Maize as a model for the evolution of plant nuclear genomes. Proc Natl Acad Sci USA 97:7008–7015

Gnerre S, MacCallum I, Przybylski D, Ribeiro FJ, Burton JN, Walker BJ, Sharpe T, Hall G, Shea TP, Sykes S, Berlin AM, Aird D, Costello M, Daza R, Williams L, Nicol R, Gnirke A, Nusbaum C, Lander ES, Jaffe DB (2011) High-quality draft assemblies of mammalian genomes from massively parallel sequence data. Proc Natl Acad Sci USA 108:1513–1518

Gonzalez GE, Poggio L (2011) Karyotype of *Zea luxurians* and *Z. mays* subsp *mays* using FISH/DAPI, and analysis of meiotic behavior of hybrids. Genome 54:26–32

Gossmann TI, Song BH, Windsor AJ, Mitchell-Olds T, Dixon CJ, Kapralov MV, Filatov DA, Eyre-Walker A (2010) Genome wide analyses reveal little evidence for adaptive evolution in many plant species. Mol Biol Evol 27:1822–1832

Grotkopp E, Rejmanek M, Sanderson MJ, Rost TL (2004) Evolution of genome size in pines (Pinus) and its life-history correlates: supertree analyses. Evolution 58:1705–1729

Guillin EA, Poggio L, Naranjo CA (1992) Genome size in annual species of Zea. Relation with cellular parameters and altitude. Maize Genet Coop Newslett 66:59–60

Hanson M, Gaut B, Stec A, Fuerstenberg S, Goodman M, Coe E, Doebley J (1996) Evolution of anthocyanin biosynthesis in maize kernels: the role of regulatory and enzymatic loci. Genetics 143:1395–1407

Hawkins J, Kim H, Nason J, Wing R, Wendel J (2006) Differential lineage-specific amplification of transposable elements is responsible for genome size variation in Gossypium. Genome Res 16:1252–1261

Hollister J, Gaut B (2009) Epigenetic silencing of transposable elements: a trade-off between reduced transposition and deleterious effects on neighboring gene expression. Genome Res 19:1419–1428

Hollister J, Smith L, Ott F, Guo Y-L, Weigel D, Gaut B (2011) Transposable elements and small RNAs contribute to gene expression divergence between *Arabidopsis thaliana* and *Arabidopsis lyrata*. Proc Natl Acad Sci USA 108:2322–2327

Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, Haberer G, Hollister JD, Ossowski S, Ottilar RP, Salamov AA, Schneeberger K, Spannagl M, Wang X, Yang L, Nasrallah ME, Bergelson J, Carrington JC, Gaut BS, Schmutz J, Mayer KF, Van de Peer Y, Grigoriev IV, Nordborg M, Weigel D, Guo YL (2011) The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat Genet 43:476–481

Initiative TAG (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

Knight CA, Molinari NA, Petrov DA (2005) The large genome constraint hypothesis: evolution, ecology and phenotype. Ann Bot 95:177–190

Laurie D, Bennett M (1985) Nuclear DNA content in the genera Zea and Sorghum—intergeneric, interspecific and intraspecific variation. Heredity 55:307–313

Lavergne S, Muenke NJ, Molofsky J (2010) Genome size reduction can trigger rapid phenotypic evolution in invasive plants. Ann Bot 105:109–116

Li R, Fan W, Tian G, Zhu H, He L, Cai J, Huang Q, Cai Q, Li B, Bai Y, Zhang Z, Zhang Y, Wang W, Li J, Wei F, Li H, Jian M, Li J, Zhang Z, Nielsen R, Li D, Gu W, Yang Z, Xuan Z, Ryder OA, Leung FC, Zhou Y, Cao J, Sun X, Fu Y, Fang X, Guo X, Wang B, Hou R, Shen F, Mu B, Ni P, Lin R, Qian W, Wang G, Yu C, Nie W, Wang J, Wu Z, Liang H, Min J, Wu Q, Cheng S, Ruan J, Wang M, Shi Z, Wen M, Liu B, Ren X, Zheng H, Dong D, Cook K, Shan G, Zhang H, Kosiol C, Xie X, Lu Z, Zheng H, Li Y, Steiner CC, Lam TT, Lin S, Zhang Q, Li G, Tian J, Gong T, Liu H, Zhang D, Fang L, Ye C, Zhang J, Hu W, Xu A, Ren Y, Zhang G, Bruford MW, Li Q, Ma L, Guo Y, An N, Hu Y, Zheng Y, Shi Y, Li Z, Liu Q, Chen Y, Zhao J, Qu N, Zhao S, Tian F, Wang X, Wang H, Xu L, Liu X, Vinar T, Wang Y, Lam TW, Yiu SM, Liu S, Zhang H, Li D, Huang Y, Wang X, Yang G, Jiang Z, Wang J, Qin N, Li L, Li J, Bolund L, Kristiansen K, Wong GK, Olson M, Zhang X, Li S, Yang H, Wang J, Wang J (2010a) The sequence and *de novo* assembly of the giant panda genome. Nature 463:311–317

Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, Jiang H, Albrechtsen A, Andersen G, Cao H, Korneliussen T, Grarup N, Guo Y, Hellman I, Jin X, Li Q, Liu J, Liu X, Sparso T, Tang M, Wu H, Wu R, Yu C, Zheng H, Astrup A, Bolund L, Holmkvist J, Jorgensen T, Kristiansen K, Schmitz O, Schwartz TW, Zhang X, Li R, Yang H, Wang J, Hansen T, Pedersen O, Nielsen R, Wang J (2010b) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. Nat Genet 42:969–972

Li X, Zhu C, Lin Z, Wu Y, Zhang D, Bai G, Song W, Ma J, Muehlbauer GJ, Scanlon MJ, Zhang M, Yu J (2011) Chromosome size in diploid eukaryotic species centers on the average length with a conserved boundary. Mol Biol Evol 28:1901–1911

Lippman Z, Gendrel A, Black M, Vaughn M, Dedhia N, McCombie W, Lavine K, Mittal V, May B, Kasschau K, Carrington J, Doerge R, Colot V, Martienssen R (2004) Role of transposable elements in heterochromatin and epigenetic control. Nature 430:471–476

Lisch D (2009) Epigenetic regulation of transposable elements in plants. Annu Rev Plant Biol 60:43–66

Lister R, O'Malley R, Tonti-Filippini J, Gregory B, Berry C, Millar A, Ecker J (2008) Highly integrated single-base resolution maps of the epigenome in Arabidopsis. Cell 133:523–536

Lockton S, Gaut B (2010) The evolution of transposable elements in natural populations of self-fertilizing *Arabidopsis thaliana* and its outcrossing relative *Arabidopsis lyrata*. BMC Evol Biol 10:10

Ma J, Bennetzen J (2004) Rapid recent growth and divergence of rice nuclear genomes. Proc Natl Acad Sci USA 101:12404–12410

Matzke M, Kanno T, Daxinger L, Huettel B, Matzke A (2009) RNA-mediated chromatin-based silencing in plants. Curr Opin Cell Biol 21:367–376

Meagher TR, Vassiliadis C (2005) Phenotypic impacts of repetitive DNA in flowering plants. New Phytol 168:71–80

Meyers BC, Tingey SV, Morgante M (2001) Abundance, distribution, and transcriptional activity of repetitive elements in the maize genome. Genome Res 11:1660–1676

Ning ZM, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11:1725–1729

Oliver MJ, Petrov D, Ackerly D, Falkowski P, Schofield OM (2007) The mode and tempo of genome size evolution in eukaryotes. Genome Res 17:594–601

Peacock WJ, Dennis ES, Rhoades MM, Pryor AJ (1981) Highly repeated DNA-sequence limited to knob heterochromatin in maize. Proc Natl Acad Sci USA 78:4490–4494

Petrov DA, Sangster TA, Johnston JS, Hartl DL, Shaw KL (2000) Evidence for DNA loss as a determinant of genome size. Science 287:1060–1062

Piegu B, Guyot R, Picault N, Roulin A, Saniyal A, Kim H, Collura K, Brar D, Jackson S, Wing R, Panaud O (2006) Doubling genome size without polyploidization: dynamics of retrotransposition-driven genomic expansions in *Oryza australiensis*, a wild relative of rice. Genome Res 16:1262–1269

Piperno DR, Ranere AJ, Holst I, Iriarte J, Dickau R (2009) Starch grain and phytolith evidence for early ninth millennium BP maize from the Central Balsas River Valley, Mexico. Proc Natl Acad Sci USA 106:5019–5024

Poggio L, Rosato M, Chiavarino AM, Naranjo CA (1998) Genome size and environmental correlations in maize (*Zea mays* ssp. *mays*, Poaceae). Ann Bot 82:107–115

Ramakrishna W, Dubcovsky J, Park Y-J, Busso C, Emberton J, SanMiguel P, Bennetzen JL (2002) Different types and rates of genome evolution detected by comparative sequence analysis of orthologous segments from four cereal genomes. Genetics 162:1389–1400

Rayburn AL, Auger JA (1990) Nuclear-DNA content variation in the ancient indigenous races of mexican maize. Acta Bot Neerlandica 39(2):197–202

Rayburn A, Price H, Smith J, Gold J (1985) C-Band heterochromatin and DNA content in *Zea mays*. Am J Bot 72:1610–1617

Rayburn A, Biradar D, Bullock D, McMurphy L (1993) Nuclear DNA content in F1 hybrids of maize. Heredity 70:294–300

Ross-Ibarra J, Wright S, Foxe J, Kawabe A, DeRose-Wilson L, Gos G, Charlesworth D, Gaut B (2008) Patterns of polymorphism and demographic history in natural populations of *Arabidopsis lyrata*. PLoS One 3:e2411

Ross-Ibarra J, Tenaillon M, Gaut B (2009) Historical divergence and gene flow in the genus zea. Genetics 181:1399–1413

San Miguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Nat Genet 82:37–44

Schmitz RJ, Schultz MD, Lewsey MG, O'Malley RC, Urich MA, Libiger O, Schork NJ, Ecker JR (2011) Transgenerational epigenetic instability is a source of novel methylation variants. Science 334:369–373

Schnable PS, Ware D, Fulton RS, Stein JC, Wei F, Pasternak S, Liang C, Zhang J, Fulton L, Graves TA, Minx P, Reily AD, Courtney L, Kruchowski SS, Tomlinson C, Strong C, Delehaunty K, Fronick C, Courtney B, Rock SM, Belter E, Du F, Kim K, Abbott RM, Cotton M, Levy A, Marchetto P, Ochoa K, Jackson SM, Gillam B, Chen W, Yan L, Higginbotham J, Cardenas M, Waligorski J, Applebaum E, Phelps L, Falcone J, Kanchi K, Thane T, Scimone A, Thane N, Henke J, Wang T, Ruppert J, Shah N, Rotter K, Hodges J, Ingenthron E, Cordes M, Kohlberg S, Sgro J, Delgado B, Mead K, Chinwalla A, Leonard S, Crouse K, Collura K, Kudrna D, Currie J, He R, Angelova A, Rajasekar S, Mueller T, Lomeli R, Scara G, Ko A, Delaney K, Wissotski M, Lopez G, Campos D, Braidotti M, Ashley E, Golser W, Kim H, Lee S, Lin J, Dujmic Z, Kim W, Talag J, Zuccolo A, Fan C, Sebastian A, Kramer M, Spiegel L, Nascimento L, Zutavern T, Miller B, Ambroise C, Muller S, Spooner W, Narechania A, Ren L, Wei S, Kumari S, Faga B, Levy MJ, McMahan L, Van Buren P, Vaughn MW, Ying K, Yeh CT, Emrich SJ, Jia Y, Kalyanaraman A, Hsia AP, Barbazuk WB, Baucom RS, Brutnell TP, Carpita NC, Chaparro C, Chia JM, Deragon JM, Estill JC, Fu Y, Jeddeloh JA, Han Y, Lee H, Li P, Lisch DR, Liu S, Liu Z, Nagel DH, McCann MC, SanMiguel P, Myers AM, Nettleton D, Nguyen J, Penning BW, Ponnala L, Schneider KL, Schwartz DC, Sharma A, Soderlund C, Springer NM, Sun Q, Wang H, Waterman M, Westerman R, Wolfgruber TK, Yang L, Yu Y, Zhang L, Zhou S, Zhu Q, Bennetzen JL, Dawe RK, Jiang J, Jiang N, Presting GG, Wessler SR, Aluru S, Martienssen RA, Clifton SW, McCombie WR, Wing RA, Wilson RK (2009) The B73 maize genome: complexity, diversity, and dynamics. Science 326:1112–1115

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N, Henz SR, Huson DH, Weigel D (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc Natl Acad Sci USA 108:10249–10254

Slotkin R, Freeling M, Lisch D (2005) Heritable transposon silencing initiated by a naturally occurring transposon inverted duplication. Nat Genet 37:641–644

Teixeira FK, Colot V (2009) Gene body DNA methylation in plants: a means to an end or an end to a means? EMBO J 28:997–998

Tenaillon M, Hollister J, Gaut B (2010) A triptych of the evolution of plant transposable elements. Trends Plant Sci 15:471–478

Tenaillon MI, Hufford MB, Gaut BS, Ross-Ibarra J (2011) Genome size and transposable element content as determined by high-throughput sequencing in maize and *Zea luxurians*. Genome Biol Evol 3:219–229

Tenaillon O, Rodriguez-Verdugo A, Gaut RL, McDonald P, Bennett AF, Long AD, Gaut BS (2012) The molecular diversity of adaptive convergence. Science 335:457–461

Tito CM, Poggio L, Naranjo CA (1991) Cytogenetic studies in the genus Zea. 3. DNA content and heterochromatin in species and hybrids. Theor Appl Genet 83:58–64

Vitte C, Bennetzen J (2006) Analysis of retrotransposon structural diversity uncovers properties and propensities in angiosperm genome evolution. Proc Natl Acad Sci USA 103:17638–17643

Vitte C, Panaud O (2003) Formation of solo-LTRs through unequal homologous recombination counterbalances amplifications of LTR retrotransposons in rice *Oryza sativa* L. Mol Biol Evol 20:528–540

Wang Q, Dooner H (2006) Remarkable variation in maize genome structure inferred from haplotype diversity at the bz locus. Proc Natl Acad Sci USA 103:17644–17649

Whitney KD, Baack EJ, Hamrick JL, Godt MJW, Barringer BC, Bennett MD, Eckert CG, Goodwillie C, Kalisz S, Leitch IJ, Ross-Ibarra J (2010) A role for nonadaptive processes in plant genome size evolution? Evolution 64:2097–2109

Wright SI, Le QH, Schoen DJ, Bureau TE (2001) Population dynamics of an Ac-like transposable element in self- and cross-pollinating arabidopsis. Genetics 158:1279–1288

Xu JJ, Zhao QA, Du PN, Xu CW, Wang BH, Feng Q, Liu QQ, Tang SZ, Gu MH, Han B, Liang GH (2010) Developing high throughput genotyped chromosome segment substitution lines based on population whole-genome re-sequencing in rice (*Oryza sativa* L.). BMC Genomics 11:656

Zhang X (2008) The epigenetic landscape of plants. Science 320(5875):489–492

Zilberman D, Henikoff S (2007) Genome-wide analysis of DNA methylation patterns. Development 134:3959–3965