2015

Mechanisms and impact of alternative transposition-induced segmental duplications

Tao Zuo
Iowa State University

Follow this and additional works at: http://lib.dr.iastate.edu/etd
Part of the Bioinformatics Commons, and the Genetics Commons

Recommended Citation

This Dissertation is brought to you for free and open access by the Graduate College at Digital Repository @ Iowa State University. It has been accepted for inclusion in Graduate Theses and Dissertations by an authorized administrator of Digital Repository @ Iowa State University. For more information, please contact digirep@iastate.edu.
Mechanisms and impact of alternative transposition-induced segmental duplications

by

Tao Zuo

A dissertation submitted to the graduate faculty
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Co-Majors: Genetics; Bioinformatics and Computational Biology

Program of Study Committee:
Thomas Peterson, Co-Major Professor
Dan Nettleton, Co-Major Professor
Nick Lauter
Roger Wise
Erik Vollbrecht

Iowa State University
Ames, Iowa
2015

Copyright © Tao Zuo, 2015. All rights reserved.
DEDICATION

With special love and thanks to my family:

my parents Xintai Zuo and Taiying Su,

my wife Jing Zhao,

and my daughter Qiaoxi Zuo.
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>v</td>
</tr>
<tr>
<td><strong>CHAPTER 1. GENERAL INTRODUCTION</strong></td>
<td>1</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>Dissertation Organization</td>
<td>9</td>
</tr>
<tr>
<td>References</td>
<td>10</td>
</tr>
<tr>
<td><strong>CHAPTER 2. GENERATION OF TANDEM DIRECT DUPLICATIONS BY REVERSED-ENDS</strong></td>
<td>16</td>
</tr>
<tr>
<td>Abstract</td>
<td>16</td>
</tr>
<tr>
<td>Introduction</td>
<td>17</td>
</tr>
<tr>
<td>Results</td>
<td>20</td>
</tr>
<tr>
<td>Discussion</td>
<td>26</td>
</tr>
<tr>
<td>Methods</td>
<td>30</td>
</tr>
<tr>
<td>References</td>
<td>33</td>
</tr>
<tr>
<td><strong>CHAPTER 3. TRANSPOSITION-MEDIATED DNA RE-REPLICATION IN MAIZE</strong></td>
<td>52</td>
</tr>
<tr>
<td>Abstract</td>
<td>52</td>
</tr>
<tr>
<td>Introduction</td>
<td>52</td>
</tr>
<tr>
<td>Results</td>
<td>54</td>
</tr>
<tr>
<td>Discussion</td>
<td>62</td>
</tr>
<tr>
<td>Material and methods</td>
<td>66</td>
</tr>
<tr>
<td>References</td>
<td>69</td>
</tr>
<tr>
<td><strong>CHAPTER 4. THE FORMATION OF DNA TRANSPONIN-INDUCED TANDEM DIRECT</strong></td>
<td>89</td>
</tr>
<tr>
<td>DUPLICATIONS AND THEIR OCCURRENCE ACROSS 22 PLANT SPECIES</td>
<td>89</td>
</tr>
<tr>
<td>Abstract</td>
<td>89</td>
</tr>
<tr>
<td>Introduction</td>
<td>90</td>
</tr>
<tr>
<td>Results</td>
<td>92</td>
</tr>
<tr>
<td>Discussion</td>
<td>98</td>
</tr>
<tr>
<td>Methods</td>
<td>102</td>
</tr>
<tr>
<td>References</td>
<td>103</td>
</tr>
</tbody>
</table>
CHAPTER 5. GENES AND SMALL RNA TRANSCRIPTS EXHIBIT DOSAGE-DEPENDENT EXPRESSION PATTERN IN MAIZE COPY-NUMBER ALTERATIONS ................................................................. 117

Abstract ........................................................................................................ 117
Introduction ..................................................................................................... 118
Results .............................................................................................................. 122
Discussion ....................................................................................................... 132
Methods .......................................................................................................... 139
References ...................................................................................................... 144

CHAPTER 6. GENERAL CONCLUSIONS ............................................. 167

Conclusions ................................................................................................... 167
References ..................................................................................................... 171

ACKNOWLEDGEMENTS ............................................................................. 174
ABSTRACT

The mechanism of standard transposition involves two termini of the same transposon, distributing the transposon around the genome. Alternative transposition specifies a transposition mechanism engaging the ends of different transposons, which could induce chromosome breakage and various chromosome rearrangements, including deletions, inverted duplications, inversions and translocations. However, the formation of direct duplications by alternative transposition is still untested, and the genetic impact of the induced rearrangements remains largely unknown. A variety of genetic, molecular and bioinformatics tools have been applied in my thesis study to gain a better understanding of the mechanism and genetic impact of alternative transposition-induced duplications.

Reversed-ends transposition (RET) of Ac/Ds elements can generate deletion, inversion and translocation, but whether RET can lead to duplication is not yet studied. In this study, we predicted that RET involving two sister chromatids may produce reciprocal deletions and duplications. We isolated two twin sectors that carry reciprocal deletions and direct duplications, and also characterized 7 additional direct duplications. PCR and Southern experiments confirmed the structures of the duplication and deletion alleles. Sequence analysis of the breakpoint junctions supported the origin of duplication/deletion from RET mechanism. These results extend the significance of DNA transposons in reconstructing genome structures.

Transposition of Ac/Ds transposons often occurs during or shortly after DNA synthesis. The mechanism to increase DNA transposon copy number commonly occurs by excision from a replicated position and then insertion into an un-replicated target. Similarly, the target of alternative transposition is likely not replicated yet, causing more complex genome reorganizations. We found that RET in un-replicated target can induce a second round replication
of transposons and their flanking sequence. The size of re-replicated segment varies from 4.5 kb to 23 kb, depending on the timing of the abortion of re-replication. The re-replicated segment would form a composite insertion, including both transposon sequences and flanking sequences, and be placed into a new site in the genome. This work significantly expands our current knowledge regarding the contribution of DNA transposons to rapid genome expansion.

To investigate the evolutionary impact of the RET mechanism, we developed a bioinformatics program (*STRAND: Search for Transposon-Induced Tandem Direct Duplications*) to detect DNA transposon-associated tandem direct duplications. We applied the *STRAND* pipeline to 22 plant species with genome sequence available. The results show that other super families of DNA transposon including *Tc1/Mariner*, *Mutator* and *PIF/Harbinger* can undergo RET to induce tandem direct duplications. We identified 62 transposon-associated tandem duplications in 11 out of 22 plant species examined. A significant number (29) were derived from RET, and only 11 out of 62 were produced by non-allelic homologous recombination (NAHR) between two transposons, suggesting that RET was more frequent than NAHR in inducing tandem duplications mediated by DNA transposon. This study suggests that RET has played a significant role in genome evolution.

Finally, we are interested in understanding the immediate impact of duplications. Most duplications in current eukaryotic genomes are relatively ancient. By taking advantage of alternative transposition, we isolated a number of recent duplications, and conducted phenotypic and expression analyses on one case (*p1-ww714*), which contains an inverted duplication on chromosome 1S. The region duplicated in *p1-ww714* is 14.6 Mb in size and is predicted to contain approximately 300 gene models. Plants homozygous for *p1-ww714* (i.e., four copies of the duplicated region) are significantly shorter and have smaller ears than normal siblings.
Whereas, heterozygous plants ($p1$-ww714/normal; three copies of the duplicated region) are intermediate in height and ear size, suggesting that the segmental duplication in $p1$-ww714 exerts a dosage-dependent effect on phenotype. We implemented both GeneChip (new Affymetrix Maize WT 100K array) and high throughput sequencing (RNA-Seq) approaches to study the relationship between copy number variation and transcript accumulation. Results show that both genes and small RNA transcripts exhibit dosage-dependent expression pattern. This research provides insight into the transcriptional expression and phenotypic effect of a specific and recent maize duplication.
CHAPTER 1. GENERAL INTRODUCTION

Introduction

Structural variations have been revealed to be widespread in eukaryotic genomes, and have played critical roles in genome evolution, phenotypic variations, and environmental adaption. There are many studies showing that non-allelic homologous recombination (NAHR), non-homologous end joining (NHEJ), fork stalling and template switching (FoSTeS), and retro-transposition are the major mechanisms to generate genomic structural variations (Hastings et al. 2009; Zhang et al. 2009). However, a recent report suggested that transposable elements (TE) are also an important source of structural variations in the cucumber genome (Zhang et al. 2015). Additionally, research in maize revealed that duplicates in maize genome are often precisely bounded by transposons (Wicker et al. 2010). These observations suggest that TEs might play a larger role in generating structural variations than previously thought.

Introduction to Transposable Elements

A transposable element (TE) is defined as a mobile element that can transpose from one location of the genome to another location. It was first discovered by Barbara McClintock in the 1940’s (McClintock 1948; McClintock 1950), and since then has been revealed as a major component of many eukaryotic genomes; e.g. about 45% of the human genome (Lander et al. 2001), and 85% of the maize genome is composed of TEs (Schnable et al. 2009). TEs can be further classified into two sub-classes, Class I TE and Class II TE, based on the mechanism of transposition. Class I TEs are often called retrotransposons, which use a RNA-mediated copy-and-paste transposition mechanism to duplicate their copies, whereas Class II TEs are DNA transposons that utilize a DNA-mediated cut-and-paste mechanism to change their positions.
The maize *Ac* (*Activator*) and *Ds* (*Dissociation*) is a two-element DNA transposable element system first discovered by Barbara McClintock when she was studying the source of maize chromosome breakage (McClintock 1948; McClintock 1950; McClintock 1956). She found that the *Ac* element is an autonomous transposon that can perform transposition by itself, whereas *Ds* is a non-autonomous transposon that requires the presence of the *Ac* element for transposition. Molecular experiments conducted in the 1980s revealed that the *Ac* element is 4565 bp in size, and contains 11-bp terminal inverted repeats (TIRs) and approximately 240-bp 5' and 3' sub-terminal regions (Sutton et al. 1984). The mRNA transcribed from *Ac* is 3.5 kb long, which encodes a transposase (Tpase) composed of 807 amino acids (Kunze et al. 1987). The crystal structure of *Ac* protein is still unknown, but the structure of the related *Hermes* transposon from *Musca domestica* has recently been determined (Hickman et al. 2014). The transposases form an octameric ring to bind the TE by recognizing binding sites located on the TIRs and sub-terminal region, nick the transposon ends, form a hairpin structure, and then transfer the TE to a new target (Hickman et al. 2014). Following transposon excision, a footprint characterized by few basepair changes (Rinehart et al. 1997) is commonly left at the *Ac/Ds* donor site, and 8-bp target site duplications (TSD) will be generated flanking *Ac/Ds* at the new position.

*Ac*'s Negative dosage effect

The negative dosage effect of *Ac* is a phenomenon described as a negative correlation between *Ac* copy number and its transposition activity. It was first described by Barbara McClintock when she studied the effects of dosage of *Ac* in 1948. She found that an increase in *Ac* copy number in maize endosperm resulted in delayed development of *Ds* mutation and decreased transposition frequency (McClintock 1948). This phenomena can be easily observed
by performing a testcross of \textit{Ac} to the maize \textit{r1-sc:m3} reporter allele; one copy of \textit{Ac} in the maize endosperm results in coarsely spotted phenotype in maize aleurone, whereas two or more copies of \textit{Ac} result in a fine spotted aleurone (Conrad and Brutnell 2005).

The negative dosage effect of \textit{Ac} might reflect a means to self regulate its own activity, although the underlying mechanism is still unknown. A study of an \textit{Ac} dosage series revealed that \textit{Ac} transcript and protein levels were both increased in proportion to \textit{Ac} copy (Kunze et al. 1987), suggesting that a post-translational mechanism is likely responsible for regulating \textit{Ac} activity (Kunze and Starlinger 1989). However, the \textit{Ac} dosage response is not always negative; it could exhibit a positive dosage effect on certain \textit{Ac} alleles (Heinlein 1996; Brutnell et al. 1997). The observed positive and negative dosage effect of \textit{Ac} activity indicates a dynamic expressional pattern of \textit{Ac} during development, and suggests that a certain level of transposase accumulation is required to perform transposition (Heinlein 1996; Brutnell et al. 1997).

**Alternative transposition-induced genome rearrangements in maize**

The standard transposition of \textit{Ac/Ds} elements involves the 5' and 3' termini of the same \textit{Ac/Ds} element, resulting in a simple excision and reinsertion of the transposon (Figure 1A). However, this mechanism alone is not enough to explain the frequent occurrence of chromosome breakage observed at, for example, the original maize \textit{Dissociation} (\textit{Ds}) locus (McClintock 1950). In addition to chromosome breakage induced at the \textit{Ds} locus, McClintock also recognized various genome reorganizations including deletions, inversions, duplications and reciprocal translocations. Subsequent analysis revealed that the original chromosome-breaking \textit{doubleDs} element contains two copies of a simple \textit{Ds} element, with one \textit{Ds} inserted into the other in an opposite direction (Doring et al. 1984; MartinezFerez and Dooner 1997). Further molecular
studies indicated that chromosome breakage induced by *doubleDs* elements is caused by a transposition event that involves the termini of two *Ds* elements in different sister chromatids (English et al. 1993; English et al. 1995). This alternative transposition mechanism is termed Sister Chromatid Transposition (SCT) (Figure 1B). Besides the *doubleDs* structure, certain configurations of nearby *Ac/Ds* elements can also induce chromosome breakage. Weil and Wessler discovered that two *Ds* elements separated by about 1 kb in opposite orientations at the maize *Waxy* locus could also induce chromosome breakage, and the frequency of breakage is dependent on the distance between the two *Ds* elements (Weil and Wessler 1993). In 1999, Zhang and Peterson isolated and characterized stable alleles of the maize *p1* gene derived from twin sectors produced by SCT. The detailed sequence signatures such as “footprints” and TSDs from twin alleles provided convincing evidence that reciprocal inverted duplications and deletions can be derived from single SCT events (Zhang and Peterson 1999). Further research showed that alternative transposition by *Ac/Ds* elements in opposite orientation can be used to efficiently generate interstitial deletions, with sizes ranging from 12 kb to a few Mb (Zhang and Peterson 2005).

Although two *Ac/Ds* elements in opposite orientation could induce chromosome breakage, deletions and inverted duplications, the SCT mechanism cannot fully explain the formation of other types of genome rearrangements such as inversions and translocations. Would two *Ac/Ds* elements located in direct orientation produce chromosome breakage and other genome rearrangements? To test this, Zhang and colleagues conducted genetic studies on *P1-ovov454*, *P1-rr910* and *P1-rr11* alleles. These three alleles each contain one intact *Ac* element and another fractured *Ac* element (*fAc*, lost 2526 bp of the *Ac 5'* sequence), with *Ac 5'* ends sitting in the reversed orientation of the 3' end of the *fAc* element. The distances between the *Ac*
5' terminus and the \( f_{Ac} \) 3' end in \( P1\text{-}ovov454 \), \( P1\text{-}rr910 \) and \( P1\text{-}rr11 \) are 823 bp, 8919 bp and 13,175 bp, respectively. Within this configuration, transposase could bind the termini from two different transposons but in the same chromatid to undergo alternative transposition (Figure 1C).

They found that \( Ac/Ds \) elements in the reversed end configuration could also generate chromosome breakage and rearrangements, including deletions, inversions and reciprocal translocations (Zhang and Peterson. 2004; Zhang et al. 2009; Yu et al. 2010; Yu et al. 2011).

Since this alternative transposition involves two reversed ends from the same chromatid, it is termed reversed-ends transposition (RET).

![Figure 1](image)

**Figure 1** Three different configurations of the \( Ac \) element(s). The open and solid arrows represent 5' and 3' end of \( Ac \), respectively. Each solid green circle indicates binding to an \( Ac \) end sequence by the transposase produced by \( Ac \). A) A single transposon undergoes standard transposition. B) Transposasate(s) bind to termini from two transposons in different sister chromatids to perform sister chromatid transposition. C) Transposase binds to the ends of two transposons in the same chromatid to perform reversed-ends transposition.
Alternative transposition in other species

The study of alternative transposition mechanism suggests that similar mechanisms could occur with other types of DNA transposons in various species. The study from English and colleagues provided evidence that doubleDs and related structures are capable of inducing chromosome breakage in tobacco (English et al. 1993). Similar genome rearrangements can also be produced by alternative transposition in other plants, such as Arabidopsis (Krishnaswamy et al. 2008), and rice (Xuan et al. 2011; Yu et al. 2012). In bacteria, composite structures of Tn5 and ISShai transposons have been observed to create deletions and inversions via a similar transposition mechanism (Goryshin et al. 2003; Watanabe et al. 2007). A hybrid element formed by two copies of P-element from sister chromatids or homologous chromosomes in Drosophila can undergo alternative transposition, leading to various genome rearrangements (Gray et al. 1996; Preston et al. 1996; Parks et al. 2004). Accordingly, alternative transposition of DNA transposons might have played an important role in the history of genome evolution.

The maize p1 gene, p1 alleles and p2 gene

The maize p1 gene is frequently used as a visible selection marker while studying the mechanism of Ac/Ds transposon-induced genome rearrangements. It contains three exons, and encodes an R2R3 Myb-type transcription activator that controls the biosynthetic pathway of red phlobaphene pigment in various maize tissues, such as kernel pericarp and cob glume (Grotewold et al. 1994). The names of p1 alleles are usually followed by a two-letter suffix, which is used to indicate the expression of p1 in the pericarp and cob glumes. For example, P1-rr indicates a P1 allele that can express in pericarp and cob glumes, specifying red pericarp and red cob; p1-ww indicates white pericarp and white cob; P1-wr represents white pericarp and red
cob; and $P1$-*ovov* specifies orange-variegated pericarp and cob. In this study, $P1$-*ovov454* was used as the progenitor allele to generate and screen genomic rearrangements induced by alternative $Ac/Ds$ transposition.

The maize $p2$ gene is a paralog of the $p1$ gene; more than 95% of the $p2$ coding sequences are the same as that in $p1$. $p2$ encodes the same Myb-like transcription factor, but its transcripts are mainly detected in silk and anther, but not pericarp and cob glumes. The $p1$ and $p2$ genes originated from a segmental duplication of an ancestral $P$ gene approximately 2.75 million years ago (Zhang et al. 2000). Subsequent retrotransposon insertions separated $p1$ and $p2$ by about 70 kb, and result in differential expression patterns (Zhang et al. 2000). Molecular studies indicate that the solo expression of $p1$ in pericarp tissue is mainly caused by enhancer elements that are present only in the $p1$ gene (Sidorenko et al. 1999; Sidorenko et al. 2000). Interestingly, the function of $p2$ can be reactivated in pericarp and cob glumes via fusion with $p1$ derived by alternative transposition events (Zhang et al. 2006).

**Introduction to genome variations**

Recent high throughput genomic analyses via array-based comparative genomic hybridization (CGH) and sequence-based whole genome re-sequencing have begun to reveal significant levels of structural variations between individuals in a great many genomes. Copy number variation (CNV), one kind of structural variations with the size ranging from 1 kilobase (kb) to 1 megabase (Mb), has been demonstrated to be widespread across a large number of species including mammals and plants. In human, it is reported that more than 10% of its genome consists of CNVs (Iafrate et al. 2004; Sebat et al. 2004; Redon et al. 2006; Stankiewicz and Lupski 2010). In plants, a large proportion of CNV has also been revealed in *Arabidopsis*

The increased understanding of the universality of CNV in eukaryotic genomes has brought more attention on investigating the functional impact of CNV. Recent studies found that the phenotypic impact of CNV can be beneficial, neutral or detrimental. The human 1q21.1 locus is a genomic region shown to be associated with developmental abnormalities (O'Donovan et al. 2008). Changes in copy number of the 1q21.1 locus will greatly increase the risk of mental disorders (Stefansson et al. 2008; Stefansson et al. 2009). Similarly, gain or loss of one copy of the human 16p11.2 locus has an association with autism (Weiss et al. 2008). The phenotypic impact of CNV in several plant species has been disclosed as well. For example, a recent study in wheat showed that duplicated copies of *Rht-D1b* derived from a tandem duplication (>1 Mb) had a significant effect on reducing plant height (Li et al. 2012); also in wheat, copy number variation of two genes (*Ppd-B1* and *Vrn-A1*) contributed to wheat adaptation through altering the timing of flowering (Diaz et al. 2012). In soybean, a haplotype that has been used extensively for pathogen resistance contains 10 tandem copies of a 31 kb segment (Cook et al. 2012).

Even though much progress has been made in understanding the functional impacts of copy-number alterations, the underlying molecular mechanisms that lead to phenotypic differences are still not clear. Two models have been proposed to explain the relation between gene copy number and phenotypes (Tang and Amon 2013). The gene-specific effects model states that phenotypic differences are caused by changes in expression of a few specific genes, such as the cases described in human (Golzio et al. 2012) and plants (Cook et al. 2013; Li et al.
2012), whereas, the general effects model invokes the cumulative effects of many genes with altered dosage. Studying the response of gene expression to copy-number change might help to determine the mechanism. The usual responses to dosage change are: 1) no change on expression (dosage compensation); 2) proportional change of expression (dosage effect). In animals, autosomal genes predominantly exhibit dosage effect, while sex chromosomal genes are subject to dosage compensation (Kahlem et al. 2004; Vacik et al. 2005; Henrichsen et al. 2009; Pavelka et al. 2010; Stingele et al. 2012). In plants, dosage effect is more common for a small segment (<100 kb), but large duplications often exhibit dosage compensation (Birchler 1979; Guo and Birchler 1994; Huettel et al. 2008; Miclaus et al. 2011), suggesting the size of a duplication might affect the regulatory response.

**Dissertation Organization**

This dissertation contains six chapters, including general introduction (CHAPTER 1), four research papers (CHAPTER 2, 3, 4 and 5), and general summary (CHAPTER 6). CHAPTER 1 provides some basic background materials and presents the questions addressed in this research. Among the four research papers, CHAPTER 2 and CHAPTER 3 present new alternative transposition mechanisms that result in tandem duplications and composite insertions. In CHAPTER 2, a reversed-ends transposition (RET) model was proposed to explain the formation of reciprocal tandem direct duplication and deletions identified from the maize *p1-ovov454* allele. Molecular studies provide sufficient evidence to support the conclusion that RET of *Ac/Ds* elements can generate tandem direct duplications. This is the first report describing the mechanism of producing tandem direct duplications by transposition. CHAPTER 3 demonstrates another novel mechanism in which RET during DNA synthesis can induce a second round of
replication of the transposons and their adjacent sequences. The direct outcome of this transposition-mediated DNA re-replication is the generation of a composite insertion that includes the transposon and flanking sequences. CHAPTER 4 studies the evolutionary impact of transposon-induced duplications. A bioinformatics program was developed to identify DNA transposon-associated tandem direct duplications across 22 plant species. Further data analysis shows that the RET mechanism has played an important role in genome evolution. CHAPTER 5 investigates the phenotypic and transcriptional impact of recently-generated segmental duplications. A 14.6 Mb duplication in maize chromosome 1 has dramatic impact on plant height, ear length and flowering time. Expression analysis using GeneChip and RNA-sequencing reveal a dosage-dependent expression pattern of genes and small RNA transcripts in dosage series. The elucidation of the dosage-dependent response of small RNA transcripts to dosage change is completely novel. Finally, CHAPTER 6 summarizes the major findings and conclusions of this thesis project.

References


MartinezFerez I, and Dooner H. 1997. Sesqui-Ds, the chromosome-breaking insertion at *bz-m1*, links double *Ds* to the original *Ds* element. *Molecular and General Genetics* 255: 580-586.


CHAPTER 2. GENERATION OF TANDEM DIRECT DUPLICATIONS BY REVERSED-ENDS TRANSPOSITION OF MAIZE Ac ELEMENTS


Jianbo Zhang*, Tao Zuo* and Thomas Peterson

Department of Genetics, Development and Cell Biology, and Department of Agronomy,

Iowa State University, Ames, Iowa 50011

*These authors contributed equally to this work

Abstract

Tandem direct duplications are a common feature of the genomes of eukaryotes ranging from yeast to human, where they comprise a significant fraction of copy number variations. The prevailing model for the formation of tandem direct duplications is non-allelic homologous recombination (NAHR). Here we report the isolation of a series of duplications and reciprocal deletions isolated *de novo* from a maize allele containing two Class II Ac/Ds transposons. The duplication/deletion structures suggest that they were generated by alternative transposition reactions involving the termini of two nearby transposable elements. The deletion/duplication breakpoint junctions contain 8 bp target site duplications characteristic of *Ac/Ds* transposition events, confirming their formation directly by an alternative transposition mechanism. Tandem direct duplications and reciprocal deletions were generated at a relatively high frequency (~0.5 to 1%) in the materials examined here in which transposons are positioned nearby each other in appropriate orientation; frequencies would likely be much lower in other genotypes. To test whether this mechanism may have contributed to maize genome evolution, we analyzed sequences flanking *Ac/Ds* and other hAT family transposons and identified three small tandem
direct duplications with the structural features predicted by the alternative transposition mechanism. Together these results show that some class II transposons are capable of directly inducing tandem sequence duplications, and that this activity has contributed to the evolution of the maize genome.

**Author Summary**

The recent explosion of genome sequence data has greatly increased the need to understand the forces that shape eukaryotic genomes. A common feature of higher plant genomes is the presence of large numbers of duplications, often occurring as tandem repeats of thousands of base pairs. Despite the importance of gene duplications in evolution and disease, the precise mechanism(s) that generate tandem duplications are still unclear. In this study we identified nine new spontaneous duplications that arose flanking elements of the Ac transposon system. These duplications range in size from 8 kb to >5,000 kb, and all cases exhibit features characteristic of Ac transposition. Using similar criteria in a bioinformatics search, we identified three smaller duplications adjacent to other hAT family transposons in the maize B73 reference genome sequence. Our results show that transposable elements can directly generate tandem duplications via alternative transposition, and that this mechanism is responsible for at least some of the duplications present in the maize B73 genome. This work extends the significance of Barbara McClintock’s discovery of transposable elements by demonstrating how they can act as agents of genome expansion.

**Introduction**

In addition to generating additional copies of coding sequences that can be used as substrates for gene evolution (Ohno 1970), gene duplication may also cause immediate
phenotypic impacts such as human disease (Zhang et al. 2009b). Segmental duplications (SD)—two or more chromosomal segments with high homology—are common in higher plant and animal genomes. In humans and mice, ~5% of the genome is composed of segmental duplications (≥90% in identity and ≥1 kb in length); tandem duplications (direct and inverted) account for 35.2% and 21.6% of the total duplications in the mice and human genomes, respectively (Bailey et al. 2002; She et al. 2008). Many plants contain an even higher percentage of duplicated sequences. In rice, segmental duplications comprise 15–62% of the genome, depending on the sequences compared and classification criteria employed (Vandepoele et al. 2003; Paterson et al. 2004; Wang et al. 2005; Lin et al. 2006). Moreover, ca. 29% of rice genes are arranged in tandem repeats (International-Rice-Genome-Sequencing-Project 2005). Recent studies by others have also confirmed the presence of numerous duplicated sequences in the maize genome (Ahn et al. 1993; Moore et al. 1995; Gaut 2001; Odland et al. 2006; Schnable et al. 2009). Comparison of genome sequences from different individuals of the same species revealed that copy number variation (CNV) is widespread, and that tandem duplications account for a significant proportion of the observed CNV. In Arabidopsis and maize, more than 50% of CNV segments contain tandem duplications (Cannon et al. 2004; Springer et al. 2009; DeBolt 2010). In cattle and mice, copy number “gain” CNVs are predominantly associated with tandem local duplications, rather than interspersed duplications (Hou et al. 2011). These observations indicate that CNVs and associated tandem duplications are contributing to rapid genome evolution.

There are several mechanisms proposed to generate tandem duplications, including 1) non-allelic homologous recombination (NAHR) between short repeats flanking a DNA segment (Hastings et al. 2009b; Zhang et al. 2009a); 2) break-induced replication (BIR) (Morrow et al.
1997; Kraus et al. 2001) which can be mediated by short microhomology regions (Hastings et al. 2009a; Hastings et al. 2009b); and 3) fork stalling and template switching (FoSTeS) (Lee et al. 2007). Here, we investigated the potential role of Class 2 transposable elements in directly generating tandem sequence duplications via aberrant transposition reactions.

The standard model for transposition of DNA elements involves excision of the termini of a single transposon from a donor locus and reinsertion into a target site; the net effect is the movement of the element, without any other changes to the genome. In contrast, Alternative Transposition (AT) events involve the termini of two separate, usually nearby elements. AT reactions can generate a variety of genome rearrangements; for example, the Drosophila P element system can undergo Hybrid Element Insertion (HEI) events that produce a wide array of flanking rearrangements (Gray et al. 1996; Preston et al. 1996; Gray 2000). In maize, the Ac/Ds transposable element system is known to undergo at least two types of AT events that lead to genome rearrangements. First, Sister Chromatid Transposition (SCT) involves the directly-oriented 5' and 3' termini of closely-linked elements located on sister chromatids. Depending on the location of the transposition target site, SCT can generate chromatid bridges and breaks (Weil and Wessler 1993; Yu et al. 2010), as well as flanking inverted duplications and deletions (Zhang and Peterson 1999). Second, Reversed Ends Transposition (RET) involves the reversely-oriented 5' and 3' termini of two elements located nearby each other on the same chromatid. In addition to bridges and breaks (Yu et al. 2010), RET can generate flanking inversions, deletions, permutations, and reciprocal translocations (Zhang and Peterson 2004; Huang and Dooner 2008). An additional type of AT event termed Single Chromatid Transposition (SLCT) which involves the directly-oriented 5' and 3' termini of nearby elements on the same chromatid has been
observed in transgenic rice containing maize Ac/Ds elements, but this reaction was not detected in maize (Xuan et al. 2011).

We predicted that RET may also generate tandem direct duplications. Here we show that a single pair of reversed Ac termini induced a series of nine flanking tandem duplications ranging in size from 8157 bp to ~5.3 Mb. The structures of these tandem duplications and their associated deletions strongly indicate that they were indeed generated by reversed Ac ends transposition. Moreover, we identified three tandem duplications with the features predicted by RET flanking other hAT transposons in the maize B73 reference genome sequence.

Results

Identification of duplication candidates from maize twin sectors

To detect newly-formed duplications, we screened maize materials that contain elements of the Ac/Ds transposon system inserted into the p1 gene that controls kernel pericarp (seed coat) pigmentation. We initiated the screen with the progenitor allele P1-ovov454, which carries a pair of reversely-oriented Ac termini in the p1 gene intron 2 (Figure 1A). If transposition of the reversed Ac ends occurs during DNA replication and the excised termini insert into the sister chromatid, two unequal chromatids can be generated: one chromatid contains a tandem direct duplication, and the other contains a corresponding deletion (Figure 1D, lower and upper chromatids, respectively; for animated version please see Movie S1). These two chromatids will segregate into two adjacent daughter cells at mitosis; further mitotic divisions could generate a visible twinned sector. The new mutant chromosomes can be transmitted through meiosis to the kernels within the sectors and subsequently propagated as heritable alleles. Because the P1-ovov454 allele specifies orange variegated pericarp and orange variegated cob, both gains and losses of p1 expression can be recognized. The sector containing the deletion chromosome
(white twin, \textit{p1-ww-Twin}) would have white (colorless) pericarp due to loss of \textit{p1} gene exons 1 and 2, while the sector with the duplication chromosome (red twin, \textit{P1-rr-Twin}) would contain two copies of \textit{Ac} and exhibit fewer red and white stripes due to the negative \textit{Ac} dosage effect (McClintock 1948; McClintock 1951) (see Methods for details). We screened \textasciitilde2000 \textit{P1-ovov454/p1-ww} ears and identified six ears with this type of twinned sector. Two such ears which gave rise to duplication alleles \textit{P1-rr-T1} and \textit{P1-rr-T481} are shown in Figure 2; the remaining four twin sector ears gave rise to more complex rearrangements which are still under investigation.

**The red twin carries a tandem direct duplication**

The RET model (Figure 1) predicts that the breakpoints of the duplication alleles (sequence \textit{a} in Figure 1\textit{D}) should be adjacent to \textit{Ac} and \textit{p1} sequences. Therefore we used \textit{Ac} casting (Singh et al. 2003; Zhang et al. 2009c) and inverse PCR to isolate the sequences at the junction of the two duplication segments (Text S1). Comparison with the maize B73 genome sequence (Release 5b.60) (Schnable et al. 2009) indicates that the breakpoints in \textit{P1-rr-T1} and \textit{P1-rr-T481} are located \textasciitilde460 kb and \textasciitilde5.3 Mb proximal to \textit{p1}, respectively. For each allele we designed two new primers (1 and 2, Figure 1) flanking the predicted insertion sites and used these in PCR together with \textit{Ac}-specific primer Ac5. Primers 1+2 amplified products containing the intact insertion sites, and primers 1+Ac5 amplified the duplication junctions of sequence \textit{a} with 5’ \textit{Ac} (Figure 3); the results indicate that the breakpoint sequence is duplicated in both \textit{P1-rr-T1} and \textit{P1-rr-T481}. Previous semi-quantitative PCR analysis indicated that the \textit{p1} sequence proximal to \textit{Ac} is duplicated; hence these alleles carry duplications. To determine the relative orientations of the duplicated segments, we performed PCR with primers 1+3 which flank the
duplication junction of each allele. As shown in Figure 1D, primers 1 and 3 are separated by a 4565 bp Ac element at the duplication. By use of short PCR cycle times we could preferentially amplify products derived from somatic excision of Ac. PCR bands with sizes expected from Ac excision were amplified from both P1-rr-T1 and P1-rr-T481; sequencing of the PCR products shows that the sequence a of each breakpoint allele is linked to p1 gene sequences via a short footprint sequence typical of an Ac excision (Figure S1), and that the duplicated segments are in direct orientation as shown in Figure 1D. Together these results confirm the conclusion that P1-rr-T1 and P1-rr-T481 each carry a large segmental duplication of the sequence proximal to p1, in direct orientation.

The white twin p1-ww-T1 carries a reciprocal deletion

Another prediction of the RET model (Figure 1) is that the white twin alleles (p1-ww-T1 and p1-ww-T481) should each carry a deletion as the reciprocal product of their corresponding red duplication twins. To test this, PCR analysis was performed with primer pairs 2 + Ac3 and 1 + Ac5 which are specific for the predicted deletion and duplication junctions, respectively (Figure 4). Products of the expected sizes were amplified from p1-ww-T1 and P1-RR-T1 (Figure 4B). Importantly, sequencing of the PCR products showed that the 8 bp sequences immediately flanking the fAc 3’ end in p1-ww-T1 and the Ac 5’ end in P1-rr-T1 are identical, indicating their origin as a target site duplication (Figure 4C), the hallmark of Ac/Ds transposition. This result confirms that the twinned duplication/deletion alleles P1-rr-T1 and p1-ww-T1 originated as reciprocal products of a single reversed Ac ends transposition event.

We attempted to isolate the p1-ww-T481 allele, but none of the plants grown from the seven kernels within the white twin sector carried the expected deletion; all carried a standard p1-ww allele derived from the normal homologous chromosome. Because the duplication in the
corresponding red twin is 5.3 Mb, and a deletion of this size is most likely gametophyte lethal, we suspect that female gametophytes that received the deletion chromosome in meiosis had aborted and thus were not represented in the mature sector. This idea is consistent with the fact that the white sector contained fewer kernels than its red co-twin (*P1-rr-T481*; Figure 2). DNA gel blotting was conducted to further test the structures of the candidate duplication alleles (Figure 5). Genomic DNAs were digested with *Sal*I, and the blot was hybridized with *p*l-specific probe 15. The progenitor allele *P1-ovov454* shows three probe 15-hybridizing bands: a 5451 bp band containing *fAc*, a 2693 bp band located proximal to *Ac*, and a 1269 bp band which is present on both sides flanking *p*l and hence has a two-fold intensity on the blot. In the *P1-rr-T1* and *P1-rr-T481* samples, the 2693 bp band is twice the intensity of the 5451 bp band, consistent with a duplication of this proximal segment. In the *p1-ww-T1* lane the 2693 band is deleted, and the 5451 bp band is absent and has shifted to a new band of ~12 kb due to the deletion. An additional band of 1075 bp present in the *P1-ovov454* and *p1-ww-T1* lanes is derived from the *p1-ww* allele that is present in heterozygous condition in these samples (Figure 5).

**Isolation of additional duplication alleles**

As described above, the *P1-rr-T1* and *P1-rr-T481* duplication alleles were isolated from twin sectors with a pericarp phenotype distinct from the parental allele. Multikernel twin sectors are produced by transpositions that occur during a narrow window of early ear development and thus are relatively rare. Therefore we asked whether additional duplication alleles could be isolated from whole ears that exhibited a similar phenotype as that of the red co-twins (i.e. less red/white pericarp variegation). These whole-ear cases could have originated from reversed-ends transposition events that occurred either earlier in embryo development (such that the red
twin sector encompassed the entire ear), or as pre-meiotic events. Approximately ~80 ears of this type were identified among the ~2000 p1-ovov454/p1-ww ears screened. Plants grown from these whole-ear cases were analyzed by semi-quantitative PCR (Figure S2) to detect changes in copy number of the p1-proximal sequences. In this way we identified 13 additional candidate duplication alleles. The breakpoints of 11 duplication candidates were cloned via Ac casting or inverse PCR (iPCR); sequencing the PCR products revealed that the breakpoints were located at various sites up to 3.3 Mb proximal to the p1 gene on chromosome 1 (Text S1). Based on the breakpoint sequences and the maize genome sequences, new primers 1 and 2 specific for each candidate allele were designed and used in PCR together with Ac primer Ac5. The results of PCR using primers 1 + 2 + Ac5 (Figure S3) confirmed that seven of the 11 candidates carried tandem direct duplications ranging in size from 8157 bp to 3.3 Mb (Table 1). PCR using primers 1+3 flanking the presumed duplication breakpoint confirms that all of the seven alleles derived from whole ears contain tandem direct duplications. The structures of the other four alleles are more complex and are under further investigation.

These seven candidate duplication alleles were also subject to DNA gel blot analysis (Figure S4); the results show a higher relative intensity of the 2693 bp fragment in all of the candidate alleles except for P1-rr-E20, whose 8157 bp duplication does not extend into the 2693 bp fragment detected by the probe. Together the DNA gel blot results confirm the allele structures predicted from the duplication breakpoint sequences. The DNA gel blot results and semi-quantitative PCR indicated that P1-rr-E301 and P1-rr-E336 also contain duplications, but their breakpoints are not yet cloned.
Identification of Transposon-induced Duplications in the Maize Genome

The experiments described above identified nine tandem direct duplication alleles apparently generated *de novo* by RET of *Ac/Ds* elements. If this mechanism has contributed to genome evolution, one would expect to find evidence of transposon-induced duplications in the maize genome sequence. Therefore we conducted a bioinformatics search of the maize B73 reference genome for duplications with the structural features predicted by the RET model. First we identified sequences flanking known *hAT* family transposons and compared the flanking sequences to detect duplications; we then analyzed these candidate duplications for the sequence features predicted by the RET model. In total, 26 known maize *hAT* family transposons, including *Ac/Ds* element and 25 *dhAT* family elements identified in the lab of Dr. Jinsheng Lai, China Agricultural University (personal communication), were used to search for associated duplications in maize B73 reference genome (*ZmB73_RefGen_V2*). In this way, we identified three small duplicated segments (Figure 6) that have the sequence features predicted by the RET model (Figure 1). These three tandem duplications are associated with 3 different *dhAT* family elements, *dhAT*-Zm1, *dhAT*-Zm13 and *dhAT*-Zm24. The first duplication is located on chromosome 1 and contains two tandem direct repeats of 147 bp and 148 bp that are 93% identical. The duplicated segments are initiated by two *dhAT*-Zm1 elements with 95% sequence identity (Figure 6). The second duplication is located on chromosome 7 and contains two tandem direct repeats of 1262 bp and 1257 bp that are 96% identical. The duplicated segments are initiated by two *dhAT*-Zm13 elements with 95% sequence identity; one is intact (568 bp) and the other has a deletion of 12 bp from the 5’ TIR sequence (Figure 6). In both duplications, the first *dhAT* element is flanked by 8 bp direct repeats that represent the Target Site Duplications (TSDs) generated by *hAT* element insertion. Whereas, the second *hAT* element is flanked on one side by
the same TSD as the first element, but the other terminus does not have a matching TSD. This is exactly the structure predicted by the RET model (Figure 1) and observed in the Ac-induced duplications (Figure 4): the first transposon has TSDs derived from the original insertion of the transposon (pre-duplication); the second transposon copy has the same TSD on one end, but the other end has a non-matching flanking sequence because it represents the subsequent RET event that generated the duplication. The third duplication (on chromosome 6) has a somewhat different structure, but is still consistent with the predictions of the RET model. This case contains direct repeats of 116 bp and 118 bp that are 99% identical; these repeats are initiated by two fractured dhAT-Zm24 elements with 96% identity. The intact dhAT-Zm24 element is 904 bp long, whereas these fractured elements contain only 288 bp and 289 bp from the 3’ end. A duplication with these structural features could also be formed by a mechanism of RET as shown in Figure S5 (Movie S2).

Discussion

By taking advantage of a visual screen to identify chromosome rearrangements associated with Ac transposition events, we have isolated and characterized nine tandem duplications that arose de novo from a single progenitor allele. The endpoints of all nine duplications coincide precisely with Ac termini. Two duplications were isolated from phenotypic twinned sectors, and in one case we were able to recover and characterize a complementary deletion allele. Importantly, the endpoints of the twinned duplication/deletion alleles share a matching 8 bp TSD which is a hallmark of Ac transposition. These results indicate that the duplications originated through reversed Ac ends transpositions (RET) that occurred during or shortly after DNA replication; the excised Ac/fAc ends inserted into sites in
the sister chromatid, resulting in reciprocal chromatids, one containing a tandem direct duplication, and the other bearing a corresponding deletion (Figure 1). These structures are not consistent with origin via other mechanisms. BIR and FoSTeS generally do not produce a deletion and a reciprocal duplication in the same event (Hastings et al. 2009b). NAHR can generate a deletion and a reciprocal duplication. However, if these duplications were generated by NAHR between non-allelic Ac elements, then they should contain three copies of Ac (one Ac flanking the proximal and distal duplication endpoints, and one between the duplicated segments). All of the duplications we isolated lack an Ac element at one breakpoint. Although it is formally possible that one Ac element excised after the formation of the duplication, this can be excluded because the sequences at the junctions do not contain any evidence of an Ac excision footprint. Moreover, duplications generated via NAHR are recurrent; independent NAHR events between the same repeats generate duplications of the same size. However, our duplications share only one breakpoint in intron 2 of the p1 gene; the second breakpoint is different for each of the duplications, resulting in a set of nine overlapping duplications ranging in size from 8157 bp to ~5.3 Mb.

The Drosophila P element transposon can undergo various types of alternative transposition events that can produce a multitude of rearrangement structures, depending on which transposon termini are involved in the transposition reactions, and the location of the target site ((Gray 2000) for review). In the case of the maize Ac/Ds system, fewer types of alternative transposition can occur because the transposition competence of each Ac/Ds end is dependent on strand-specific hemi-methylation of the transposon TIR. The tandem duplications described here are entirely consistent with the RET model shown in Figure 1, and with the
known restriction on transposition competence of \( Ac/Ds \) elements (Ros and Kunze 2001; Peterson and Zhang 2013).

NAHR is reported to occur at a frequency of \( 10^{-5} \) to \( 10^{-6} \) in human (Lam and Jeffreys 2006; Lam and Jeffreys 2007; Turner et al. 2008); in \( Arabidopsis \), a frequency of \( 10^{-4} \) to \( 10^{-6} \) was observed for NAHR between two ~1.2 kb repeats separated by ~4.0 kb unique DNA sequence (Molinier et al. 2004). Rates of NAHR have not, to our knowledge, been reported for maize. Our results indicate that transposition-induced duplications can occur at a relatively high frequency, depending on the presence of an active transposon system with appropriately positioned elements. From a population of approximately 2000 plants, we identified seven whole ears and two twinned-sector ears with transposition-generated tandem direct duplications. DNA gel blotting and semi-quantitative PCR results indicate that two additional cases (\( P1\text{-}rr\text{-}E301 \) and \( P1\text{-}rr\text{-}E336 \); Figure S4) also carry duplications, although we could not clone their breakpoints. The calculated duplication frequency (~0.5%) is very likely an underestimate for two reasons. First, the visual phenotype used to detect duplications (darker red pericarp and fewer purple aleurone spots) is somewhat subtle and some events may have been overlooked. Second, the screen would not have detected distal duplications because these would not alter the \( pl \) gene or \( Ac \) dose. Distal duplications would result from insertion of the excised \( Ac/fAc \) termini into a site between the \( pl \) gene and telomere (Figure S5; Movie S2), and these would be expected to occur as frequently as proximal duplications. Thus the real frequency of duplications derived from the \( P1\text{-}ovov454 \) allele may be closer to 1%. Given this high frequency, we asked whether \( Ac/Ds \)-induced tandem duplications could be detected in the maize B73 genome, which contains ~50 \( Ac/Ds \) elements (Du et al. 2011). However, we failed to find any \( Ac/Ds \) copies adjacent to a tandem duplication, possibly because the \( Ac/Ds \) elements in the
B73 genome are widely separated, and efficient reversed-ends Ac/Ds transposition requires two elements in close proximity and correct orientation (Yu et al. 2010).

In addition to the Drosophila P element and Ac/Ds systems, the Antirhinnnum Tam3 element, a founding member of the hAT transposon superfamily, is known to induce flanking genome rearrangements (Martin et al. 1988; Martin and Lister 1989; Lister et al. 1993), possibly via alternative transposition mechanism(s). This suggested that other transposons, in particular hAT family elements, may be capable of undergoing alternative transposition to mediate genomic rearrangements. Therefore we extended our bioinformatics searches for transposon-associated tandem duplications to a set of 25 other hAT family elements previously identified in the maize B73 reference genome (personal communication). These searches returned a total of 7611 hAT element insertions, and among these we identified three tandem direct duplications with the key structural features predicted by the RET model: First, they have exactly two repeated copies, and each repeat is initiated precisely by the transposon. Moreover, in two of the duplications the first hAT element is flanked by 8 bp TSDs, while the second (middle) element is flanked by only one of these 8 bp sequences. These features are not expected from other duplication mechanisms such as NAHR, BIR and FoSTeS, but they are perfectly predicted by the RET model. Although the duplications observed are relatively short and their frequency is low, it is possible that some examples may not have been detected for various reasons. First, the maize B73 reference genome sequence still has numerous gaps and uncertainties in the order and orientations of many sequence contigs, and these ambiguities will interfere with the identification of duplications, especially larger ones. Second, those more recent and therefore nearly identical duplications may be under-represented in the reference sequence due to collapse during sequence assembly (Emrich et al. 2007; Phillippy et al. 2008). Third, those duplications in which either one of the
TEs excised after formation of the duplication would not be detected by our search criteria. Nevertheless, we conclude from these results that RET-induced tandem duplication has occurred in maize evolutionary history. Given the high frequency and diversity of Class II transposons present in many eukaryotic species, the impact of this mechanism in eukaryotic genome evolution may be significant. The RET model described here provides the conceptual basis for additional bioinformatics searches that will be necessary to assess the actual impact of this mechanism in different species.

Methods

Genetic Stocks

The maize p1 gene encodes a Myb-like transcription factor controlling the pigmentation of floral tissues, including kernel pericarp (seed coat) and cob. The suffix of a p1 allele indicates its expression pattern in pericarp and cob, e.g., P1-rr specifies red pericarp and red cob, p1-ww specifies white (colorless) pericarp white (colorless) cob, and P1-ovov specifies orange variegated pericarp (seed coat) and orange variegated cob. The numeral following the suffix indicates the origin of the allele; alleles with the same phenotype but different numeral may have different structures. The P1-ovov454 allele conditions a high frequency of colorless sectors, presumably resulting from alternative transposition events which interrupt or delete the p1 gene (Yu et al. 2011). The p1-ww-[4Co63] allele is from the maize inbred line 4Co63 (Goettel and Messing 2010); heterozygous plants of genotype P1-ovov454/p1-ww-[4Co63] were fertilized with pollen from plants of genotype C1, r1-m3::Ds [4Co63]. Ac induces excision of Ds from r1-m3::Ds, resulting in restoration of r1 gene function and purple aleurone sectors. Ac/Ds transposition is subject to the negative Ac dosage effect (McClintock 1948; McClintock 1951), in which increases in Ac copy number result in a developmental delay in Ac/Ds transposition. If
reversed Ac ends transposition occurs as shown in Figure 1, two different sister chromatids would be produced: one carrying a tandem direct duplication, and the other a reciprocal deletion (Figure 1D). These chromatids will separate into two adjacent daughter cells at mitosis, forming a twinned sector after successive rounds of cell division. The sector with the deletion chromosome has lost Ac and exons 1 and 2 of the pl gene, and thus should have colorless pericarp with no purple aleurone sectors. The sector with the duplication chromosome retains a functional Pl-ovo454 gene and two copies of Ac, and thus should exhibit fewer colorless pericarp sectors, and smaller kernel aleurone sectors.

**Genomic DNA extractions, Southern blot hybridization**

Total genomic DNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) extraction protocol (Porebski et al. 1997). Agarose gel electrophoresis and Southern hybridizations were performed according to Sambrook et al (Sambrook et al. 1989), except hybridization buffers contained 250 mM NaHPO4, pH 7.2, 7% SDS, and wash buffers contained 20 mM NaHPO4, pH 7.2, 1% SDS.

**PCR amplifications**

Sequences of oligonucleotide primers used in PCR reactions are given in Table 2; note that primers 1 and 2 are specific to each allele. PCR was performed using HotMaster Taq polymerase from 5 PRIME (Hamburg, Germany). Reactions were heated at 94 °C for 2 min, and then cycled 35 times at 94 °C for 20 s, 60 °C for 10 s, and 65 °C for 1 min per 1 kb length of expected PCR product, then 65 °C for 8 min. For difficult templates, 0.5-1 M betaine and 4%–8% DMSO were added. The band amplified was purified from an agarose gel and sequenced directly. Sequencing was done by the DNA Synthesis and Sequencing Facility, Iowa State
University, Ames, Iowa, United States. Ac casting and inverse PCR were performed as described previously [36].

**Bioinformatics**

The sequences of 26 hAT family transposable elements were used as queries to search for homologous elements in the maize B73 reference genome (ZmB73_RefGen_V2) via local BLASTN with default parameters. Two types of homologous sequences were identified: 1) intact elements, which contained both 5’ and 3’ termini; 2) fractured elements, which contained one terminal end (either 5’ or 3’) but having lengths greater than 100 bp. A PERL script was developed to extract two 100 bp segments flanking each transposon, one 5’ adjacent and one 3’ adjacent. Pair-wise comparisons were performed among the segments flanking the same terminal end within each individual hAT family. Two hAT family members with the same orientation, less than 100 kb apart, and with homologous sequences flanking one terminal end but not the other end were retained for further structural analysis. Such cases were examined manually for the following features: 1) the duplication comprises the complete sequence between the two hAT elements, and 2) the duplication is initiated by the transposable element insertion. Sequences that met the above criteria were considered as putative duplications generated by alternative transposition and were examined further for the presence of TSDs as described in the text.
Acknowledgements

We thank Dr. Jinsheng Lai (China Agricultural University, Beijing, China) and Jinliang Yang (Iowa State University, Ames, IA) who graciously provided the sequences of 25 unpublished dhAT elements. We thank Terry Olson for assistance in molecular biology, and Douglas Baker for field work.

Author Contributions

Conceived and designed the experiments: J.Z., T.Z., and T.P.; Performed the experiments: J.Z., T.Z.; Analyzed the data: J.Z., T.Z., and T.P.; Contributed new reagents/materials/analysis tools: J.Z., T.Z., and T.P.; Wrote the paper: J.Z., T.Z., and T.P.; Designed software for use in analysis: T.Z.;

References


Table 1. The lengths of the duplications

<table>
<thead>
<tr>
<th>Alleles</th>
<th>Distance</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1-rr-E20</td>
<td>8157 bp</td>
</tr>
<tr>
<td>P1-rr-E45</td>
<td>203 kb</td>
</tr>
<tr>
<td>P1-rr-E43</td>
<td>332 kb</td>
</tr>
<tr>
<td>P1-rr-T1</td>
<td>460 kb</td>
</tr>
<tr>
<td>P1-rr-E317</td>
<td>459 kb</td>
</tr>
<tr>
<td>P1-rr-E3</td>
<td>802 kb</td>
</tr>
<tr>
<td>P1-rr-E70</td>
<td>1.0 Mb</td>
</tr>
<tr>
<td>P1-rr-E10</td>
<td>3.3 Mb</td>
</tr>
<tr>
<td>P1-rr-T481</td>
<td>5.3 Mb</td>
</tr>
</tbody>
</table>

Table 2. PCR primers

<table>
<thead>
<tr>
<th>Primer 1</th>
<th>Primer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1-rr-E20</td>
<td>TAGATTTCCGTCTCTTGTTGTA</td>
</tr>
<tr>
<td>P1-rr-E45</td>
<td>GCAACACCTCAAAGACCACACGG</td>
</tr>
<tr>
<td>P1-rr-E43</td>
<td>GCGTAGCGACGTCTCCTCAC</td>
</tr>
<tr>
<td>P1-rr-T1</td>
<td>GCCGAGCGTTCCGTGATGCTGACTC</td>
</tr>
<tr>
<td>P1-rr-E317</td>
<td>GTGAAAATCGTGAGGAGTGTTG</td>
</tr>
<tr>
<td>P1-rr-E3</td>
<td>GTCCCACGCGCGCTCGAAGA</td>
</tr>
<tr>
<td>P1-rr-E70</td>
<td>GTTGCTAGTGGAAGTTCAGGCA</td>
</tr>
<tr>
<td>P1-rr-E10</td>
<td>CAGCAGTCATGCTGTCGGATG</td>
</tr>
<tr>
<td>P1-rr-T481</td>
<td>ACAGTAGAGAGCTTCGCTTACTTCTC</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Ac3</th>
<th>Primer 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>GATTACCGTATTTATCCCGTTTTCGTTTTTC</td>
<td></td>
</tr>
<tr>
<td>Ac5</td>
<td>GCTATCAAACAGACCAACGGAGAGAAT</td>
</tr>
<tr>
<td>Primer 3</td>
<td>CCCCCCGTTCCGGTTTTCGTTTTCG</td>
</tr>
</tbody>
</table>

Ac3
| Ac5               | GCTATCAAACAGACCAACGGAGAGAAT |
| Ac5               | CCCCCCGTTCCGGTTTTCGTTTTCG |
FIGURES and LEGENDS

A

B

C

D
**Figure 1. Reversed Ac ends transposition generates direct duplication.**

The two lines indicate sister chromatids of maize chromosome 1, joined at the centromere (black). The blue boxes are exons 3, 2, and 1 (left to right) of the *p1* gene. Red lines with arrowhead(s) indicate Ac/fAc insertions, and the open and solid arrowheads indicate the 3' and 5' ends, respectively, of *Ac/fAc*. The short horizontal arrows show the orientations and approximate positions of PCR primers, and the numbers below are the primer names. The green/black triangles indicate the transposon target site sequences and target site duplications.

(A) *Ac* transposase cleaves the lower chromatid at the 3’ end of *fAc* and the 5’ end of *Ac* (arrows).

(B) Following transposase cleavage, the internal *p1* genomic sequences are joined to form a circle. Dotted lines indicate the insertion of the *fAc* and *Ac* termini into the a/b site on the sister chromatid.

(C) Transposon ends insert into the upper sister chromatid at a proximal site.

(D) The *Ac* 5’ end joins to the distal side (green) of the target site and the *fAc* 3’ end joins to the proximal side (black) of the target site to generate a proximal deletion (upper chromatid) and a direct duplication (lower chromatid). The shaded arrows encompass the duplicated segments.

For animation, please see Movie S1.
Figure 2. Ears with twinned sectors T1 (left) and T481 (right).

The white and red phenotypic twinned sectors are outlined. The remainder of the ear has the orange-variegated phenotype specified by the progenitor *P1-ovov454* allele.

Figure 3. PCR analysis of the twinned alleles with primers 1+2+Ac5.

Lane 1: DNA ladder; lanes 2 and 5: water (negative control); lanes 3 and 4: *P1-ovov454, P1-rr-T1*; lanes 6 and 7: *P1-ovov454, P1-rr-T481*. Note that primers 1 and 2 are specific for each allele.
Figure 4. Breakpoint sequences of reciprocal duplication/deletion alleles $P1$-$rr$-$T1$ and $p1$-$ww$-$T1$ generated by Reversed Ends Transposition.

(A) Diagram shows the structure of the progenitor $P1$-$ovov454$ allele prior to RET. Two sister chromatids are shown, with symbols as in Figure 1. The dotted box shows the a/b target site region, whose sequence is indicated above. The color of the letters in the sequences matches the chromatid line color.


(C) Sister chromatid structures of $p1$-$ww$-$T1$ (upper) and $P1$-$rr$-$T1$ (lower). Sequences of the deletion and duplication breakpoints (dotted boxes) are shown in color matching the chromatid line color. Note that each breakpoint has a copy of the 8 bp TSD GCGCTTTA which is present in a single copy at the a/b target site in the progenitor allele.
Figure 5. DNA gel blot analysis of the twinned alleles.

(A) Structure of progenitor $P1$-$ovov454$ allele (upper), and predicted structures of the $P1$-$rr$-$Twin$ (duplication) and $p1$-$ww$-$Twin$ (deletion) alleles (lower) generated via reversed $Ac$ ends transposition. Blue lines and boxes indicate $p1$ sequences, green lines indicate sequences proximal to $p1$, and gray boxes indicate sequences homologous to probe 15. The short vertical black lines indicate $SalI$ sites, and red asterisks (*) mark methylated $SalI$ sites. The other symbols have the same meaning as in Figure 1.

(B) DNA gel blot. Genomic DNA was digested with $SalI$ and hybridized with $p1$ genomic probe 15 (gray boxes in Figure 5A). Lane 1: $p1$-$ww$[$4Co63$], Lane 2: $P1$-$ovov454$/p1-$ww$[4Co63], Lane 3: $P1$-$rr$-$T1$, Lane 4: $p1$-$ww$-$T1$/$p1$-$ww$[4Co63], Lane 5: $P1$-$rr$-$T481$
Figure 6. Tandem direct duplications in maize generated by Reversed-Ends Transposition.

Red lines with arrowhead(s) indicate the dhAT family elements associated with each duplication; solid and open arrowheads indicate the transposon 5’ and 3’ ends, respectively. The truncated solid arrowhead in dhAT-Zm13 indicates a deletion of 12 bp from the 5’ TIR. The blue lines represent duplicated segments. The blue triangles indicate the transposon target site duplications. Numbers above each line indicate the length of that segment. Sequences and genomic positions are shown in Text S2.
SUPPLEMENTAL FIGURES and LEGENDS

A

Somatic Ac excision

B

P1-rr-T1 (primer 3 was used for sequencing, the lower strand in Figure S1A was read in the chromatogram below)

Original sequence (both strands are shown)
5’ TTGCGCTGTATCATGCCGCCGCTTTA
3’ AACAGACTAGTAAAGCCGCCGCAAAAT
After Ac excision (only the lower strand is shown)
3’ qACGGGGAGACAGGTTAATTGCAAGGTTGCCAGCTAGCAGACGCGGGGTTGGG
5’ qACGGGGAGACAGGTTAATTGCAAGGTTGCCAGCTAGCAGACGCGGGGTTGGG

P1-rr-T481 (primer 1 was used for sequencing, the upper strand in Figure S1A was read in the chromatogram below)

Original sequence (both strands are shown)
5’ ATCTGGTCGGTACAGCCGAGTGCAGGTTGGAGTTGACC
3’ TAGACGAGACGGGGTGCTCGCCATTTGACGCACTACGTCTCACTCTG
After Ac excision (only the upper strand is shown, d = A or G or T)
5’ fGCCCCGGCCGCAATTGAGCTGCCAAGGCAAGGATCGTTATCGTT
5’ fGCCCCGGCCGCAATTGAGCTGCCAAGGCAAGGATCGTTATCGTT
Figure S1. Determining the orientation of duplications by PCR and sequence analysis.

(A) Excision of \( Ac \) from the duplication chromosome brings primers 1 and 3 into close proximity, enabling the amplification of PCR products containing the excision site and footprint (indicated by \( \times \)).

(B) Sequencing chromatograms of \( Ac \) excision products from \( P1-rr-T1 \) (upper) and \( P1-rr-T481 \) (lower). \( Ac \) excision is commonly accompanied by minor sequence changes (footprints); the most common \( Ac/Ds \) footprints in plants are small (1 – 2) base substitutions or deletions. Genomic DNA prepared from leaf tissue of a single plant may contain molecules from multiple independent somatic \( Ac \) excision events; these can be detected by direct sequencing of the PCR products (amplified using primers 1 and 3), resulting in multiple peaks beginning at the excision site as seen in the upper chromatogram. The sequences flanking the footprint are identical in different PCR products, hence it is often possible to infer the sequences of the individual major products. The signals giving rise to multiple peaks at each position are recorded in italic letters to aid in interpreting the figure. In \( P1-rr-T1 \), primer 3 was the sequencing primer, and two types of footprints were identified. In \( P1-rr-T481 \), primer 1 was the sequencing primer, and two types of footprints are indicated in italics. The presence of a third footprint species can be inferred from the observation of three distinct peaks (A, T and G) at position 81; this third species would be identical to one of the two sequences shown in italics, except for the base at position 81.
Figure S2. Semi-quantitative PCR of duplication alleles in comparison with parental *P1-ovov454*.

*(A)* Structure of *p1* (blue) and *p2* (green) genes in a tandem duplication allele. The *p2* gene is a paralog of *p1* located ~70 kb proximal to *p1* in *P1-ovov454* and its derivatives. The boxes indicate exons of *p1/p2*, and the blue and green bars under the gene structure indicate the approximate positions of the PCR products. Other symbols have the same meaning as in Figure 1.

*(B)* PCR gel. The primer pair used here amplifies a 343 bp band from *p1* and a 420 bp band from *p2*. By comparing the *p2/p1* product band intensities, the *p2/p1* gene ratio of each genotype was estimated and is indicated at the bottom (note that the smaller *p1* band appears to amplify somewhat more efficiently than the *p2* band, hence the product intensity ratio is not identical with the inferred gene copy ratio). The 4Co63 inbred line (lane 1) contains a *p2* gene but lacks *p1*; its *p2/p1* ratio is indicated as 2/0. The progenitor allele *P1-ovov454* and all its derivatives are heterozygous with 4Co63. For *P1-ovov454* (lane 2), the ratio of *p2/p1* is 2/1 (two copies of *p2*-one copy from 4Co63 and one copy from the *P1-ovov454* chromosome: one copy of *p1* from the *P1-ovov454* chromosome). Lanes 3-6 are *P1-rr-E43*, *P1-rr-E10*, *P1-rr-E20*, and *P1-rr-E70*, respectively. A *p2/p1* ratio of 2/2 (lane 5) indicates a duplication that does not include *p2* (only *p1* in the *P1-ovov454*-carrying chromosome was duplicated); whereas a 3/2 ratio (lanes 3, 4, 6) indicates a duplication that extends beyond *p2* (both *p1* and *p2* in the *P1-ovov454*-carrying chromosome were duplicated).
Figure S3. PCR analysis of progenitor P1-ovov454 and duplication alleles obtained from whole ears. PCR was performed using primers 1 + 2 + Ac5 (see Figure 1 for primer locations; note that primers 1 and 2 are specific for each allele). Primers 1 + 2 amplify the transposition target site (present in the progenitor P1-ovov454 and each duplication allele), while primers 1 + Ac5 amplify the duplication junction (present only in the duplication alleles). Lane 1: DNA ladder; lanes 2-4: H2O, P1-ovov454, P1-rr-T1; lanes 5-7: H2O, P1-ovov454, P1-rr-T481; lanes 8-10: H2O, P1-ovov454, P1-rr-E70; lanes 11-13: H2O, P1-ovov454, P1-rr-E3; lanes 14-16: H2O, P1-ovov454, P1-rr-E43; lanes 17-19: H2O, P1-ovov454, P1-rr-E45. Six of the nine alleles isolated from whole ears are analyzed here. The other three alleles were not included because they were heterozygous with 4Co63 which can produce a product of the same size as that in P1-ovov454 using primers 1 + 2.
Figure S4. DNA gel blot analysis of progenitor *P1-ovov454* and duplication alleles.

Genomic DNA was digested with *Sal*I and hybridized with genomic probe 15. See Figure 5 for allele structures and probe locations. Lane 1: *p1-ww[4Co63]*; Lane 2: *P1-ovov454/p1-ww[4Co63]*; Lane 3: *P1-rr-T1*; Lane 4: *p1-ww-T1/p1-ww[4Co63]*; Lane 5: *P1-rr-T481*; Lane 6: *P1-rr-E10/p1-ww[4Co63]*; Lane 7: *P1-rr-E70/p1-ww[4Co63]*; Lane 8: *P1-rr-E3/p1-ww[4Co63]*; Lane 9: *P1-rr-E317/p1-ww[4Co63]*; Lane 10: *P1-rr-E43/p1-ww[4Co63]*; Lane 11: *P1-rr-E45/p1-ww[4Co63]*; Lane 12: *P1-rr-E20/p1-ww[4Co63]*; Lane 13: *P1-rr-E336/p1-ww[4Co63]*; lane 14 *P1-rr-E301/p1-ww[4Co63]*
Figure S5. Reversed Ac ends transposition generates distal direct duplication. The two lines indicate sister chromatids joined at the centromere (black circle). All the symbols have the same meaning as in Figure 1 except the green line now indicates the pl-distal segment.

(A) Ac transposase cleaves the lower chromatid at the 3’ end of fAc and the 5’ end of Ac.

(B) Following transposase cleavage, the inter-transposon segment is joined to form a circle.

(C) Transposon ends insert into the sister chromatid at a distal site. The 3’ end of fAc joins to the green segment (b) to generate a distal direct duplication, while the 5’ Ac end joins to the black segment (a) to generate a distal deletion. Note that both chromatids carry one copy of Ac element.
CHAPTER 3. TRANSPOSITION-MEDIATED DNA RE-REPLICATION IN MAIZE

A paper published in eLife 2014; 3: e03724

Jianbo Zhang*, Tao Zuo*, Dafang Wang, Thomas Peterson

Department of Genetics, Development and Cell Biology, Department of Agronomy,

Iowa State University, Ames, Iowa 50011

*These authors contributed equally to this work.

Abstract

Every DNA segment in a eukaryotic genome normally replicates once and only once per cell cycle to maintain genome stability. We show here that this restriction can be bypassed through alternative transposition, a transposition reaction that utilizes the termini of two separate, nearby transposable elements (TEs). Our results suggest that alternative transposition during S phase can induce re-replication of the TEs and their flanking sequences. The DNA re-replication can spontaneously abort to generate double-strand breaks, which can be repaired to generate Composite Insertions composed of transposon termini flanking segmental duplications of various lengths. These results show how alternative transposition coupled with DNA replication and repair can significantly alter genome structure and may have contributed to rapid genome evolution in maize and possibly other eukaryotes.

Introduction

Initiation of DNA replication in eukaryotic cells is controlled by the replication licensing system (Blow, 1993; Blow and Dutta, 2005; Truong and Wu, 2011), which ensures that each segment of the genome is replicated only once per cell cycle. The expression and activity of the replication licensing factors are precisely regulated, and misexpression or mutation of these
factors can lead to DNA re-replication, genome instability, major chromosomal rearrangements, and tumorigenesis (Green et al., 2010; Green and Li, 2005; Hook et al., 2007; Lontos et al., 2007; Melixetian et al., 2004; Rice et al., 2005; Sugimoto et al., 2009). Misregulation of some histone methyltransferases can also result in DNA re-replication in plants and animals (Fu et al., 2013; Jacob et al., 2010; Tardat et al., 2010).

Although DNA replication is strictly controlled, some DNA segments can escape this restriction and replicate more than once in a single cell cycle in normal cells. For example, some Class II DNA transposons, including the maize Ac/Ds system, *E. coli* TN10, and *E. coli* TN7, are known to transpose during DNA replication (Chen et al., 1987; Peters and Craig, 2001; Roberts et al., 1985). If a replicated transposon excises and reinserts into an unreplicated site, the transposon can undergo one additional replication in the same S phase; the re-replication, however, is limited to the TE itself, and does not extend into the TE-flanking regions.

We and others have previously shown that a pair of Ac termini in reversed orientation can undergo transposition, generating major chromosomal rearrangements such as deletions, inversions, permutations, duplications, and reciprocal translocations (Huang and Dooner, 2008; Zhang and Peterson, 2004; Zhang et al., 2009; Zhang et al., 2006; Zhang et al., 2013), this transposition reaction is termed reversed Ac ends transposition (RET). All the RET-generated genome rearrangements described to date are fully explained by models in which the excised TE termini inserted into target sites that had completed DNA replication. However, it seems reasonable to expect that RET, like standard Ac/Ds transposition, may also occur during DNA replication, and that the excised reversed Ac termini could insert into unreplicated target sites. Here we show that such events do occur, and that they can induce re-replication of the TE and its
flanking sequences. This process generates novel structures termed Composite Insertions (CIs) which contain TE sequences and variable lengths of the flanking genomic DNA.

**Results**

**Model of transposition-mediated DNA re-replication**

The allele *P1-ovov454* (GenBank accession # KM013692) carries an intact *Ac* element and a fractured *Ac* (*fAc*) element inserted in the second intron of the maize *p1* gene; the 5’ terminus of *Ac* and the 3’ terminus of *fAc* are present in reversed orientation with respect to each other, and separated by an 822 bp inter-transposon segment (Figure 1A) (Yu et al., 2011). Our recent work showed that the *P1-ovov454* allele undergoes RET to generate derivative alleles containing either deletions or Tandem Direct Duplications (TDDs; Figure 1—Figure supplement 1). These are formed as a direct consequence of transposition of the *Ac/fAc* termini into a replicated target site on the sister chromatid (Zhang et al., 2013). The deletions and TDDs vary in size depending on the position of the insertion site (green/black triangle in Figure 1—Figure supplement 1); the TDDs previously characterized range in size from 8 kb to 5.3 Mb (Zhang et al., 2013).

Here, we asked: what are the consequences of RET events that occur during DNA replication? We developed and tested models in which replicated *Ac/fAc* termini are excised by RET and inserted into unreplicated target sites. As shown in Figure 1 (see also the animation Movie 1), this type of transposition reaction places already-replicated DNA in front of a replication fork where it may undergo a second round of replication. We propose that the re-replication fork may spontaneously abort, yielding two chromatid fragments terminated by double-strand breaks (DSBs); fusion of the DSBs restores the chromosome linearity, and
generates CIs containing $Ac/fAc$ and their flanking sequences at the duplication breakpoints. By comparing RET events involving insertion sites that are unreplicated (Figure 1) versus replicated (Figure 1—Figure supplement 1), we can see that both types of events generate TDDs whose sizes are determined by the transposon insertion site. However, only events with unreplicated insertion sites also generate CIs via re-replication of the $Ac/fAc$ and their flanking sequences; the resulting products are termed TDDCI alleles. Because the formation of TDDs was described in detail previously (Zhang et al., 2013), here we will focus on the origin and characterization of the CI of the TDDCI alleles.

**Identification of alleles with Composite Insertions (CIs)**

Both TDD and TDDCI alleles contain similar duplication structures and should exhibit similar phenotypes. Therefore we screened maize ears as described previously to visually identify putative TDD-containing alleles (Zhang et al., 2013). We identified 25 candidate alleles, and cloned and sequenced the duplication/$Ac$ junctions (the green segment flanking the $Ac$ 5’ end in Figure 1E and Figure 2A) from 16 of the 25 TDD/TDDCI candidates via $Ac$ casting (Singh et al., 2003; Wang and Peterson, 2013) or inverse PCR (iPCR) (see Zhang et al (2013) for detailed screening and cloning methods). To identify the TDDCI alleles, we designed PCR primers that flank the progenitor insertion target sites for each allele (Figure 2A, primers 1 and 2). Primers 1 + Ac5 can amplify a product from both TDD and TDDCI while primers 2 + Ac3 can amplify a product only from TDDCI since the latter contains an additional CI (Figure 2A). As expected, PCR using primers 1 + Ac5 produced bands of the expected sizes in all the 16 alleles (Figure 2B, upper panel; seven examples are shown here). Whereas, primers 2 + Ac3 produced bands with expected sizes from only seven alleles (Figure 2B, lower panel). Sequencing of the PCR products obtained from primers 1 + Ac5 and 2 + Ac3 revealed that these seven TDDCI candidates have
duplication/insertion breakpoints located from 13,392 bp to 1.7 Mb proximal to the \textit{p}1 locus on chromosome 1 (Table 1). Importantly, the \textit{Ac} termini are flanked by 8 bp target site duplications (TSDs; green/black triangles in Figure 1E) as predicted by the model in Figure 1 (See Supplementary file 1 for sequences containing TSDs).

Of particular importance are the results derived from three red/white twinned sectors, in which a sector of red kernel pericarp (seed coat) is twinned with an adjacent white pericarp sector (Figure 3). From each red pericarp sector we isolated \textit{P1-rr} alleles (\textit{P1-rr-T21}, \textit{P1-rr-T22} and \textit{P1-rr-T24}), and from each white twin sector we isolated corresponding \textit{p1-ww} alleles (\textit{p1-ww-T21}, \textit{p1-ww-T22}, and \textit{p1-ww-T24}). Similar types of twinned pericarp sectors have been shown to arise from the reciprocal products of standard \textit{Ac} transposition events (Chen et al., 1992; Greenblatt and Brink, 1962). Here, we propose that each pair of red/white twinned alleles are derived from the reciprocal TDDCI/deletion products of RET (sister chromatids shown in Figure 1E). This was tested by PCR using primers 2 + Ac3; as shown in Figure 2B (lower panel), these primers produced bands of the same size for each set of twinned alleles. Moreover, for each pair of red/white co-twins, the sequences of the PCR products obtained using primers 2 + Ac3 are identical (Supplementary file 1). Together these results are consistent with the model of RET during DNA replication as shown in Figure 1.

**Structures of the TDDCI alleles and Composite Insertions**

Because PCR only provides information on rearrangement junctions, we further analyzed the structures of the candidate TDDCI alleles by DNA gel blot. Genomic DNA was digested with \textit{SacI} and the blot was hybridized with probe 8B (gray boxes in Figure 2A). This probe detects the \textit{p}1 gene (12.7 kb band), the paralogous \textit{p}2 gene (4.7 kb band), and the \textit{p1-ww[4Co63]} allele (5.0 kb band) (Goettel and Messing, 2010) on the homologous chromosome. First, the 12.7
kb $p1$ band is absent in the three twinned $p1$-ww alleles ($p1$-ww-$T22$, $p1$-ww-$T24$, and $p1$-ww-$T21$; Figure 2C, lanes 4, 6 and 10, respectively). This result confirms the presence of a deletion as predicted by the model shown in Figure 1. Second, the alleles $P1$-$rr$-$T24$, $P1$-$rr$-$E17$, and $P1$-$rr$-$E340$ show a more intense 4.7 kb $p2$ band, in comparison with the 5.0 kb band (Figure 2C, lanes 5, 7, 8). This result is also expected because these three alleles have duplications of >70 kb (Table 1) which generate additional copies of the $p2$ gene located ~70 kb proximal to $p1$. Third, alleles $P1$-$rr$-$T22$, $P1$-$rr$-$T21$, and $P1$-$rr$-$E5$ (Figure 2C, lanes 4, 9 and 12, respectively) exhibit one or two new bands hybridizing with probe 8B. This is consistent with the presence of a CI which contains a newly-generated copy of the 8B sequence (Figure 2A). In $P1$-$rr$-$T22$, the duplication/insertion breakpoint occurred in the $p2$ band containing probe 8B, resulting in a shift of the 4.7 kb band to ~8 kb (Figure 2C, lane 3). Moreover, this ~8 kb band is more intense than the 5.0 kb $p1$-ww[4Co63] band and the 12.7 kb $p1$ band in $P1$-$rr$-$T22$ (lane 3 in Figure 2C). The model in Figure 1 and our analyses indicate that the intense ~8 kb band is actually a triplet containing two copies of a new 8,461 bp $p1$ fragment (one from the TDD, and a second from the CI, see below) and one copy of a 8,127 bp $p2$ fragment from the rearrangement junction. Further DNA gel blot analyses with a different $p1$ probe (not shown) confirm that $P1$-$rr$-$T22$ contains a TDD. All together, these results indicate that these four alleles-- $P1$-$rr$-$T22$, $P1$-$rr$-$T24$, $P1$-$rr$-$E17$, and $P1$-$rr$-$E340$--contain the TDDCI structure.

We then characterized the structures of the CIs in the four TDDCI alleles. The model in Figure 1 predicts that the insertion size and structure are determined by where re-replication aborts and how the resulting DSBs are repaired (Figure 1D). The structures of the CIs were determined by PCR using a series of divergent primer pairs flanking the $Ac/fAc$ insertions (δ and π, the blue arrows in Figure 2A). These primers will not amplify products from the progenitor
*P1-ovov454* allele because they point away from each other (Figure 2A). However, if the CI is formed by re-replication and the *Ac/fAc* flanking segments are fused as shown in Figure 1E and Figure 2A, then these primers will be oriented towards each other and can amplify the internal sequence of the insertion. In this way we obtained the internal sequences carried by the CIs in *P1-rr-T22* and *P1-rr-E17*.

The CI in *P1-rr-T22* is 23,238 bp in length (GenBank accession # KM013690), consisting of 14,484 bp of *fAc* and its distal flanking sequence and 8754 bp of *Ac* and its proximal flanking sequence (Figure 4); these two fragments are joined at a 4 bp microhomology sequence consistent with DSB repair via non-homologous end joining (NHEJ). In addition to the CI, the *P1-rr-T22* allele carries a 70 kb TDD (Table 1), and its white co-twin *p1-ww-T22* carries a reciprocal 70 kb deletion; moreover, the breakpoints of both the *P1-rr-T22* duplication and *p1-ww-T22* deletion contain 8 bp target site duplications. All of these features are predicted by the RET/re-replication model shown in Figure 1.

The CI in *P1-rr-E17* (GenBank accession # KM013689) is 19,341 bp in length (Figure 4); its structure suggests that the DSBs predicted in Figure 1D were repaired via homologous recombination (HR) between two direct repeat sequences that flank the *pl* gene in *P1-ovov454* (Lechelt et al., 1989). These repeats (hatched boxes in Figure 2A) are 5248 bp in length; the proximal copy is 4555 bp from the *Ac* element while the distal copy is 2934 bp from *fAc* (Figure 2A). If re-replication continued beyond the *Ac* and *fAc* segments and into the flanking 5248 bp repeats before aborting, then the DSBs could be repaired via HR to generate the observed structures (Figure 5). The two repeat copies flanking *P1-ovov454* differ at six SNPs in the distal half of the repeats (Figure 2A, red vertical short lines in the hatched box). Sequences of the *P1-rr-E17* allele show that the repeat in the CI is identical to the proximal copy. These results
suggest that the HR crossover occurred between the proximal halves of the two repeats (Figure 5).

For P1-rr-T24, no product could be amplified using the divergent primer strategy described above. However, a band of ~5.0 kb could be amplified using primers 1+2 which flank the insertion site. This band was sequenced and found to contain an intact Ac element (Figure 4). It seems very unlikely that this Ac was inserted through a simple transposition event, because the insertion site is located precisely at the duplication junction that is generated by RET, and an independent Ac transposition would not be expected to insert into precisely the same site. We suggest that the Ac insertion in P1-rr-T24 was produced by HR between the re-replicated Ac and fAc segments as they share 2039 bp of sequence identity (Figure 6 and Movie 2). Finally, the structure of the CI in P1-rr-E340 is still unknown; DNA gel blotting (not shown) indicated that the Ac-proximal fragment is in the range of 18-90 kb and the fAc-distal fragment is greater than 18 kb, resulting in a CI of at least 36 kb in length.

**RET-mediated DNA re-replication can generate solo-CI**

PCR results show that the P1-rr-E311, P1-rr-T21 and P1-rr-E5 alleles contain junctions consistent with the presence of CI (Figure 2B, lanes 12 – 18). However, DNA gel blot analysis suggests that these same alleles do not contain TDDs (Figure 2C, lanes 9 - 12). Importantly, the CI in P1-rr-T21 is flanked by a target site duplication, and the CI insertion site is identical to the deletion breakpoint in the co-twin p1-ww-T21; these results strongly suggest that these twinned alleles were generated as the reciprocal products of an alternative transposition mechanism. We propose that the solo-CI alleles were formed by a mechanism similar to that shown in Figure 1, except that the termination of replication (Figure 1C) resulted in release and loss of the rolling circle. Because the TDD originates from the DNA included in the rolling circle, release of the
rolling circle and subsequent DSB repair will result in a chromatid that carries only the CI (Figure 7 and Movie 3). The CI structures of these three alleles were characterized via PCR using primers δ and π as described above, and are diagrammed in Figure 4.

In *P1-rr-T21*, the CI is 14,287 bp in length and contains a 3 bp microhomology region at the internal junction (GenBank accession # KM013688), consistent with DSB repair via NHEJ. For *P1-rr-E311*, the CI is 23,647 bp in length and has no apparent microhomology sequence at the internal junction (GenBank accession # KM013691), which is not uncommon for NHEJ-mediated repair (Kramer et al., 1994; Lloyd et al., 2012; Wu et al., 1999). *P1-rr-E311* does not contain a TDD, and its CI does not include fragment 8B; therefore the DNA gel blotting pattern in *P1-rr-E311* is the same as its progenitor *P1-ovov454* (lane 2 and lane 11 in Figure 2C).

Finally, the CI in *P1-rr-E5* is 19,341 bp; its structure is identical to that in *P1-rr-E17* (Figure 4), even though these alleles arose independently and have the CI in different positions (16,497 bp and 1.7 Mb proximal to the *Ac* element in *P1-ovov454*, respectively; Table 1). We propose that both cases were produced via HR between the 5248 bp *p1*-flanking repeat sequences as described above and shown in Figure 5.

**RET mediates DNA re-replication at other genomic loci in maize**

In addition to the above alleles, we identified another allele (*P1P2-3*, Figure 8A) that contains a CI but which was derived from a different progenitor allele (*p1-vv-D103*). The structure of *p1-vv-D103* is similar to that of *P1-ovov454*, except that the *fAc* element is shorter (779 bp vs. 2039 bp in *P1-ovov454*) and the sequence distal to *fAc* has been replaced by chromosome 10 due to a chromosome 1-10 reciprocal translocation (in preparation). Like the examples described above, the *P1P2-3* allele arose in a single generation from *p1-vv-D103*; it contains a TDD of 80 kb, and a CI of 10,191 bp composed of 5017 bp of *Ac* and *Ac*-proximal
flanking sequence and 5174 bp of fAc and fAc-distal flanking sequence. This structure is the same as that predicted by the model in Figure 1. The internal breakpoint junction of the CI contains a 9 bp homologous sequence, consistent with DSB repair via a microhomology-mediated end joining (MMEJ) mechanism (Ma et al., 2003; McVey and Lee, 2008).

If alternative Ac/Ds transposition can induce DNA re-replication and the formation of linked duplications and Composite Insertions, one may be able to detect these products at other loci. Interestingly, Barbara McClintock isolated an allele of the maize bronze1 gene (bz1-m4-D6856) (McClintock, 1956) which has a complex structure consisting of three TDDs of bz1 and its flanking sequence, separated by Ds elements (Figure 8B) (Dowe et al., 1990; Klein et al., 1988). The third repeat is not complete; its proximal side (including the bz1 coding sequence) is truncated and joined to a truncated Ds sequence. This structure is similar to that of P1-rr-T22, P1-rr-E17, and P1P2-3 described above: two intact Tandem Direct Duplications (p1 vs. bz1 sequence), separated by TE (Ac vs. Ds), adjacent to a CI. In the case of bz1-m4-D6856, the CI contains the truncated copy of the tandem duplication and the truncated Ds, and is flanked by 8 bp target site duplications. We propose that bz1-m4-D6856 originated via a mechanism very similar to that shown in Figure 1: RET of two Ds elements located distal to the bz1 gene, followed by insertion of the excised Ds termini into an unreplicated target site in the bz1 5’ UTR region. The three tandem repeats would have been formed by rolling circle replication; one replication fork would have dissociated from the circle distal to the bz1 coding region to generate the incomplete repeat, while the other fork would have dissociated from the Ds element to generate a truncated Ds (Movie 4). This model presupposes the existence of a Ds element (the leftmost element in Figure 8B) distal to the tandem repeats in bz1-m4-D6856 and its progenitor allele. No such element was reported on the original bz1-m4-D6856 genomic clones (Dowe et
al., 1990; Klein et al., 1988). Efforts in our lab to identify a Ds element in this position in bz1-m4-D6856 and related stocks have been unsuccessful. However, McClintock's description of the origin of bz1-m4-D6856 (as reported in Klein et al. 1988) indicates that the bz1-m4 progenitor produced a high frequency of dicentric chromosomes, while the bz1-m4-D6856 derivative exhibited low dicentric frequency. Dicentric chromosome formation is a characteristic feature of alternative transposition reactions, such as RET, involving two nearby Ac/Ds elements (Huang and Dooner, 2008; Yu et al., 2010). The switch from high to low dicentric frequency observed by McClintock would be consistent with excision of the "missing" Ds shortly after the formation of the bz1-m4-D6856 allele.

**Discussion**

We have identified a new pathway leading to re-replication of specific chromosome segments in maize. This pathway is initiated by transposase-induced excision of the replicated termini of nearby transposons, followed by insertion of the excised transposon ends into an unreplicated target site. Re-replication begins when chromosomal replication forks reach the transposon, and may continue for considerable distances into the flanking DNA before aborting. The two resulting chromatid ends are joined together to restore chromosome linearity. This re-replication pathway is localized to the transposons and their flanking sequences, and does not require origin re-initiation. In contrast, deregulating licensing factor activity results in re-firing of replication origin(s), leading to re-replication at multiple dispersed origins (Green et al., 2006).

Although little is known about termination of eukaryotic DNA replication, studies in yeast indicate that termination does not require specific terminator sites, but occurs wherever two replication forks converge (McGuffee et al., 2013). Here, we propose that alternative transposition reactions can interrupt normal fork convergence. For example, Figure 1 shows that
converging replication forks α and β are separated from each other by alternative transposition (Figure 1C); if not terminated by other factors, replication fork β could in principle continue until the end of the chromosome, which is ~48 Mb from the p1 locus. However, our results suggest that DNA re-replication tends to abort after relatively short distances. The re-replicated segments generated from a single replication fork range in size from 4781 bp to 18,866 bp; the structure of the insertion in P1-rr-E340 is unknown, but DNA gel blotting analysis suggests a size of at least 36 kb. Thus the total extent of DNA re-replication is less than 19 kb in eight of nine alleles examined. In contrast, break-induced replication in yeast is capable of replicating from the site of a DSB to the end of the chromosome (Kraus et al., 2001). What causes termination of re-replication following alternative transposition in maize? One possibility is fork chasing and head-to-tail fork collision (rear-ending), which has been shown to cause fork collapse and termination of DNA re-replication in Xenopus (Davidson et al., 2006). Alternatively, re-replication may spontaneously stall and abort due to compromised fork progression as reported in yeast (Green et al., 2010).

Our model proposes that DNA re-replication aborts to produce chromatids terminated by broken ends, which are joined together to restore chromosome linearity (Figures 1, 6, 7, and 8). If the chromatid DSBs were not repaired, the cell would die and that event would not be recovered in our screen. From a population of ~2000 plants, we isolated 16 alleles that carry a duplication and/or insertion structure. Nine of these 16 alleles (56%) have only a duplication (Zhang et al., 2013), which indicates that the target site was replicated at the time of RET (Figure 1—Figure supplement 1); whereas seven alleles have an insertion, which indicates the target site was unreplicated (Figures 1, 6, and 7). The frequency of insertion into an unreplicated target site is 7/16 (44%), which is similar to a previous estimate of Ac insertion into unreplicated sites
(Greenblatt and Brink, 1962). Thus the products of insertion into unreplicated target sites are not significantly under-represented in our sample, suggesting that repair of re-replication-generated DSBs is quite efficient in mitotic S phase cells.

DNA lesions caused by replication fork stalling and collapse can be repaired by HR, NHEJ, MMEJ, replication slippage, FoSTeS (fork stalling and template switching), BIR (break-induce replication), MMBIR (microhomology-mediated break-induced replication), MMIR (microhomology/microsatellite-induced replication), and other mechanisms (Hastings et al., 2009a; Hastings et al., 2009b; Kraus et al., 2001; Lee et al., 2007; Ma et al., 2003; McVey and Lee, 2008; Payen et al., 2008). In mammalian cells, replication fork-associated DSBs are predominantly repaired via HR (Arnaudeau et al., 2001). Among the six CI alleles sequenced here, three were repaired by HR and three by NHEJ, indicating that these two repair pathways have relatively similar activities during the S phase of mitosis in maize.

An important advantage of the maize system is the ability to identify genetically twinned sectors, and to propagate and analyze their corresponding alleles. Because twinned alleles are the reciprocal products of a single event (Greenblatt and Brink, 1962), their structures should reflect a single parsimonious mechanism of origin. This allows us to distinguish among a variety of possible mechanisms for formation of segmental duplications. For example, non-allelic homologous recombination (NAHR) could generate a TDD joined and flanked by Ac as observed in P1-rr-T24 if there were a p1-proximal Ac element in the progenitor allele P1-ovov454 (Figure 9); however, such an NAHR event cannot explain the observed structure of the white co-twin p1-ww-T24 (compare upper chromatids of Figure 6F and 9C). Similarly, re-replication induced gene amplifications (RRIGA, a mechanism that couples NAHR and DNA re-replication) can also generate chromosome structures very similar to that of P1-rr-T24 (Finn and
Like NAHR, RRIGA would also require a \( p1 \)-proximal \( Ac \) element as in Figure 9B. However, the reciprocal product of an RRIGA-generated TDD would be a chromosomal fragment that lacks a centromere and telomeres, which would be lost in subsequent cell divisions. Therefore neither NAHR nor RRIGA can generate the white co-twin \( p1-ww-T24 \).

In contrast, the actual structure of the white co-twin \( p1-ww-T24 \) is exactly as predicted by the RET re-replication model shown in Figure 6. The structures of the other TDDCI alleles are also inconsistent with NAHR and RRIGA: the duplicated segments are flanked by \( Ac \) on the left side and a CI on the right side (Figure 1E, lower chromatid), while NAHR/RRIGA-induced duplications would be flanked by identical \( Ac \) copies (Figure 9C, lower chromatid). Finally, NAHR and RRIGA generate TDDs of the same structure recurrently. In contrast, all of the TDDCI alleles we have isolated to date have different duplication breakpoints. This is consistent with their origin via alternative transposition, because the duplication endpoints are determined by the position of the transposon insertion site, which is expected to differ for each transposition event. Moreover, the RET reinsertion sites have the same characteristic features as for standard \( Ac/Ds \) transposition, including preferential insertion into nearby, hypomethylated, gene-rich regions (Greenblatt and Brink, 1962; Chen et al., 1992; Vollbrecht et al., 2010), and formation of 8-bp Target Site Duplications lacking sequence specificity (Vollbrecht et al., 2010). Taken together, our results consistently support the proposed mechanism of alternative transposition, re-replication and repair.

In summary, we show here that reversed \( Ac \) ends transposition can generate TDDs and CIs. The TDDs range in size from several kb to >1 Mb and thus can increase the copy number of multiple linked genes and their regulatory sequences. The CIs we have discovered may be 20 kb or more in length. These are produced as a consequence of \( Ac \) transposition during DNA
replication, and they exhibit a number of interesting features. First, the internal portions contain sequences that were originally flanking the donor Ac/fAc elements; the relative positions of these sequences are now switched, and they are fused together at a new junction. Because Ac/Ds elements are commonly inserted within or near genic sequences in plants, CI formation may shuffle the coding and/or regulatory sequences of the formerly flanking genes to create novel products. Moreover, the CIs are bordered by transposition-competent Ac/fAc 5’ and 3’ termini, hence the entire CI has the structure of a macrotransposon (Huang and Dooner, 2008; Yu et al. 2010) which could subsequently transpose to new sites and increase in copy number. Eukaryotic genomes contain significant portions of Tandem Direct Duplications, dispersed segmental duplications, and tandem multi-copy arrays (Bailey et al., 2004; Bailey et al., 2008; Bailey et al., 2003; Dujon, 2010; Rizzon et al., 2006; Shoja and Zhang, 2006; Tremblay Savard et al., 2011); our results suggest that transposition-induced DNA re-replication may have played an important role in generating these segmental expansions during genome evolution.

Materials and methods

Genetic stocks, other materials and methods used here are similar to those described previously (Zhang et al., 2013). Following is a condensed description, for full details see Zhang et al (2013).

Genetic stocks

The maize p1 gene encodes an R2R3-Myb transcription factor which regulates kernel pericarp (seed coat) and cob coloration. The phenotype conferred by each p1 allele is indicated by the particular suffix: P1-rr specifies red pericarp and red cob, p1-ww specifies white (colorless) pericarp white (colorless) cob, and P1-ovov specifies orange variegated pericarp (seed
coat) and orange variegated cob. *P1-ovov454* confers orange/red pericarp with frequent colorless sectors attributed to alternative transposition events which abolish *p1* function (Yu et al., 2011). The *p1-ww[4Co63]* allele is from the maize inbred line 4Co63 (Goettel and Messing, 2010).

Ears of plants of genotype *P1-ovov454/p1-ww[4Co63]* were fertilized with pollen from plants of genotype *C1, r1-m3::Ds, p1-ww[4Co63]*. The *r1-m3::Ds* allele is an *Ac* reporter allele: *Ac*-encoded transposase excises *Ds* from *r1-m3::Ds*, resulting in *r1* reversion and purple aleurone sectors. Changes in *Ac* copy number can be inferred by the negative *Ac* dosage effect: increased copy number of *Ac* delays the developmental timing of *Ac/Ds* transposition and reduces the frequency of early transposition events, generally producing variegated patterns with fewer, later transposition events (McClintock, 1948; McClintock, 1951). Reversed *Ac* ends transposition (Figure 1) can generate two non-identical sister chromatids: one carries a TDDCI, and the other a reciprocal deletion (Figure 1E). At mitosis these chromatids will segregate into adjacent daughter cells, forming an incipient twinned sector. The sector with the deletion chromosome has lost *Ac* and exons 1 and 2 of the *p1* gene; loss of *Ac* and *p1* functions will specify kernels with colorless pericarp and no purple aleurone sectors. The sector with the duplication chromosome retains a functional *P1-ovov454* gene and three copies of *Ac*; the predicted kernel phenotype will be orange/red pericarp with fewer colorless pericarp sectors, and fewer/smaller kernel aleurone sectors. Similar twinned sectors can also be formed via the mechanism in Figure 6 or 7. Mature ears were screened for multi-kernel twinned sectors with these characteristics; kernels from selected sectors were grown and analyzed. Alleles derived from twinned sectors or whole ears are indicated by a “T” or “E”, respectively, prior to the allele number.
Genomic DNA extractions, DNA gel blot hybridizations

Total genomic DNA was extracted using a modified cetyltrimethylammonium bromide (CTAB) extraction protocol (Porebski et al., 1997). Restriction enzyme digestions and agarose gel electrophoresis were performed according to manufacturers’ protocols and Sambrook et al (1989). DNA gel blots and hybridizations were performed as described (Sambrook et al., 1989), except hybridization buffers contained 250 mM NaHPO4, pH 7.2, 7% SDS, and wash buffers contained 20 mM NaHPO4, pH 7.2, 1% SDS.

PCR amplifications

Sequences of oligonucleotide primers are shown in Table 2; note that primers 1 and 2 are specific to each allele, depending upon the flanking sequences. PCR was performed using HotMaster Taq polymerase from 5 PRIME (Hamburg, Germany). Reactions were heated at 94°C for 2 min, and then cycled 35 times at 94°C for 20 s, 60°C for 10 s, and 65°C for 1 min per 1 kb length of expected PCR product, then 65°C for 8 min. In some reactions 0.5–1 M betaine and 4%–8% DMSO were added to improve yield. PCR products were separated on agarose gels, purified and sequenced directly by the DNA Synthesis and Sequencing Facility, Iowa State University, Ames, Iowa, United States. Ac casting and inverse PCR were used to isolate sequences flanking Ac insertions; these were performed as described previously (Zhang et al., 2009).

Acknowledgements

This research is supported by NSF award 0923826 to TP and JZ. We thank co-PIs David F. Weber (Illinois State University) and Dan Nettleton (Iowa State University) for their
participation in this NSF-supported project. We thank Terry Olson for technical assistance, and Douglas Baker for field work.

**Author Contributions**

J.Z., T.Z., Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; D.W., Acquisition of data, Analysis and interpretation of data, Drafting or revising the article; T.P. Conception and design, Analysis and interpretation of data, Drafting or revising the article;

**References:**


Table 1. Features of alleles generated by RET-induced DNA re-replication.

<table>
<thead>
<tr>
<th>Allele Number</th>
<th>Allele Type</th>
<th>Distance from donor locus to CI*</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1-rr-T21</td>
<td>Solo-CI</td>
<td>13,392 bp</td>
</tr>
<tr>
<td>P1-rr-E5</td>
<td>Solo-CI</td>
<td>16,497 bp</td>
</tr>
<tr>
<td>P1-rr-T22</td>
<td>TDDCI</td>
<td>70 kb</td>
</tr>
<tr>
<td>P1-rr-T24</td>
<td>TDDCI</td>
<td>80 kb</td>
</tr>
<tr>
<td>P1-rr-E340</td>
<td>TDDCI</td>
<td>447 kb</td>
</tr>
<tr>
<td>P1-rr-E311</td>
<td>Solo-CI</td>
<td>563 kb</td>
</tr>
<tr>
<td>P1-rr-E17</td>
<td>TDDCI</td>
<td>1.7 Mb</td>
</tr>
</tbody>
</table>

*Distance given is from the 5’ end of Ac in the progenitor P1-ovov454 allele, to the point of insertion of the CI; i.e. the distance between the TDD and CI insertion points in Figure 2A. In TDDCI alleles this distance is also the length of the duplicated segment. Except for the fully sequenced alleles P1-rr-T21 and P1-rr-E5, the values given are based on the B73 reference genome sequence (Schnable et al., 2009), which likely differs from the genotype used in these experiments.
Table 2. Primer sequences

<table>
<thead>
<tr>
<th>Primer 1</th>
<th>Primer 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>P1-rr-T22</strong></td>
<td><strong>P1-rr-T22</strong></td>
</tr>
<tr>
<td>CTGTGGTCGCTCTGCTCCG</td>
<td>AGAAGCTACTGGAACCTCGCACCTCA</td>
</tr>
<tr>
<td><strong>P1-rr-E17</strong></td>
<td><strong>P1-rr-E17</strong></td>
</tr>
<tr>
<td>AGATTTGACAGAAGAGCCGACAC</td>
<td>CCAGAGTATAGGGTGCTGGAGCC</td>
</tr>
<tr>
<td><strong>P1-rr-T24</strong></td>
<td><strong>P1-rr-T24</strong></td>
</tr>
<tr>
<td>GGTACGCCCATAATAAAAACAATAC</td>
<td>GCGTCCTCTATCCATTCACTTTCA</td>
</tr>
<tr>
<td><strong>P1-rr-E340</strong></td>
<td><strong>P1-rr-E340</strong></td>
</tr>
<tr>
<td>AACCGTGCTCATCATCATCAGTGT</td>
<td>TTTATGAGCGGCTGAATCGC</td>
</tr>
<tr>
<td><strong>P1-rr-T21</strong></td>
<td><strong>P1-rr-T21</strong></td>
</tr>
<tr>
<td>GGTGTTGTTGCTGCCTCC</td>
<td>CCGATGCTCTTTTCCTTCTCCTTCC</td>
</tr>
<tr>
<td><strong>P1-rr-E311</strong></td>
<td><strong>P1-rr-E311</strong></td>
</tr>
<tr>
<td>TCGTTCTGTGGGGTTCGTCGT</td>
<td>GCGATGCTATCAGTTAGACCAGGC</td>
</tr>
<tr>
<td><strong>P1-rr-E5</strong></td>
<td><strong>P1-rr-E5</strong></td>
</tr>
<tr>
<td>ATTTGGTCCTCCCTCCCT</td>
<td>CGCCGAACTTTACCTGCTCTGCTA</td>
</tr>
<tr>
<td>Ac3</td>
<td>Ac3</td>
</tr>
<tr>
<td>GATTACCGTATTTATCCGTTGTTTTC</td>
<td>GATTACCAGTATTTATCCGTTGTTTTC</td>
</tr>
<tr>
<td>Ac5</td>
<td>Ac5</td>
</tr>
<tr>
<td>CCCGTTCGTCCTCCGTTTTCGT</td>
<td>CCCGTTCGTCCTCCGTTTTCGT</td>
</tr>
</tbody>
</table>
Figure 1. Reversed Ac ends transposition (RET) during DNA replication generates Tandem Direct Duplication (TDD) and Composite Insertion (CI). Lines indicate a replicating chromosome, hexagons indicate replicons. The blue boxes are exons 1, 2, and 3 (right to left) of the pI gene, and the green/black triangles are the transposition target site. Red lines with arrow(s) indicate Ac/fAc insertions, and the open and solid arrowheads indicate Ac/fAc 3’ and 5’ ends, respectively. Two replication forks considered here are marked α and β. (A) The locus containing fAc/Ac is replicated. Vertical arrows indicate the sites of Ac transposase cuts at the fAc 3’ and Ac 5’ ends. (B) Transposase cleaves and the inter-transposon segment is ligated to form a circle. The excised transposon ends will insert into an unreplicated target site indicated as the green/black triangle. Like standard Ac/Ds transposition, insertion of the Ac/fAc termini into the target site generates an 8 bp target site duplication (TSD; green/black triangle). (C) Insertion of the excised transposon termini places fAc and fAc-flanking DNA ahead of replication fork β (upper chromatid), and Ac and Ac-flanking DNA ahead of replication fork α to generate a rolling circle replicon (lower chromatid). DNA replication continues. (D) Following re-replication of fAc, Ac, and a portion of the flanking sequences, DNA replication forks α and β stall and abort, resulting in chromatids terminated by broken ends (the red > or < symbol) (Michel et al., 1997). The dotted red line connects the two broken ends which will fuse together. (E) Chromatid fusion produces a chromosome with two unequal sister chromatids: The upper chromatid contains a deletion of the segment from fAc to the a/b target site. The lower chromatid contains a TDD (left-hand loop), as well as a new CI (right-hand loop). The TDD contains the DNA deleted from the upper chromatid; the CI contains the re-replicated Ac, fAc and flanking sequences. The junction where broken chromatid ends were joined is indicated by the red ×).
**Figure S1.** Reversed *Ac* ends transposition after DNA replication generates Tandem Direct Duplications (TDDs).

The two lines indicate sister chromatids of fully replicated maize chromosome 1, joined at the centromere (black). The blue boxes are exons 3, 2, and 1 (left to right) of the p1 gene. Red lines with arrowhead(s) indicate *Ac/fAc* insertions, and the open and solid arrowheads indicate the 3′ and 5′ ends, respectively, of *Ac/fAc*. The short horizontal arrows show the orientations and approximate positions of PCR primers, and the numbers below are the primer names. The green/black triangles indicate the transposon target site sequences and target site duplications. 

(A) *Ac* transposase cleaves the lower chromatid at the 3′ end of *fAc* and the 5′ end of *Ac* (arrows). 

(B) Following transposase cleavage, the internal *p1* genomic sequences are joined to form a circle. Dotted lines indicate the insertion of the *fAc* and *Ac* termini into the *a/b* site on the sister chromatid. (C) Transposon ends insert into the upper sister chromatid at the *a/b* target site. (D) The *Ac* 5′ end joins to the distal side (green) of the target site and the *fAc* 3′ end joins to the proximal side (black) of the target site to generate a proximal deletion (upper chromatid) and a direct duplication (lower chromatid). The shaded arrows encompass the duplicated segments.

Note: This figure is adopted from Figure 1 of Zhang et al (2013).
**Figure 2.** PCR screening and DNA gel blotting of candidate TDDCI alleles.

(A) Detailed structures of *P1-ovov454* (progenitor) and RET-generated *P1-rr-twin/p1-ww-twin* (TDDCI/Deletion) alleles deduced from Figure 1. The horizontal blue lines are *p1* gene sequence while the green lines are *p1* proximal sequences, including the *p2* gene sequence (*p1* paralog, ~70 kb proximal to *p1*); the blue and green boxes are exons 1, 2 and 3 (right to left) of *p1* and *p2*, respectively. The small horizontal arrows indicate the orientation and the approximate position of the PCR primers. The gray boxes indicate probe 8B used in DNA gel blot analysis,
the short vertical black lines are SacI sites, and the numbers between the SacI sites indicate the lengths of those fragments detected by probe 8B. The hatched boxes represent the distal (black) and proximal (green) 5248 bp repeats flanking the p1 locus. These repeats are identical except for six SNPs, indicated by short red vertical lines inside the green hatched box; (SNPs 3 and 4 are only 43 bp apart). Other symbols have the same meaning as in Figure 1.

**B** PCR products obtained using primers 1+Ac5 (upper) or 2+Ac3 (lower). Lane 1, 1 kb DNA ladder; lane 2, P1-ovo454; Lane 3, P1-rr-T22; 4, p1-ww-T22; lane 5, P1-ovo454; Lane 6, P1-rr-T24; 7, p1-ww-T24; lane 8, P1-ovo454; Lane 9, P1-rr-E17; lane10, P1-ovo454; Lane 11, P1-rr-E340; lane 12, P1-ovo454; Lane 13, P1-rr-T21; 14, p1-ww-T21; lane 15, P1-ovo454; Lane 16, P1-rr-E5; lane 17, P1-ovo454; Lane 18, P1-rr-E311. Note: the sequences of primers 1 and 2 are specific for each allele.

**C** DNA gel blot analysis of the TDDCI/deletion alleles. Genomic DNA was digested with SacI and the blot was hybridized with probe 8B (see Figure 2A for the position of the probe). Lane 1: p1-ww[4Co63], Lane 2: P1-ovo454/p1-ww[4Co63], Lane 3: P1-rr-T22/p1-ww[4Co63], Lane 4: p1-ww-T22/p1-ww[4Co63], Lane 5: P1-rr-T24/p1-ww[4Co63], Lane 6: p1-ww-T24/p1-ww[4Co63], Lane 7: P1-rr-E17/p1-ww[4Co63], Lane 8: P1-rr-E340/p1-ww[4Co63], Lane 9: P1-rr-T21/p1-ww[4Co63], Lane 10: p1-ww-T21/p1-ww[4Co63], lane 11: P1-rr-E311/p1-ww[4Co63], Lane 12: P1-rr-E5/p1-ww[4Co63].
Figure 3. An ear with twinned sectors. The photo shows two sides of the same ear. Left-side view has a large area with parental $P1$-ovo454 phenotype (orange pericarp with frequent colorless sectors), while the right-side view shows a large area with typical $P1$-rr-Twin phenotype (dark red pericarp with few colorless sectors). A single large $p1$-ww-Twin sector (kernels with mostly colorless pericarp) is visible in both views. The solid purple kernels present in all the sectors result from an independent germinal reversion of the $r1$-m3::Ds allele, and can be ignored.
Figure 4. The structures and sizes of Composite Insertions (CIs). The double-headed arrows (left side) indicate Ac elements, while the single-headed arrows (right side) indicate fAc. The red × symbol indicates the junction of the two re-replicated segments in the insertion. Other symbols have the same meaning as in Figure 1 and Figure 2A.
Figure 5. RET followed by homologous recombination generates identical 19,341 bp Composite Insertions in P1-rr-E17 and P1-rr-E5. (A) Structure of the chromosome 1S segment containing the progenitor P1-ovov454 allele, prior to RET. (B) Drawing shows the RET stage corresponding to Figure 1D. Recombination between the 5248 bp repeats near the two DSBs (marked by > or <) generates a Composite Insertion. (C) Structure of P1-rr-E17 containing TDD (left-hand triangle) and Composite Insertion (right-hand triangle). All the symbols have the same meaning as in Figure 2. Note: P1-rr-E5 contains the 19,341 bp CI, but does not contain the TDD. See text for details.
Figure 6. RET followed by homologous recombination generates a simple $Ac$ insertion in $PI$-$rr$-$T24$. (A), (B), (C), and (D) are the same as in Figure 1. (E) Homologous recombination occurs between the re-replicated $Ac$ and $fAc$. (F) Two new chromatids are formed: The lower chromatid contains a Tandem Direct Duplication and an $Ac$ insertion, and the upper chromatid carries a reciprocal deletion. For animated version, see Movie 2.
Figure 7. Generation of a Composite Insertion in the absence of a duplication. (A), (B), and (C) are the same as in Figure 1. (D) Upper chromatid contains deletion; in lower chromatid stalling and abortion of rolling circle replication fork releases the circle. (E) The two chromatids fuse to form a new chromatid containing a Composite Insertion. For animated version, see Movie 3.
Figure 8. Two additional maize alleles likely generated by RET and re-replication.

A. Structure of progenitor allele *p1-vv-D103* (upper) and TDDCI allele *P1P2-3* (lower). The *p1-vv-D103* allele is carried on a chromosome 1-10 translocation; the brown line indicates DNA segment from chromosome 10. See text for details. Other symbols have the same meaning as in previous figures. B. TDDCI structure of *bz1-m4-D6856*. The bronze-colored boxes indicate exons 1 and 2 (right to left) of the *bronze1* gene on maize chromosome 9. The baseline shows the predicted structure of the progenitor of *bz1-m4-D6856*. The dashed box encloses a hypothetical Ds element proposed to have been involved in the generation of *bz1-m4-D6856* via RET. For animation, see Movie 4. Other symbols as in previous figures. The structure of *bz1-m4-D6856* is deduced from Dowe et al (1990) and Klein et al (1988).
Figure 9. NAHR generates Tandem Direct Duplications. All the symbols have the same meanings as in Figure 1. (A) Ac transposes to a site between $a$ and $b$. (B) Homologous recombination between two non-allelic Ac elements on sister chromatids generates a deletion (upper chromatid) and a TDD (lower chromatid) in (C).
CHAPTER 4. THE FORMATION OF DNA TRANSPOSON-INDUCED TANDEM DIRECT DUPLICATIONS AND THEIR OCCURRENCE ACROSS 22 PLANT SPECIES

Tao Zuo, Hsien-chao Chou, and Thomas Peterson

Abstract

Recent whole genome sequencing has revealed significant levels of tandem direct duplications (TDDs) in many eukaryotic genomes. The formation of TDDs is commonly explained by non-allelic homologous recombination (NAHR); for example by unequal crossover between two transposable element (TE) copies located in non-allelic positions: if the TEs are in direct orientation, the segment between them will be duplicated. However, recent studies in maize and rice showed that reversed-ends transposition (RET) of Ac/Ds elements is also capable of inducing small to large (10 kb to > 5 Mb) TDDs at a frequency of 0.5% to 1%, when the two Ac/Ds elements are less than 1 kb apart. Still unclear is the overall extent and relative impact of DNA transposon-induced TDD in plant genomes. To address this issue, we developed and implemented a bioinformatics pipeline termed STRAND: Search for Transposon-Induced Tandem Direct Duplications. STRAND utilizes available TE databases and reference genome sequences to quickly and efficiently identify transposon-associated TDDs. Our results show that 14 out of 22 plant genomes analyzed contain one or more duplicated sequences initiated by TEs. Sequences were analyzed to distinguish, where possible, TDDs induced by RET vs. NAHR. The frequency of TDDs generated by RET and NAHR varies among plant species, but RET-induced TDDs are most often found at greater frequency than those induced by NAHR. Our results indicate that RET could have played an important role in generating TDDs and other genomic rearrangements during plant genome evolution.
Introduction

Segmental duplications--two or more copies of nearly identical DNA sequence with the size ranging from 1 to >200 kb--are prevalent in both plant and animal genomes, and have played important roles in genome evolution (Samonte and Eichler 2002; Flagel and Wendel 2009). Segmental duplications account for as much as 5.2% of the human genome (Bailey et al. 2002). In plants, because of their inability to move and historical polyploidization events, the proportions of segmental duplication in plant genomes are even greater; e.g. 65-85% of *Arabidopsis* genes have one or more homologous copies in the genome (Cannon et al. 2004). Tandem duplication is a type of segmental duplication in which homologous sequences are clustered at a single locus. Tandemly arrayed genes (TAGs) comprise about 10% of the duplicated genes in *Arabidopsis* and rice (Rizzon et al. 2006), and approximately one third of duplicated genes in human, mouse and rat genomes (Shoja and Zhang 2006).

Gene duplications could have dramatic impacts on organismal phenotype, since the increased gene copy number could lead to elevated levels of RNA transcripts due to gene dosage effects (Conrad and Antonarakis 2007). A number of recent studies have shown that gene duplications have played important roles in adaptation, domestication and human selection. For example, the adaptation of dogs to a starch-rich diet was facilitated by increased copy number of an amylase gene (Axelsson et al. 2013). In tomato, the change from round to elongated fruit shape was due to the early- and over-expression of a *SUN* gene included within a segmental duplication mediated by a *copia*-like retrotransposon (Xiao et al. 2008). In soybean, the *Rhg1-b* allele that confers resistance to nematode infection contains 10 tandem copies of a 31 kb segment (Cook et al. 2012). Finally, grain length in rice is significantly increased via the duplication of the *GL7* locus (Wang et al. 2015b).
In parallel with the growing understanding of the functional importance of gene
duplications, studies on the mechanisms that generate duplications have received more attention.
The *de novo* duplication of DNA sequences is commonly thought to occur via one of four major
mechanisms: non-allelic homologous recombination (NAHR), non-homologous end joining
(NHEJ), fork stalling and template switching (FoSTeS), and retrotransposition (Hastings et al.
2009b; Zhang et al. 2009a). Recent studies have revealed a new duplication mechanism based on
the activity of DNA transposons. In particular, the maize *Ac/Ds* transposon system can undergo
alternative transposition reactions that involve one end from each of two nearby elements.
Alternative transposition can induce reciprocal inverted duplications and deletions via Sister
Chromatid Transposition (Zhang and Peterson 1999; Zhang and Peterson 2005), and can
generate deletions, inversions, translocations and tandem direct duplications (TDDs) via
Reversed-Ends Transposition (RET) of *Ac/Ds* elements (Figure 1) (Zhang and Peterson 2004;
Similar discoveries have been described in rice as well (Xuan et al. 2011; Yu et al. 2012).
Analysis of features of duplicated segments in a number of plant genomes disclosed that
transposable elements are frequently located precisely at the border of these duplications (Wicker
et al. 2010), suggesting that transposon-mediated mechanisms such as NAHR and alternative
transposition might play an important role in generating genome duplications (Figure 2).

Because plant genomes contain significant portions of DNA transposons that are
potentially capable of undergoing RET, it is reasonable to test the impact of this mechanism
more broadly by asking several questions: 1) is there any evidence of RET-induced duplications
in other plant genomes? 2) Does this mechanism occur with other DNA transposons outside the
*hAT* family? 3) What is the relative impact of RET mechanism in genome evolution? To address
these questions, we developed a computational program (STRAND: Search for Transposon-Induced Tandem Direct Duplications) that identifies tandem duplications precisely bordered by transposable elements. We applied the STRAND program to maize and 21 other plant genomes followed by further data analysis and structural characterization of identified duplications.

Results

The algorithm of STRAND: Search for Transposon-Induced Tandem Direct Duplications

Detection of tandem direct duplications within fully sequenced genomes is computationally challenging, especially from genomes of large size or containing a high proportion of repeated sequence elements. To efficiently identify transposon-induced tandem direct duplications (TI-TDDs), we designed an algorithm that is based on their unique features. This algorithm takes known TE sequences as input data to search for tandem duplications initiated by these TEs. Using the TE sequences as queries, the search function of this algorithm can avoid computationally intensive and unproductive whole genome searching.

The details of the algorithm are depicted in Figure 3. In the first step, TE sequences are used as queries in BLAST searches of a reference genome to identify TE insertion sites. Secondly, two 100 bp segments flanking each transposon, one 5' adjacent and one 3' adjacent, are extracted and checked for sequence similarity. In step 3, a selection procedure is enforced: Two transposon family members with the same orientation, less than 50 kb apart and with homologous sequences flanking one terminal end but not the other end are retained for further structural analysis. Step 4 uses the unmatched 100 bp flanking segment located between the two TE copies to search its neighboring homologous region; this defines the endpoint of the transposon-associated duplication. In step 5, global sequence alignment is performed for non-TE duplicated sequences after the duplication endpoint is defined. Candidate loci contain two or
more TE copies in the same orientation within a 50 kb interval, as well as >60% coverage on TE-flanking duplications. These cases are then manually inspected in step 6 to determine whether they possess the other features of TE-induced TDDs, such as the presence of flanking target site duplications (TSDs) and the existence and position of a third element of the same family.

**DNA transposon-induced TDD discovery in maize**

We automated this searching process via implementing the algorithm in Perl. We first applied this pipeline to the maize B73 reference genome (ZmB73 RefGen_v3) that is well known for its high proportion of transposons (Schnable et al. 2009). In this study, we mainly focused on identifying the tandem direct duplications that are associated with miniature inverted-repeat transposable elements (MITEs), because MITEs are prevalent in eukaryotic genomes (Feschotte and Pritham 2007), and also because a database of plant MITEs is already established (Chen et al. 2014).

The maize MITE database contains representatives of 252 DNA transposon families, which belong to five DNA TE super-families, including *Tc1/Mariner, Mutator, hAT, CACTA* and *PIF/Harbinger*. We downloaded all of the maize seed sequences (496 consensus sequences) from the P-MITE database ([http://pmite.hzau.edu.cn/django/mite/](http://pmite.hzau.edu.cn/django/mite/)) and then loaded them into our pipeline to search for associated duplications in the maize B73 genome. The program returned 27 candidate duplications. An example of the graphic output is shown in Figure 4A. Further manual analysis found that 9 of 27 candidates came from a single event which contains eight tandem direct copies of the transposon itself, while two hits were derived from a single locus containing > 2 repeats. Of the remaining 16 genomic TDDs, 3 cases involved three or more tandem copies and were not considered further. The remaining 13 cases contained two copies of each duplication, with the size of the non-TE repeat ranging from 158 bp to 47 kb. These 13 cases
include two of the three hAT element-induced TDDs previously identified by Zhang and colleagues (Zhang et al. 2013); the third case (dhAT-Zm24) identified in the previous study was not returned in the initial output because the element dhAT-Zm24 has no significant homology with the 496 consensus transposon sequences, hence it was not represented in the search query. When the element sequence was added to the TE database, the pipeline successfully identified this case as a TDD, making a total of 14 TDDs identified in the maize B73 reference genome.

Importantly, TE-induced TDDs identified by this program are not only derived from the RET mechanism. Structurally similar TDDs could be formed via Non-Allelic Homologous Recombination (NAHR), for example by unequal crossover between two TE copies as shown in Figure 2A. TDDs derived from either NAHR or RET will share certain structural features, including 1) duplications will be in head to tail (direct) orientation; 2) duplication endpoints coincide with the TE copies; 3) the TE copy located between the duplicated segments will lack one of the TSDs that flank the other TE copy (Figure 2B). However, some features can distinguish TDDs induced by NAHR vs. RET. NAHR produces a TDD by a crossover between two TE copies inserted at sites that flank the incipient duplication. The resulting TDD contains three TE copies: the two parental copies flanking the TDD, and a third TE located between the duplicated segments (Figure 2A). Because the two parental TEs would have resulted from independent insertions, they will often contain sequence polymorphisms. Moreover, the central TE copy is a recombinant and should contain SNPs derived from both parental TEs, depending on the location of the crossover. Subsequent to formation of the TDD, one or more of the TE copies could excise, leaving behind a footprint derived from the original TSDs.

In contrast, the RET mechanism produces a TDD containing two TE copies at the duplication boundaries. These two transposons will be identical in sequence, because they are
derived from sister chromatid sequences (Figure 1). Following TDD formation, the TE copies and the associated duplicated segments will undergo mutation. However, the TEs and the duplicated segments will likely have similar percent identities because they were both generated in the same RET event. The RET mechanism also requires a third copy of transposon located nearby to initiate alternative transposition (Figure 1 and Figure 2B), and the distance between the third TE and the duplication could be up to 100 kb (Huang and Dooner 2008). Thus, the location of the third transposon copy is critical to distinguish NAHR from RET-induced events.

To characterize the origin of the 14 maize TDDs, we manually checked each case for the presence of a third TE copy. None of the 14 maize TDDs contained a third TE copy or TE excision footprints at the duplication endpoints, suggesting that NAHR was not involved in their formation. Among these 14 TDDs, 6 were excluded from further analysis due to the presence of sequence gaps within the duplications. Four of the remaining 8 TDDs were associated with partial elements that lack one terminus required for simple excision. NAHR can be excluded for these cases because the third partial TE copy is unable to excise and should be retained at the duplication endpoint (Figure S1), thus they are attributed to RET (Table 1). Among the remaining 4 cases, one case (SE220300214) has a 12 bp deletion from the 5' TIR of the middle TE copy, most likely the result of a loss of terminal sequences during the RET process (Zhang et al. 2013). A second case (SE220500133) has 100% identity between the two TE copies, which is expected via RET-induced duplication (Wang et al. 2015a) (Table 1). A third case (SE220300125) is associated with hAT transposons that will leave footprint after excision. The lack of footprint at the end of this duplication suggests no transposon excision event ever happened; its origin is classified as RET. Also we were unable to ascertain the origin of the fourth duplication (SE220500012) because of the absence of a third TE copy within a 100 kb
range; the third fully functional copy could transpose away after the formation of duplication in either model.

Importantly, our results show that other DNA transposons besides \textit{hAT} family elements can induce TDDs in maize via the RET mechanism. Among the 7 maize TDDs induced by RET, 3 cases were associated with \textit{PIF/Harbinger} elements (Table 1), suggesting that \textit{PIF/Harbinger} elements could also perform alternative transposition to generate genome rearrangements including the formation of a new chimeric gene (Wang et al. 2015a).

**The presence of DNA TE-induced TDDs in other maize lines**

Recent studies have revealed tremendous diversity among the genomes of different maize inbred lines (Liu et al. 2003; Buckler et al. 2006). Copy number variations (CNVs) and presence and absence variations (PAVs) are pervasive among maize genotypes, and these structural variations likely contribute to important phenotypic differences (Lai et al. 2010; Chia et al. 2012; Lu et al. 2015). Thus it would be interesting to determine whether the 8 TE-induced TDDs identified in the maize B73 genome are present in other maize lines. At this time the sequence of only one additional maize inbred line is available, that of Mo17, whose genome was assembled using the B73 sequence as a reference (Xin et al. 2013). We searched the Mo17 genome sequence using the same TE queries and parameters as used previously for B73; the Mo17 search did not yield any new TDDs. Interestingly, only four of the eight TDDs identified in B73 are also present in the Mo17 genome sequence (Table 1). One might suspect that the duplications unshared between B73 and Mo17 occurred since the separation of the two lines. However, the B73-specific duplications did not exhibit higher sequence similarity than the unshared cases, as would be expected for more recent duplications.
**Discovery of DNA TE-induced TDDs in other plant species**

To understand the broader impact of RET on plant genome evolution, we further applied this program to another 21 plant species including rice, sorghum, millet and potato (Table S1). The genome sequences for these were obtained from EnsemblGenomes and other websites (Table S2), and the representative MITE sequences for each species were collected from the PMITE database ([http://pmite.hzau.edu.cn/django/mite/](http://pmite.hzau.edu.cn/django/mite/)). Our results show that 13 out of 21 plant genomes contain one or more TDDs that were initiated by DNA TEes (Table 2). Among these 13 plant genomes, 10 of them have at least one 2-copy TDD; the total number of 2-copy TDDs detected is 48.

We further characterized the origin of the 2-copy TDDs based on their sequence and structural features using the same criteria as for maize B73 TDDs. We found both RET- and NAHR-induced cases among the 48 TDDs analyzed. A clear example of a NAHR-induced duplication in potato is shown in Figure 5. The duplicated segments are flanked by three similar hAT transposons; the two outside elements have 8 bp TSDs, produced by independent insertion events. Whereas, the middle element is a recombinant that carries SNPs and one TSD derived from each of the flanking elements. The frequency of TDDs generated by RET and NAHR varies among plant species. For example, we identified two 2-copy TDDs in sorghum (*Sorghum bicolor*) genome, both of which were produced by NAHR. In contrast, apple (*Malus domestica*) contains two duplications induced by RET. The *Medicago truncatula* and *Cajanus cajan* genomes contain similar numbers of duplications generated by each mechanism. Overall, RET-induced TDDs are found at similar or greater frequency than those induced by NAHR: among the 48 2-copy TDDs detected across 21 plant genomes, 22 were induced by RET, 11 were characterized as NAHR, and the other 15 could not be classified due to sequence erosion.
Among the 22 duplications derived from RET, 12 out of 22 were initiated by *Tcl/Mariner* elements, 5 were associated with *PIF/Harbinger elements*, 3 were generated from *hAT* transposons, and 2 were from *Mutator*. These results indicate that RET could have played an important role in generating TDDs and other genomic rearrangements during plant genome evolution.

**Discussion**

Our recent studies in maize showed that RET of *Ac/Ds* elements can induce tandem direct duplications and composite insertions (Zhang et al. 2013; Zhang et al. 2014). Similarly, other research groups demonstrated that various DNA transposons in different species could induce genomic rearrangements by alternative transposition mechanisms. For instance, paired Drosophila *P* elements could produce duplications and deletions via a mechanism termed “Hybrid Element Insertion” (Gray et al. 1996). In bacteria, alternative transposition of *Tn5* and *ISShal* transposons could induce deletions and inversions (Goryshin et al. 2003; Watanabe et al. 2007). Considering the fact that most eukaryotic genomes contain a high proportion of tandem direct duplications, and the finding that transposable elements in plants are often precisely located at the borders of tandem direct duplications (Wicker et al. 2010), it is reasonable to hypothesize that RET of DNA transposons might have contributed significantly to genome evolution.

In this study, we aimed to test this hypothesis by developing a computational program to detect TDD events that were derived from RET in maize and other plant species. A key feature of RET-induced TDDs is that each duplication copy begins with a DNA transposon, and then is followed by the duplication of other non-TE sequences (Zhang et al. 2013). Based on this feature, we developed and implemented a search algorithm to identify transposon-induced
TDDs. Applying the program to 22 sequenced plant genomes, we found clear evidence of DNA transposon-induced TDDs produced via both RET and NAHR. In total, we discovered 62 two-copy DNA transposon-induced TDDs in 11 of 22 plant genomes examined; 29 are derived from RET, and 11 are induced by NAHR. The potato (*Solanum tuberosum*) genome has the highest number (22) of TI-TDDs, followed by maize with 14 TI-TDDs. No TI-TDDs were detected in Arabidopsis and ten other plant genomes. The number of TI-TDDs might be related to the number of TEs in each genome (Table S1); genomes that have more MITEs tend to contain higher numbers of TI-TDDs.

Other mechanisms including NHEJ (Moore and Haber 1996), FoSTeS (Lee et al. 2007) and microhomology-mediated break-induced replication (MMBIR) (Hastings et al. 2009a) could also induce tandem direct duplications, but the duplication structures produced would be different from those identified here. Each of these mechanisms uses a non-homologous repair mechanism to rejoin the double strand breaks. The repair junctions are usually at random genome positions, often require micro-homology, and frequently lead to small deletions and possible insertion of foreign DNA. However, the duplications we found in this study are precisely bounded by the MITEs; i.e. the TE is located exactly at the beginning of each duplication. Also, both TE and duplicated sequences within the duplication are intact, lacking filler DNA or other sequence aberrations, suggesting that these duplications are less likely to be induced by any of these other mechanisms.

Considering the large amount of duplications discovered in plant genomes, the frequency of DNA transposon-induced TDDs identified in our current study might appear low. However, it is noteworthy to consider that a large proportion of the duplications in current plant genomes are the remnants of ancient polyploidization events (Blanc et al. 2003; Flagel and Wendel 2009),
whose copies could be present in different place of the modern genome. Also, the conditions applied here to characterize tandem direct duplications are rather stringent, requiring high sequence similarity, precise duplication structures, matched TSDs, etc.; i.e. we only focused on detecting the most recent duplications whose mechanism of origin can be determined; older duplication structures would be degraded due to sequence erosion. The similarities between the two duplicated segments in the majority of cases we identified are higher than 90%. Moreover, the sequence quality of the reference genomes would greatly affect the detection of duplications. For example, the presence of sequence gaps and assembly errors would separate the original duplication structures, rendering those events undetectable. Identical duplications resulting from very recent events may often be absent from genome assemblies due to sequence collapse (Emrich et al. 2007; Phillippy et al. 2008). In addition, the completeness of the DNA transposon database employed in the initial search is critical. In the current study, duplications that are associated with transposons not represented in the MITE database we used would not be detected. For example, the absence of dhAT-Zm24 in the maize MITE database prevented its inclusion in the initial pipeline output. Finally, it should be noted that duplications derived from RET may undergo subsequent homologous recombination, as they would contain an active TE located between the two repeated segments. This configuration has been shown to induce homologous recombination in both maize and Arabidopsis, leading to loss of the duplication structure (Xiao et al. 2000; Xiao and Peterson 2000).

Surprisingly, after characterizing the origin of the 62 identified duplications, we found that nearly half (29) of the 62 duplications were evidently derived from RET mechanism, and less than 1/5 (11) of the duplications were clearly produced by NAHR. Certainly, the fewer cases of NAHR we found in this study should not underestimate the role of NAHR in driving genome
evolution, because NAHR could indeed occur between any two homologous sequences. Here, all the duplications we identified are associated with Class II (DNA) transposons; the large number of Class I transposons (retroelements) in many plant genomes may also be substrates of NAHR. Interestingly, our study shows that RET induces a higher frequency of TDDs than NAHR, at least when considering DNA transposons. This observation is not inconsistent with reported frequencies of NAHR and RET. Studies in human and Arabidopsis reported NAHR between two non-allelic repeats occurs at a frequency of approximately $10^{-4}$ to $10^{-6}$ (Molinier et al. 2004; Lam and Jeffreys 2006; Lam and Jeffreys 2007; Turner et al. 2008); whereas RET can occur at a much higher frequency, approximately $10^{-3}$ when two transposons are less than 1 kb apart (Zhang et al. 2013). Moreover, RET is catalyzed directly by transposase and is not dependent on meiosis-specific events such as chromosome pairing or the expression of specific recombination enzymes. Thus RET can occur during somatic cell divisions, and in plants the resulting duplications can be included in the reproductive cell lineages and be transmitted to progeny (Zhang et al. 2013). Finally, NAHR between TE sequences may be less likely due to their frequent hypermethylation, which may reduce recombination (Yelina et al. 2015).

In plants, previous research has shown that RET of hAT family transposons could induce genomic rearrangements, but less is known about other DNA transposons. In our study, we found that other DNA transposons, including Tc1/Mariner, Mutator and PIF/Harbinger, could also use induce duplications via the RET mechanism, suggesting that RET may have broader impact on plant genome evolution. We did not find any TI-TDDs induced by CACTA transposon family, possibly due to the relatively low number of CACTA MITEs in the TE database we used to query the STRAND program. Eventually, the program used in this study could be applied to
detect TDDs induced by other genomic elements such as retroelement sequences or other non-TE repeat sequences.

Methods

The bioinformatics pipeline

The STRAND program to detect DNA transposon-induced TDDs is a UNIX pipeline written with Perl scripts. This computational program takes the known transposon sequence(s) and a given genome sequence as input, and outputs a text that lists all transposon-induced tandem direct duplications and visualization graphs showing the structure of each case. All the outputs used for downstream analysis were generated by the default parameters. However, the program also has 10 options that allow users to adjust parameters according to personal preference. Details of the options can be found at the package (Supplemental information). The algorithm implemented by the program is presented in Figure 3.

Databases

The genome sequences of 22 plant species used in this study were obtained from the EnsemblGenomes (ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/) and other genome websites (Table S2). The MITE families across 22 plant species were identified by Chen and colleagues (Chen et al. 2014), and their consensus sequences can be downloaded from the P-MITE database (http://pmite.hzau.edu.cn/django/mite/). The maize Mo17 reference genome was downloaded from MaizeGDB (http://ftp.maizegdb.org/MaizeGDB/FTP/Mo17/2013_XinEtAl_PlantCell/).
Author Contributions

T.Z. and T.P. conceived and designed the research; T.Z. designed the algorithm; T.Z. and H.C. developed the STRAND pipeline; T.Z. and T.P. analyzed the data; and T.Z. and T.P. wrote the paper.

Reference


Table 1. DNA TE-induced TDDs identified in maize B73 genome.

<table>
<thead>
<tr>
<th>TE</th>
<th>TE family</th>
<th>TE_1</th>
<th>TE_2</th>
<th>Similarity (TEs)</th>
<th>Similarity (repeats)</th>
<th>Size of dup (bp)</th>
<th>Mechanism</th>
<th>Mo17</th>
</tr>
</thead>
<tbody>
<tr>
<td>SE220300125</td>
<td>hAT</td>
<td>F_TSD</td>
<td>F_noTSD</td>
<td>93%</td>
<td>94%</td>
<td>769</td>
<td>RET</td>
<td>Yes</td>
</tr>
<tr>
<td>SE220300214</td>
<td>hAT</td>
<td>F_TSD</td>
<td>F_noTSD</td>
<td>95%</td>
<td>96%</td>
<td>3637</td>
<td>RET</td>
<td>No</td>
</tr>
<tr>
<td>SE220300505</td>
<td>hAT</td>
<td>P</td>
<td>P</td>
<td>97%</td>
<td>94%</td>
<td>2704</td>
<td>RET</td>
<td>Yes</td>
</tr>
<tr>
<td>dhAT-Zm24</td>
<td>hAT</td>
<td>P</td>
<td>P</td>
<td>96%</td>
<td>99%</td>
<td>811</td>
<td>RET</td>
<td>No</td>
</tr>
<tr>
<td>SE220500133</td>
<td>PIF/Harbinger</td>
<td>F_TSD</td>
<td>F_TSD</td>
<td>100%</td>
<td>99%</td>
<td>2112</td>
<td>RET</td>
<td>Yes</td>
</tr>
<tr>
<td>SE220500193</td>
<td>PIF/Harbinger</td>
<td>P</td>
<td>P</td>
<td>97%</td>
<td>99%</td>
<td>1109</td>
<td>RET</td>
<td>No</td>
</tr>
<tr>
<td>SE220500208</td>
<td>PIF/Harbinger</td>
<td>P</td>
<td>P</td>
<td>96%</td>
<td>98%</td>
<td>1447</td>
<td>RET</td>
<td>No</td>
</tr>
<tr>
<td>SE220500012</td>
<td>PIF/Harbinger</td>
<td>F_TSD</td>
<td>F_TSD</td>
<td>98%</td>
<td>99%</td>
<td>1955</td>
<td>Unknown</td>
<td>Yes</td>
</tr>
</tbody>
</table>


Table 2. DNA TE-induced TDDs identified across the other 21 plant species.

<table>
<thead>
<tr>
<th>Species</th>
<th>Total # of dups</th>
<th># of 2-copy dups</th>
<th>RET</th>
<th>NAHR</th>
<th>Unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycine max</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Brassica rapa</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Prunus persica</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>5</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Cajanus cajan</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Malus domestica</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>4</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Setaria italicla</td>
<td>8</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Citrus sinensis</td>
<td>6</td>
<td>6</td>
<td>3</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>26</td>
<td>22</td>
<td>13</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>Total</td>
<td>70</td>
<td>48</td>
<td>22</td>
<td>11</td>
<td>15</td>
</tr>
</tbody>
</table>
Table S1. The genome size, MITE elements and TI-TDDs in 22 plant species examined.

<table>
<thead>
<tr>
<th>Species</th>
<th>Genome size (Mb)</th>
<th>Total length of MITE (Mb)</th>
<th># of 2-copy dups</th>
<th>RET</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabidopsis thaliana</td>
<td>119.67</td>
<td>0.85</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cucumis sativus</td>
<td>203.06</td>
<td>2.02</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Musa acuminata</td>
<td>472.96</td>
<td>2.22</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Arabidopsis lyrata</td>
<td>206.67</td>
<td>4.64</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Populus trichocarpa</td>
<td>417.14</td>
<td>7.49</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Prunus persica</td>
<td>227.25</td>
<td>8.84</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Fragaria vesca</td>
<td>206.89</td>
<td>8.97</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Citrus sinensis</td>
<td>327.94</td>
<td>11.35</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>Brachypodium distachyon</td>
<td>283.84</td>
<td>11.49</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Vitis vinifera</td>
<td>486.19</td>
<td>14.69</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Setaria italica</td>
<td>405.78</td>
<td>15.6</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Citrullus lanatus</td>
<td>353.47</td>
<td>19.55</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Medicago truncatula</td>
<td>307.48</td>
<td>25.24</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>Solanum lycopersicum</td>
<td>781.67</td>
<td>26.89</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Glycine max</td>
<td>973.34</td>
<td>27.69</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Sorghum bicolor</td>
<td>738.58</td>
<td>29.63</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Cajanus cajan</td>
<td>605.78</td>
<td>31.06</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>373.25</td>
<td>37.27</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Solanum tuberosum</td>
<td>797.83</td>
<td>38.65</td>
<td>22</td>
<td>13</td>
</tr>
<tr>
<td>Zea mays</td>
<td>2058.58</td>
<td>40.36</td>
<td>14</td>
<td>7</td>
</tr>
<tr>
<td>Malus domestica</td>
<td>881.28</td>
<td>44.63</td>
<td>2</td>
<td>2</td>
</tr>
</tbody>
</table>
Table S2. Website sources of genome sequences of 22 plant species used in this study.

<table>
<thead>
<tr>
<th>Species</th>
<th>Website</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Cucumis sativus</em></td>
<td>ftp://www.icugi.org/pub/genome/cucumber/PI183967/cucumber_PI183967_genome.fa.gz</td>
</tr>
<tr>
<td><em>Prunus persica</em></td>
<td>ftp://ftp.ensemblgenomes.org/pub/plants/release-22/fasta/prunus_persica/dna/Prunus_persica.GCA_000346465.1.22.dna.genome.fa.gz</td>
</tr>
<tr>
<td><em>Fragaria vesca</em></td>
<td>ftp://ftp.bioinfo.wsu.edu/species/Fragaria_vesca/Fvesca-genome.v1.1/assembly/fvesca_v1.1_pseudo.fna.gz</td>
</tr>
<tr>
<td><em>Citrus sinensis</em></td>
<td><a href="http://citrus.hzau.edu.cn/orange/download/csi.chromosome.fa.tar.gz">http://citrus.hzau.edu.cn/orange/download/csi.chromosome.fa.tar.gz</a></td>
</tr>
<tr>
<td><em>Citrullus lanatus</em></td>
<td>ftp://www.icugi.org/pub/genome/watermelon/97103/v1/watermelon_v1.genome.gz</td>
</tr>
<tr>
<td><em>Cajanus cajan</em></td>
<td><a href="http://www.icrisat.org/gt-bt/iipg/genomedata.zip">http://www.icrisat.org/gt-bt/iipg/genomedata.zip</a></td>
</tr>
<tr>
<td><em>Malus domestica</em></td>
<td><a href="http://www.rosaceae.org/sites/default/files/apple_genome/Malus_x_domestica.v1.0-primary.pseudo.fa.gz">http://www.rosaceae.org/sites/default/files/apple_genome/Malus_x_domestica.v1.0-primary.pseudo.fa.gz</a></td>
</tr>
</tbody>
</table>
FIGURES and LEGENDS

**A**

The two black lines represent two sister chromatids after DNA replication. Red lines with arrowhead(s) indicate DNA transposon insertions, and the open and solid arrowheads indicate the 3' and 5' ends, respectively, of transposon. The blue/green bars indicate the polymorphisms between the left and right TE. (A) DNA transposase (black arrows) recognizes and cleaves the 3' end of left TE and the 5' end of right TE on the lower chromatid. (B) The Inter-Transposon Segment is circularized by NHEJ at site x to form a small acentric circle which is presumably lost. (C) Transposon ends insert into the upper sister chromatid at a proximal site (a/b). Dotted lines represent the insertion of the two termini into the a/b site on the sister chromatid. (D) The 5' end of the right-hand TE joins to a, and the 3' end of the left-hand TE joins to b. This generates a tandem direct duplication (lower chromatid, shaded arrows) and a reciprocal deletion (upper chromatid).

**Figure 1. RET generates tandem direct duplication and reciprocal deletion.**
Figure 2. TDD generated by NAHR and its difference from that induced by RET.

(A) Non-Allelic Homologous Recombination occurs between nearby mis-paired TEs. The TEs are largely similar but have sequence polymorphisms as indicated by the blue and green bars. Following NAHR, the sequence between the original TE insertions is duplicated (shaded arrows), and the duplications are separated by a recombinant TE containing SNPs (green and blue boxes) from both parental TE copies.

(B) TDD induced by NAHR has polymorphic parental TE copies located precisely at the proximal and distal duplication endpoints, and a recombinant TE copy located between the duplicated segments. RET produces duplications which include identical TE copies. A third potentially polymorphic TE copy is located some variable distance nearby, but is not positioned precisely at a duplication endpoint as for NAHR-derived duplications.
Input: Genome DB

1

BLAST

2

Left flanking (100bp) Right flanking (100bp)

3

Selection 1: <= 50 kb Selection 2: only one flanking matched

4

Identify duplication border

5

Global alignment: >60% coverage

6

Manually check for TSDs and the third closest TE

Output: DNA TE associated TDDs
Figure 3. Algorithm used to identify DNA TE-induced TDDs. DNA transposons are represented by red lines with arrowheads. The solid and open arrowheads indicate the 5' and 3' TE ends, respectively. In step 1, a transposon sequence is used as a BLAST query to search a genome for related TE family members. In step 2, two 100 bp flanking sequences (left and right, blue and green lines, respectively) of each family member are extracted and compared. In step 3, two transposons less than 50 kb apart and having only one matched flanking sequence are selected. The red ovals indicate non homology of the 100 bp left flanking sequences. Step 4 identifies the endpoint of the candidate duplication. The non-matching 100 bp TE flanking sequence (blue) is used in a BLAST query to identify a homologous sequence on the other side of the TE. In step 5, an alignment is performed between the putative duplication segments demarcated by the sequences indicated in green and blue; segments with an overall alignment coverage of >60% are returned by STRAND as candidate TI-TDDs. In step 6, candidate duplications are manually checked as described in the text. The algorithm used to detect duplications induced by partial TE elements is similar, but an additional constraint is placed on step 3: the matched flanking sequence should be on the correct side of the TE copy, i.e. adjacent to the end lacking an intact TE terminus.
Figure 4. An example of RET-induced TDD identified by STRAND. Once a duplication is identified, STRAND will output 1) a text that describes the duplication position and features, and 2) a graphic displaying the duplication structure (A). This duplication is associated with a hAT transposon termed SE220300125 located on the maize chromosome 1. Red rectangles represent TEs, purple rectangles indicate the duplicated segments, and numbers indicate the chromosome coordinates of each feature. (B). Results of manual analysis of the duplication fine structure.

Large shaded arrows indicate TDDs, smaller red arrows indicate TEs, and purple bars indicate the non-TE duplicated segments. The transposon terminal sequences are shown above, and 8 bp flanking sequences are shown below. Note that the leading TE copy (on right) has identical 8 bp TSDs, while the internal TE copy is flanked on the far side by the same TSD as the leading TE copy, while the near side has a non-matching 8 bp flanking sequence which is produced by the RET insertion into the sister chromatid.
Figure 5. An example of NAHR-induced TDD identified by STRAND.

A). The graphical output of the hAT transposon SE380300018-associated TDD in potato (Solanum tuberosum) generated by STRAND. The red rectangles represent transposable elements, light blue bars indicate duplicated sequences, and numbers indicate chromosome coordinates of sequence features.

B). TDD fine structure following manual analysis. Shaded arrows indicate TDDs, smaller red arrows indicate TEs, and light blue bars indicate the duplicated segments. The transposon terminal sequences and internal SNPs are shown above, and 8 bp flanking sequences are shown below. Note that the flanking TE copies have unique 8 bp TSDs and are polymorphic for internal SNPs. The central TE copy is a recombinant, as indicated by the presence of TSDs and SNPs derived from each of the parental copies.
Figure S1. Model of NAHR-induced TDD generated by crossover between two partial TE copies. Homologous recombination between non-allelic partial TEs lacking one transposon terminus generates tandem duplications (shaded arrows). The TE copies at the duplication endpoints will be incapable of excision due to the absence of one TE terminus.
CHAPTER 5. GENES AND SMALL RNA TRANSCRIPTS EXHIBIT DOSAGE-DEPENDENT EXPRESSION PATTERN IN MAIZE COPY-NUMBER ALTERATIONS

Tao Zuo, Jianbo Zhang, Andrew Lithio, Sudhansu Dash, David F. Weber, Roger Wise, Dan Nettleton and Thomas Peterson

Abstract

Copy-number alterations are widespread in animal and plant genomes, but our understanding of their mechanistic impact on phenotypic variation is still limited. Two contrasting general responses to changes in gene copy number have been described: 1) dosage effect, in which transcript levels change in proportion to gene copy number; and 2) dosage compensation, in which transcript levels remain stable despite copy number changes. In animals, copy-number alterations usually exhibit dosage effects, except for sex chromosome genes that tend to be dosage compensated. However, in maize and other plants, analyses using relatively large segmental aneuploids have found that dosage compensation is more prevalent. Here, we compared sibling maize plants that contain two, three and four doses of a 14.6 Mb segment of chromosome 1 that contains ~300 genes. Plants containing four doses of the duplicated segment are significantly shorter, have smaller ears, and flower later than normal siblings. Whereas, plants with three copies are intermediate for most traits tested. To assess the impact of the duplication on gene expression, we determined transcript levels by GeneChip and RNA sequencing methods. Statistical tests indicate that most expressed genes and unique small RNAs within the duplicated segment exhibit gene dosage-dependent transcript levels. Furthermore, 98 genes outside the duplicated region also display dosage-dependent expression, possibly via trans-acting dosage effects of gene(s) in the duplicated segment. These results indicate that
dosage effect is the predominant gene regulatory response in the segmental dosage series we tested.

**Significance**

Segmental duplications comprise a significant proportion of eukaryotic genomes. In plants, genes within small duplications (<100 kb) often exhibit dosage-dependent expression, whereas genes present in large duplications (>50 Mb) more likely retain a diploid expression level. Less is known about expression in moderately-sized duplications, or about the response of small RNAs to dosage change. To address these questions, we performed GeneChip and RNA sequencing on a dosage series in maize that varies in the copy number of a 14.6 Mb segment of chromosome 1. Results show that dosage-dependent expression is the most common response for both genes and unique small RNA transcripts. These findings provide new insight into the regulation of genes and small RNAs in response to dosage changes.

**Introduction**

DNA copy-number alterations, whether involving chromosome segments or entire chromosomes, can have dramatic phenotypic impacts. Aneuploidy, one kind of DNA copy-number alteration (Tang and Amon, 2013), results from changes of chromosome number, i.e., gain or loss of one or more entire chromosomes. In most animals aneuploidy is detrimental and may cause severe genetic disorders. In humans, the gain of a single extra copy of chromosome 21 (trisomy 21) causes Down Syndrome; most other aneuploidies result in severe developmental disorders and do not survive to term. For instance, most spontaneous abortions and developmental abnormalities in human are caused by aneuploidy (Hassold and Hunt, 2001). In addition, many human cancer cells are highly aneuploid, although the mutual causality of
aneuploidy and tumorigenesis is still unclear (Weaver and Cleveland, 2007). However, aneuploidy in plants is typically far less detrimental. Trisomics for each of the chromosomes have been recovered in several plant species, and even monosomics for all of the 10 maize chromosomes have been recovered and are viable (Weber, 1994).

In contrast to the severe effects of aneuploidy, other copy-number alterations such as segmental aneuploidy and copy number variations (CNV) often have milder phenotypic effects, which may facilitate their retention and possible accumulation in a population. CNV, involving segments of 1 kilobase (kb) to 1 megabase (Mb) in size (Tang and Amon, 2013), is widespread across many species including mammals (Iafrate et al., 2004a; Redon et al., 2006a; Sebat et al., 2004) and major crop plants (Jiao et al., 2012; Lai et al., 2010; McHale et al., 2012; Saintenac et al., 2011; Springer et al., 2009; Swanson-Wagner et al., 2010; Yu et al., 2011b; Zheng et al., 2011). For example, more than 10% of the human genome is composed of CNVs and segmental duplications with sizes ranging from a few kb to several Mb (Iafrate et al., 2004b; Redon et al., 2006b; Sebat et al., 2004b; Stankiewicz and Lupski, 2010). In rice, tandem arrayed genes (TAGs) account for up to 20% of the duplicated genes (Rizzon et al., 2006).

In addition to the effects of CNVs and segmental duplications on creating genomic diversity, CNVs and segmental duplications in eukaryotes can dramatically impact organismal phenotype. For example, gain or loss of one copy of human 1q21.1, a genomic region associated with developmental abnormalities (O'Donovan et al., 2008), confers a higher risk of mental disorders (Stefansson et al., 2009; Stefansson et al., 2008). In addition, microdeletion or microduplication of human 16p11.2 has an association with autism (Weiss et al., 2008). CNVs can be advantageous as well. For instance, changes in copy number of an amylase gene (AMY2B) greatly facilitated the adaptation of dogs to starch-rich diets associated with domestication.
(Axelsson et al., 2013). In plants, several important traits are directly impacted by CNVs (Cook et al., 2012; Diaz et al., 2012; Li et al., 2012; Maron et al., 2013; Xiao et al., 2008). For example, a recent study in rice showed that a tandem duplication that increases copy number of the GL7 locus has significant effects on grain length (Wang et al., 2015). However, despite the fact that CNVs and segmental duplications are pervasive and often have major effects, the question of precisely how copy-number alterations affect gene expression and contribute to phenotypic diversity remains largely unanswered.

Dosage compensation and dosage sensitivity are two major responses of gene expression to changes in DNA dosage (Birchler, 2010; Birchler and Veitia, 2012; Guo et al., 1996). Several recent studies in human and plants showed that the change in the expression of genes with altered copy number is a causative factor for the phenotypic impact of CNVs (Cook et al., 2012; Golzio et al., 2012; Li et al., 2012b). Specifically, elevated expression (mRNA and/or protein levels) of genes located within the trisomy or segmental trisomy region is largely responsible for the aneuploid syndromes (Huettel et al., 2008; Kahlem et al., 2004; Pavelka et al., 2010; Stingele et al., 2012; Vacik et al., 2005; Williams et al., 2008). These results suggest that genes are expressed in proportion to their dosage. However, expression of genes located on the X chromosome of human, Caenorhabditis elegans or Drosophila is typically dosage compensated, though different mechanisms are involved in maintaining gene expression balance between male and female in these species. In addition to sex chromosomal genes, the expression of autosomal genes in Drosophila could also be dosage compensated (Sun et al., 2013). However, the situation may be more complex in plants. For example, the maize alcohol dehydrogenase (Adh1) gene showed a dosage effect in small segmental aneuploids (Birchler, 1981); when tested in larger segmental aneuploids, however, the same Adh1 gene as well as other linked genes exhibited
equivalent levels of expression when compared to diploids (Birchler, 1979; Guo and Birchler, 1994). These results suggest that the size of the duplicated sequence may affect the regulation of the included genes. In addition, genes with altered copy numbers could exert trans-effects on the expression of genes whose dosage is constant (Guo and Birchler, 1994; Makarevitch et al., 2008). Maize is one of the most diverse plant species (Buckler et al., 2006), and variation in maize CNVs is prevalent not only among modern inbred lines (Jiao et al., 2012; Lai et al., 2010; Springer et al., 2009), but also between domesticated maize and its wild progenitor (Chia et al., 2012; Hufford et al., 2012; Swanson-Wagner et al., 2010). However, less is known about the association between phenotypic divergence and CNV. Besides affecting expression of protein coding genes, CNVs could potentially affect expression of noncoding RNAs which have been recently recognized as having important roles in regulating gene expression and maintaining genomic stability. How small RNAs respond to copy-number alterations is an interesting and as-yet unexplored question.

We have shown previously that alternative transposition involving two Ac/Ds transposon termini can induce genomic rearrangement including inversion, deletion, duplication and translocation (Yu et al., 2011a; Yu et al., 2010; Zhang and Peterson, 1999; Zhang and Peterson, 2004; Zhang and Peterson, 2005). We isolated a number of such duplications and deletions, and conducted phenotypic and transcriptional analysis on one case (p1-ww714) with a tandem inverted duplication on chromosome 1S. The region duplicated in p1-ww714 is 14.6 Mb in size and is predicted to contain approximately 300 gene models in the maize filtered gene set (ZmB73_5b_FGS). Using this segmental duplication in combination with a normal chromosome 1, we produced sibling plants containing 2, 3 and 4 copies of the affected region. We then conducted phenotypic and transcriptional studies on this segmental dosage series. We
implemented both high throughput sequencing (RNA-Seq and small RNA-Seq) and GeneChip (new Affymetrix Maize WT 100K array) approaches to study the relationship between CNV and transcript accumulation. Our study provides valuable insights into disclosing CNV effects in maize and understanding how gene expression responds to copy number change.

Results

Identification of a 14.6 Mb tandem inverted duplication on maize chromosome 1

*p1-ww714* is an allele of the maize *p1* gene that encodes a Myb-like transcription factor which regulates maize kernel pericarp and cob pigmentation (Grotewold *et al.*, 1994). The *p1-ww714* allele was isolated in a screen for inverted duplications derived from the progenitor allele *p1-vv9D9A* that contains insertions of *Ac/fAc* transposable elements. Previous research has shown that the paired *Ac/fAc* elements in *p1-vv9D9A* can undergo aberrant transposition reactions termed Sister Chromatid Transposition (SCT) to generate inverted duplications that begin at the *p1* gene and extend variable lengths proximally (Zhang and Peterson, 1999). The structure of *p1-ww714* was tested by PCR of genomic DNA using a series of primers diagnostic for the presence of inverted duplications (*Fig. S1A*). These PCR tests confirm that the progenitor *p1-vv9D9A* structure was disrupted by the presence of an inverted duplication in *p1-ww714* (*Fig. S1B*). The SCT model predicts that the duplication endpoint will be adjacent to the 3’ end of the *fAc* element; therefore we used Inverse PCR (I-PCR) to isolate this junction sequence and used it in BLAST searches of the maize reference genome. The endpoint junction is located at position 62.7 Mb on maize chromosome 1, a distance of 14.6 Mb from the *p1* locus at position 48.1 Mb. Taken together, these results indicate that *p1-ww714* allele contains a ~14.6 Mb inverted duplication in the short arm of maize chromosome 1 from 48.1 Mb to 62.7 Mb, representing about 5% of chromosome 1.
Phenotypic impact of the 14.6 Mb duplication

Initial observations found that plants that carry *p1-ww714* allele exhibit significant phenotypic differences compared to their progenitor. To better study the phenotypic effect of this copy-number alteration, the original *p1-ww714* allele was backcrossed with the maize inbred B73 for five generations to achieve a near-B73 genetic background. In the succeeding generation, heterozygous *p1-ww714*/B73 plants were self-pollinated to generate sibling plants of the genotypes B73/B73, *p1-ww714*/B73 and *p1-ww714*/p1-ww714, which carry 2, 3 and 4 copies of the 14.6 Mb segment, respectively. These plants were grown to maturity at Iowa State University in summer 2012, in a randomized complete block design with three replications.

Homozygous *p1-ww714*/p1-ww714 plants exhibited delayed flowering time compared to B73/B73 (Table S1) and *p1-ww714*/B73. *p1-ww714*/p1-ww714 flowered approximately 10 days later than B73, while *p1-ww714*/B73 flowering time was intermediate: about 6 days later than B73/B73 but 4 days earlier than *p1-ww714*/p1-ww714 (Table S1). Additional phenotypic impacts on plant height and ear length were observed: homozygous *p1-ww714* plants were 46 cm and 43 cm shorter than homozygous B73 plants and heterozygous *p1-ww714*/B73 plants, respectively (Fig. 1B) with a *p*-value less than 0.0001 (Table S1). However, the average heights of *p1-ww714*/B73 and B73/B73 plants were not significantly different. The average ear length of *p1-ww714*/p1-ww714 was 4.1 cm shorter than that of sibling B73/B73 (*p*-value < 0.0001) and 1.7 cm shorter than that of sibling *p1-ww714*/B73 (*p*-value = 0.0007), while ear length of *p1-ww714*/B73 was 2.4 cm (*p*-value = 0.0001) shorter than that of B73/B73 (Fig. 1 C and D; Table S1).

Phenotypic measurements were repeated using sibling plants grown during 2012 winter nursery in Rancagua, Chile with two replications. Similarly, homozygous *p1-ww714* plants were
significantly shorter than B73/B73 and \(p1\)-\(ww714\)/\(B73\) plants; while ears of \(p1\)-\(ww714\)/\(p1\)-\(ww714\) were on average 3.9 cm (\(p\)-value = 0.0061) and 2.2 cm (\(p\)-value = 0.0159) shorter than that of sibling B73/B73 and \(p1\)-\(ww714\)/B73, respectively (Table S1). The average height of \(p1\)-\(ww714\)/B73 plants was not significantly different from that of B73/B73. In summary, \(p1\)-\(ww714\)/\(p1\)-\(ww714\) plants exhibited delayed development, shorter plant stature, and shorter ear length. Whereas, \(p1\)-\(ww714\)/B73 plants were intermediate for flowering time and ear length. Finally, ~15% of homozygous \(p1\)-\(ww714\) plants exhibited severe development defects, such as extreme dwarfing, absent or barren tassels, and other morphological aberrations. Because these aberrant plants did not appear to have a common mutant phenotype, we suggest that they may represent the heterogeneous products of genome instability, such as chromosome mis-segregation, induced by the presence of the large segmental duplication. Whatever the cause of these exceptional plants, they were excluded from the phenotypic and molecular analyses since they were not representative of the majority of homozygous \(p1\)-\(ww714\) plants.

**Differential transcript abundance in duplication genotypes**

The phenotypic differences observed among copy-number variants suggest that some differences in transcript levels might be induced by the altered copy number of the 14.6 Mb segment. To identify genes that are differentially expressed as a result of the 14.6 Mb duplication, we examined transcript levels by high-throughput sequencing of RNA samples from sibling plants of the \(p1\)-\(ww714\)/\(p1\)-\(ww714\), B73/B73 and \(p1\)-\(ww714\)/B73 genotypes. Nine libraries from 14-day old whole seedlings of each of the three genotypes, with three biological replications each, were constructed and then sequenced on an Illumina HiSeq-2000 system. In total, we generated 149.7 million 100 bp raw reads; 76.2% of them were uniquely mapped to the B73 reference genome (RefGen_V2). The number of uniquely mapped reads per sample ranged
from 9.5 million to 15.2 million (Table S2). The Pearson correlations of read counts within genotypes were greater than 98%. A series of bioinformatics tools and statistical tests were then applied to analyze the data and identify differentially expressed genes (DEGs) (see Methods). We found 253 of 23730 expressed genes were differentially expressed between p1-ww714/p1-ww714 and B73/B73 in the 14-day old seedling tissues: 125 of 253 (49%) DEGs are located within the 14.6 Mb chromosomal region, 51 DEGs are within the remainder of chromosome 1, and the remaining 77 DEGs are distributed in the other nine chromosomes (Fig. 2A). These results indicate that genes within the duplicated segment are clearly overrepresented among all DEGs detected. Among 125 DEGs located in the duplicated segment, the great majority (118/125) were over-expressed in p1-ww714/p1-ww714 compared with B73/B73 sibling plants. Whereas, among the 128 DEGs located outside the duplicated segment, less than a third (41/128) of them were over-expressed. We also compared the transcript data of p1-ww714/B73 (three copies) with that of sibling homozygous B73 (two copies) and p1-ww714 (four copies) plants. Interestingly, only one gene was identified as differentially expressed in the comparison between p1-ww714/B73 and p1-ww714/p1-ww714, and only 10 DEGs were identified in the comparison between p1-ww714/B73 and B73/B73. Most likely our statistical analysis was not sufficiently sensitive to detect expression changes caused by single-copy differences.

**Dosage-dependent expression of genes in duplicated region**

As a complement to statistical identification of individual DEGs, comparisons of transcript expression ratios between genotypes can also elucidate copy number effects, especially for those genes that exhibit smaller changes in dose-dependent expression. Consequently, we computed transcript abundance ratios among the three tested genotypes derived from RNA-Seq
data and plotted the log2 fold changes in expression with corresponding genome position. We found that the majority of maize chromosome 1 genes (Fig. 2B) as well as genes located in other chromosomes (Fig. S2) have similar expression levels across the three tested genotypes, because the expression ratio distributions are centered on or near zero. The striking exception is the duplicated segment located between 48.1 Mb and 62.7 Mb on maize chromosome 1. The \textit{p1-}
\textit{ww714/p1-ww714, p1-ww714/B73} and \textit{B73/B73} genotypes carry 4, 3 and 2 copies of this 14.6 Mb segment, respectively. The log2 ratio of gene transcripts located in this region is clearly shifted above 0 for all three pair-wise comparisons; whereas, the ratios of transcript levels for other regions of maize chromosome 1 are not similarly skewed (Fig. 2B). Further analysis of this 14.6 Mb region revealed that more than 80\% of genes have a log2 ratio value greater than 0 in the genotype containing additional copie(s) of this 14.6 Mb segment; i.e. most genes within the altered region show increased transcript number as copy-number is increased. The distribution of log2 fold change in the comparison between \textit{p1-ww714/p1-ww714} and \textit{B73/B73} is very closely centered near 1.0, which is the log ratio of copy number between homozygous \textit{p1-ww714} (4 copies) and \textit{B73} (2 copies). Similarly, the distributions of log2 fold change obtained from the other two comparisons are centered close to their corresponding log2 ratios of gene copy numbers; specifically, the gene copy number ratios (log2) of \textit{p1-ww714/p1-ww714} vs. \textit{B73/B73, p1-ww714/B73} vs. \textit{B73/B73} and \textit{p1-ww714/p1-ww714} vs. \textit{p1-ww714/B73} are 1.0, 0.585 and 0.415, respectively; and the median transcript abundance ratios (log2) are 0.91, 0.52 and 0.38, respectively. The direct correlation between gene transcript levels and copy-number suggests that most genes located within the duplicated region are expressed in proportion to their dosage.

To further study the effects of copy-number variation on gene expression, a new Affymetrix Maize 100 format whole transcriptome (WT) array containing 103,262 probe sets
was used to obtain global gene expression profiling of copy-number variant genotypes. The same nine RNA samples used for RNA-Seq were used to prepare probes for hybridization with Affymetrix Maize WT 100K arrays. Hybridization data were processed using the PLEXdb pipeline (Dash et al., 2012). Similar to the results from RNA-Seq, most genes within the duplicated segment were inclined to increase expression, as the log2 fold change distributions of each comparison are clearly centered above 0 (Fig. S3). Although transcript abundance ratios are somewhat smaller than their corresponding copy number ratios, they are still positively and proportionally correlated with gene copy number. In contrast, we did not observe any similar regional shifts in transcript levels outside the duplicated segment, either on the same or other chromosomes (Fig. S3). These GeneChip data further support the conclusion that many transcripts mapping within the duplicated chromosome region are differentially expressed in accordance with the gene dosage in each genotype.

**Trans-acting dosage effect for genes unlinked with duplicated region**

In addition to dosage effects on genes within the 14.6 Mb duplicated segment, trans-acting effects could alter the expression of unlinked target genes (Li et al., 2013). RNA-Seq experiments identified 128 DEGs located outside the duplicated segment when comparing homozygous p1-ww714 and B73. Among 128 DEGs, 41 are over-expressed while 87 are under-expressed in p1-ww714/p1-ww714, suggesting the existence of both activating and repressing trans-effects. To test whether trans-acting dosage effects (Cooper and Birchler, 2001; Guo and Birchler, 1994) are implicated, transcript levels of these 128 genes across three copy-number variants were further compared. We hypothesized that if the expression of these 128 genes are in fact subject to trans-acting dosage effects, we should observe proportional expression levels of these 128 genes; i.e., expression of these genes in the heterozygote p1-ww714/B73 should be
intermediate between that of $p1$-ww714 and B73. Examination of the log2 fold changes in expression between $p1$-ww714/B73 and B73, and $p1$-ww714 and B73, shows that all 128 genes exhibit consistent increasing or decreasing expression as the copy number of the 14.6 Mb segment changes (Fig. 2C). These results suggest that both positive and negative trans-acting dosage effects do in fact impact the expression of genes outside the duplicated region, although specific statistical analyses are required to test this conclusion (below).

**Statistical tests of cis- and trans-dosage effects**

As described above, we observed dosage-dependent expression of most genes located in the duplicated region, as well as trans-acting dosage effects for some genes outside the duplicated region. We then developed and applied statistical approaches (see Methods) to rigorously test the hypothesis of dosage effect by virtue of transcript level of genes across all three genotypes. For each gene, we compared three generalized linear models, henceforth referred to as M0, M1, and M2 in increasing order of complexity. The log of the mean read count for each gene was assumed to be a linear function of gene-specific parameters plus a sample-specific normalization factor, determined for each sample by the log of the 0.75 quantile of sample-specific read counts (Bullard et al., 2010). We found no evidence that the full model (M2) for the means is needed. In fact, the log-linear model (M1) appears to be sufficient (see Methods). According to the M1 model, the dosage compensation hypothesis is represented with the equation $\beta_1=0$. The non-dosage compensation effect (dosage effect or others) is represented with the equation $\beta_1 \neq 0$, whereas the dosage effect is represented with $\beta_1=1$. We found that 135 of 212 (63.7%) genes in the duplicated region have $\beta_1$ significantly different from 0 when controlling the false discovery rate (FDR) at 5%, indicating that these genes do not exhibit dosage compensation. The distribution of the $\beta_1$ estimate for each gene in the duplicated region is
depicted as a histogram (see Fig. 3A), where the center point is significantly different from 0 and in fact is close to 1. To further estimate the number of genes within the duplicated region that satisfy the dosage effect hypothesis ($\beta_1=1$), we constructed 95% confidence intervals (CIs) for the $\beta_1$ value for each gene. We found that the $\beta_1$ value of 108 out of 135 genes was not significantly different from the value “1” (Fig. 3C), suggesting that expression levels of 50.9% (108/212) of genes in the duplicated region were consistent with the positive dosage effect hypothesis. Expression levels of three of 135 genes exhibited $\beta_1$ estimates not significantly differently from -1, consistent with regulation by a negative dosage effect. Expression levels of the remaining 11.3% (24/212) of duplicated genes could not be explained by either the dosage compensation or dosage effect model (Fig. 3C).

A similar statistical test was performed for the 23,518 genes located outside the 14.6 Mb duplicated region. In contrast, only 154 of the 23,518 (0.65%) genes had $\beta_1$ estimates that are significantly different from 0 (FDR $\leq$ 0.05), suggesting that the expression of the majority of un-duplicated genes is not altered by the 14.6 Mb duplication in chromosome 1. The distribution of the $\beta_1$ estimate for each gene that is not in the duplicated region is clearly centered near 0 (Fig. 3B). Results from the calculated 95% CIs show that 24% (37/154) of genes were positively regulated by trans-acting dosage effects, whereas 40% (61/154) of genes were subject to negative trans-acting dosage effects (Fig. 3D). The changes of RNA level of the remaining 66 genes cannot be explained solely by trans-acting dosage effects.

**Parallel verification of DEGs in an independent duplication allele**

In addition to $p1$-ww714, we also isolated a number of shorter inverted duplications generated by Ac/Ds-induced SCT. One such allele, $p1$-wwid1, carries a 3.3 Mb duplicated segment (Zhang and Peterson, 1999), which is completely overlapped by the 14.6 Mb
duplication in $pl$-ww714. The 3.3 Mb region (chr1:48.1 Mb to 51.4 Mb) contains 42 annotated genes, 27 of which were identified as DEGs between homozygous $pl$-ww714 and B73. If expression of these 27 DEGs were in fact regulated by gene dosage, we would expect these genes should also be up-regulated in the $pl$-wwid1 allele. We randomly chose five of the 27 total DEGs for testing, and one non-DEG located outside the duplicated region as a control.

Homozygous $pl$-wwid1 and B73 plants were grown under the same conditions as for $pl$-ww714/$pl$-ww714, and transcript levels were determined by quantitative Real-Time PCR, using tissues from the same development stage as used previously for the $pl$-ww714 series. We found that all five DEGs were over-expressed in $pl$-wwid1/$pl$-wwid1 compared to B73/B73, while expression of the non-DEG control was not significantly different between the two genotypes (Fig. S4). These results utilizing an independent segmental duplication further support the conclusions discussed above.

**Small RNA dosage effect**

The additional copy of the 14.6 Mb fragment in $pl$-ww714 will not only increase the copy number of protein coding genes, but will also double the dose of inter-genic regions, transposons and small RNA genes. To investigate the effects of dosage on small RNA expression, we performed small RNA sequencing (sRNA-Seq) of $pl$-ww714/$pl$-ww714, $pl$-ww714/B73 and B73/B73 genotypes using the identical 14-day seedling tissues described above from six pooled sibling plants. Our sRNA-Seq data included 31.3 million reads for $pl$-ww714/$pl$-ww714, 33.5 million reads for $pl$-ww714/B73 and 65.5 million reads for B73/B73. We selected reads that could be uniquely and perfectly mapped to the maize B73 reference genome because the origin of these uniquely-mapped reads is more tractable. For each genotype approximately 30% of sRNA-Seq reads mapped to a single site in the maize B73 reference
The genome was subdivided into 100-kb windows across the 10 maize chromosomes, and the total number of uniquely mapped reads per 100-kb window was calculated. The levels of small RNAs in homozygous p1-ww714 and B73 are very similar in most genomic positions (Fig. S5), although p1-ww714/p1-ww714 has some dispersed regions with apparently lower expression levels, possibly due to residual sequence polymorphisms (un-replaced background). However, within chromosome 1 where the duplication is located, a large (~40 Mb) region shows substantial variation in small RNA levels between p1-ww714/p1-ww714 and B73/B73 (Fig. 4A), p1-ww714/p1-ww714 and p1-ww714/B73, as well as p1-ww714/B73 and B73/B73 (Fig. S6). We believe that many of these apparent differences in small RNA levels between p1-ww714 and B73 are actually caused by residual sequence polymorphisms in the p1-ww714 chromosome that reduce the efficiency of mapping the p1-ww714 reads to the B73 reference genome. Most likely this ~40 Mb region was not replaced during backcross to B73 due to suppression of recombination by the inverted duplication.

Therefore, instead of comparing p1-ww714 and B73, we compared p1-ww714 with two other smaller duplication alleles (p1-wwid1 and p1-wwid4). Both p1-wwid1 (id1) and p1-wwid4 (id4) were also generated by SCT from p1-vv9D9A, the same progenitor of p1-ww714; however, id1 and id4 carry smaller duplications of ~3.3 Mb and ~400 kb, respectively, which are completely overlapped by the 14.6 Mb duplication in p1-ww714. These alleles were also backcrossed to B73 in parallel with p1-ww714. During the backcross program the p1 locus was selectively retained in each generation; hence the genome region near p1 has likely been retained in each of the three duplication stocks, so comparisons among these alleles will provide a better test of small RNA expression levels. Our results show that the 3.3 Mb duplicated region in id1 displays a two-fold increase in small RNA transcripts relative to p1-vv9D9A (Fig. S7).
results (Fig. 4B) show a region of doubled small RNA expression in \textit{p1-ww714} when compared with \textit{id1}. This region of over-expression extends from \(\sim 52.4 \text{ Mb} \) to \(62.7 \text{ Mb} \) in chr.1, which is exactly the extent of the duplication in \textit{p1-ww714} that is not contained within \textit{id1}. Whereas, the 3.3 Mb region (48.1 to 52.4 Mb) which is duplicated in both \textit{p1-ww714} and \textit{id1} shows equal levels of small RNAs. Similar pair-wise comparisons were also made between homozygous \textit{p1-ww714} and \textit{id4}, and \textit{id1} and \textit{id4} (Fig. 4 \textbf{C and D}). A clear positive dosage effect of the 3.3 Mb region is now visible for both comparisons since both \textit{p1-ww714} and \textit{id1} carry an extra copy of this 3.3 Mb region with respect to \textit{id4}.

We also performed small RNA sequencing for some alleles (B73, \textit{p1-vv9D9A}, \textit{id1} and \textit{id4}) from 14-day old root tissues. The comparisons of small RNA expression from the root tissue are very similar as those from the seedling tissue: compared with \textit{p1-vv9D9A} and \textit{id4}, the \textit{id1} allele contains a 3.3 Mb region with overexpressed small RNAs that is completely coincident with the extent of the duplicated segment (Fig. S8). This indicates the consistency of dosage effect on small RNA transcripts between the leaf and root tissue.

\textbf{Discussion}

Dosage effect and dosage compensation are two contrasting mechanisms of gene regulation in response to copy-number alterations. Aneuploids in human and mouse frequently exhibit increased expression of genes located in a trisomy region, suggesting that dosage effect is relatively more common in these species. Dosage effect is also reported for relatively small duplications in plants (Cook \textit{et al.}, 2012; Wang \textit{et al.}, 2015; Xiao \textit{et al.}, 2008). However, studies of large segmental duplications in maize suggest that dosage compensation is the more prevalent mechanism (Siegel \textit{et al.}, 2012).
Here, we constructed a *de novo* series of copy-number variants in maize and compared levels of transcripts of protein coding genes and small RNAs by RNA-Seq and GeneChip. Results of RNA-Seq showed that 58.7% of protein-coding genes (125/213) within the duplicated region exhibited significantly different expression levels when comparing genotypes containing two versus four copies. However, the number of identified DEGs dropped dramatically when comparisons were made between genotypes that differ by a single copy. For example, only 1 gene was differentially expressed between \textit{p1}\textit{-ww714/p1}\textit{-ww714} and \textit{p1}\textit{-ww714/B73} (4 vs. 3 copies), and 10 genes were differentially expressed between \textit{p1}\textit{-ww714/B73} and \textit{B73/B73} (3 vs. 2 copies). Nevertheless, transcript levels of most genes within the altered region in \textit{p1}\textit{-ww714/p1}\textit{-ww714} and \textit{p1}\textit{-ww714/B73} were elevated compared to B73, even though some of the increments were not large enough to be classified as statistically significant. In order to make the conclusion of dosage effect and dosage compensation more reliable, we developed and applied a statistical procedure to test dosage effect and dosage compensation by integrating transcript data of all three genotypes. The expression of ~63.7% (135/212) of the duplicated genes were clearly not dosage compensated; 52.4% showed a proportional gene dosage effect, while the remaining 11.3% had transcript levels intermediate between the predictions of the dosage compensation and dosage effect models.

In contrast, previous studies in maize suggested that dosage compensation was the major gene regulatory response to copy-number changes (Birchler, 1979; Guo and Birchler, 1994). These conclusions were based on analysis of selected genes within large segmental aneuploids, e.g. the long arm of maize chromosome 1, produced using maize B-A translocation chromosomes. Possibly, the larger size of duplication may have influenced the gene expression pattern, since the same \textit{Adh1} gene did exhibit dosage-dependent expression in a shorter
duplication (Birchler, 1981). More recently, Makarevitch and colleagues performed large scale RNA profiling of maize segmental aneuploid stocks using the Affymetrix maize 17K GeneChip and reported both dosage compensation and dosage-dependent expression (Makarevitch and Harris, 2010; Makarevitch et al., 2008). They found about 50% or more of genes present in three copies were dosage compensated when compared to the expression of the same genes present in two copies, suggesting that dosage compensation is more common than dosage-dependent expression. Because the stocks they used contained duplications of nearly the entire short arm of maize chromosome 5 (~50 Mb), a region considerably larger than our ~15 Mb duplication, it is possible that the results are impacted by the size of duplication as noted above for Adh1. On the other hand, the number of dosage-dependent genes identified in Makarevitch et al. might be underestimated. The comparison of transcript levels was made between two genotypes (duplication-deficient vs. normal) that differ by a single copy (3 vs. 2 copies); hence dose-dependent changes in transcript abundance would be only 1.5 fold. Given this small difference, and considering the inherent high variability of gene expression, statistical methods may lack the power to reliably detect significant changes in expression between the tested genotypes. In our study, we used high-throughput sequencing to determine transcript levels in a segmental dosage series comprising three genotypes containing 2, 3 and 4 copies of the variant region. This approach provides greater statistical power to detect differences in gene expression. Interestingly, Makarevitch and colleagues found that qualitative changes in expression are common in their segmental aneuploidy material, and that some of them are directly responsible for some aspects of the aneuploid phenotype. In contrast, we did not detect any genes ectopically expressed in our experiments, despite the use of similar tissue and developmental stage in both studies.
Issues of Transcript Normalization

The comparisons of transcript levels employed in our study are based on the commonly-held assumption that total RNA levels per cell are similar among the tested genotypes. However, several recent studies reported two to three fold increased total RNA levels in certain tumor cells compared with normal cells (Lin et al., 2012; Nie et al., 2012). This observation calls into question the assumption of similar RNA expression levels, which could result in misinterpretation of GeneChip and/or RNA-Seq results if not correctly controlled (Loven et al., 2012). In our study, we found that genes within the duplicated region exhibited significantly different transcript levels, while the expression of most genes located outside the duplication was unchanged. Because these expressional differences are largely restricted to genes within the duplicated segment, our results cannot be readily explained as an artifact of global transcriptional amplification or repression.

Another recent paper reported that in Drosophila triple X metafemale flies, the expression of most genes located in the autosomal chromosomes was down-regulated by 1/3 due to increased X-chromosome number (Sun et al., 2013). This observed inverse relationship is called inverse dosage effect; the combination of inverse and gene dosage effects are proposed to result in dosage compensation (44). In our study, we found that expression of most genes within the varied chromosome segment is up-regulated, while the expression of genes outside the varied segment is largely unchanged. Considering the inverse dosage effect and the potential normalization issue proposed by Birchler (Birchler, 2010), one possible explanation of our results is that the duplicated genes were mostly dosage compensated, while genes outside the duplication were down-regulated in proportion to the change in dosage of the varied segment. In the present case the duplicated segment is 14.6 Mb, which represents only ~0.6% of the total...
maize genome; it seems unlikely that a 0.6% change in genome size would result in a global \(~1/3\) reduction in total RNA level per cell. Moreover, Guo and associates found that total RNA expression levels increase in proportion to ploidy in a maize ploidy series (Guo et al., 1996), and Stingele and colleagues reported that mRNA abundance is increased in trisomic and tetrasomic human cell lines (Stingele et al., 2012); these observations suggest that increased genome size would increase, rather than decrease, total RNA levels. Finally, we re-analyzed our data based on the methods implemented by Sun and associates (Sun et al., 2013), and found that dosage effects were still evident when the duplicated segment and other chromosome regions were treated separately. Therefore, we conclude that dosage effect, not dosage compensation, is the prevalent response in the maize copy-number alteration stocks we tested.

**Dosage effects on Non-coding RNAs**

Although the effects of copy-number alterations on mRNA and protein levels have been studied in several systems, much less is known about how copy-number alterations affect transcription of non-coding and intergenic sequences such as small RNAs and Transposable Elements (TEs). For example, \(~45\%\) of the human genome is composed of TE-derived sequences (Lander et al., 2001), whereas in maize the amount of TEs is as high as 85\% (Schnable et al., 2009). Hence, large duplications will include a substantial portion of TE sequences, and proper regulation (silencing) of these extra TEs is likely critical to maintain genome stability. In plants, small RNA-guided methylation plays an important role in deactivating transposons (Law and Jacobsen, 2010). We hypothesized that small RNAs targeted to the duplicated region would exhibit proportionally increased expression in order to silence the extra TE sequences. To test this hypothesis, we performed small RNA sequencing of our dosage series and analyzed the levels of small RNA transcripts produced from consecutive 100-kb
windows across the genome. When examining the levels of small RNA including reads that map at up to 100 different genome sites (multiply-mapped reads), there was little or no indication of over-expression of small RNA transcripts within the duplicated segment. This is not surprising considering the large number of other genomic loci outside the duplicated segment that can encode the bulk of multiply-mapped reads. However, analysis of the uniquely and perfectly mapped reads located within the duplicated segments indicated that many of these uniquely mapped small RNA transcript levels are increased in a dosage-dependent manner (Fig. 4). Whether these results extend to multi-copy TEs and other highly-repetitive heterochromatic sequences remains unknown.

Long non-coding RNAs (lncRNAs) are recently disclosed to play key roles in regulation of gene expression (Wilusz et al., 2009). A previous study identified 1704 high-confidence lncRNAs in the maize B73 reference genome (Li et al., 2014). Among them, 10 lncRNAs transcribed from the 14.6 Mb duplicated region are expressed in the seedling tissue we studied. We found that 8 of 10 lncRNAs are significantly over-expressed in the p1-ww714 background compared to B73 genotype, suggesting that, similar to protein coding genes, the majority of expressed lncRNAs in the duplicated region exhibit dosage-dependent expression.

*Trans-acting dosage effect*

The effects of aneuploidy on gene expression may be observed not only for genes with altered dosage, but also for genes unlinked to the varied chromosomal segment. This phenomenon, termed *trans*-acting dosage effect, is a mechanism to proportionally regulate genes outside the altered chromosomal segment (Guo and Birchler, 1994). Previous studies had revealed a small number of *trans*-effected genes. In our study, we provide further evidence through the identification of 34 genes up-regulated by positive *trans*-dosage effect and 64 genes
down-regulated by negative trans-acting dosage effect. Considering that the 14.6 Mb duplication studied here comprises approximately 0.5% of the total maize genome, these findings suggest that trans-acting dosage effects may be common in segmental aneuploid conditions. We hypothesize that the cumulative effects of both cis- and trans-acting dosage effects likely play important roles in producing the pronounced phenotypic effects of aneuploidy.

**Alternative transposition of Ac/Ds elements as a tool to induce maize CNVs.**

Like many other eukaryotes, the maize genome is large and complex, due to its high level of TEs (Schnable et al., 2009) and segmental duplications (Emrich et al., 2007; Schnable et al., 2009). In addition, striking differences in genome content and organization have been reported among maize inbred lines, including significant levels of copy number variation (CNV) and presence/absence variation (PAV). However, the functional impacts of segmental duplications in maize are still unclear. Although most examples of extensive copy number change are deleterious, several recent studies in yeast revealed a beneficial role of aneuploidy in stress response (Pavelka et al., 2010; Sheltzer et al., 2012; Siegel et al., 2012; Tang and Amon, 2013; Torres et al., 2007). By taking advantage of alternative transposition of Ac/Ds elements, we generated a 14.6 Mb segmental duplication in maize chromosome 1 with pronounced phenotypic and transcriptional effects. We recently showed that large tandem direct duplications can be generated at high frequency by another type of alternative transposition reaction termed reversed-ends transposition (Zhang et al., 2013). Thus, alternative transposition may be a productive mechanism to induce recent segmental duplications for research purposes, as well as a new approach to generate agronomically-beneficial duplications for maize breeding programs.
Methods

Phenotypic characterization of CNV plants

Copy-number variation plants (p1-ww714/p1-ww714, p1-ww714/B73 and B73/B73) were grown in the field season of summer 2012 in Iowa State and in the winter season of 2012 at Chile with three and two replications, respectively. Each genotype was randomly assigned to a 4-row plot in every replication. Individual plant heights were measured after flowering as the distance from the surface of soil to the top of tassel. Flowering date was recorded for each plot when more than half of its plants were shedding pollen. Ear length was measured as the distance from the bottom to top of the mature cob. A linear mixed-effects model that included fixed effects for genotypes and random effects for replications was used to test for phenotypic differences among genotypes.

Plant growth, tissue collection and RNA extraction

For mRNA sequencing, sibling seeds of three different genotypes (p1-ww714/p1-ww714, p1-ww714/B73 and B73/B73) were sown in SB 300 Universal Soil and grown in a PGW-40 growth chamber as described by Swanson-Wagner and colleagues (Swanson-Wagner et al., 2006); 15 h of light and 9 h of dark; 25°C at day time and 20°C at night time. Ten days after sowing, individual plants were genotyped by genomic PCR. At 14 days, above-ground tissues of six random plants per genotype per replication were harvested and pooled for preparation of total RNA. Nine pooled maize tissues (three genotypes x three replications per genotype) were ground in liquid nitrogen and total RNA was extracted by using RNeasy Plant Mini Kit (Qiagen) as per manufacturer’s instructions. Total RNA was further purified by performing On-column DNase digestion (Qiagen), and RNA quality and intensity was evaluated by a NanoDrop 1000
spectrophotometer and Agilent 2100 Bioanalyzer.

For small RNA sequencing, 6 plants from each of the four tested genotypes (p1-vv9D9A/p1-vv9D9A, id1/id1, id4/id4 and B73/B73) were grown in the growth chamber in the same conditions described above. Roots and seedlings were harvested at 14 days after sowing. Tissues from the same genotype were pooled together. For the p1-ww714/p1-ww714 and p1-ww714/B73 genotypes, tissue samples were taken from one of the replications used for mRNA sequencing. Total RNA was made from each pooled root and seedling tissue by the PureLink Plant RNA Reagent, followed by DNase treatment (DNase I, NEB).

qRT-PCR

Plants were grown in the same condition as previously did, with two replications, and were harvested at 14 days after sowing. Each replication included six plants that were pooled later to form a single sample. The methods and reagents used to prepare purified total RNA were the same as described for mRNA sequencing. Total RNA was reverse transcribed to cDNA using Omniscript RT kit (Qiagen). PCR was performed using Ssofast Evagreen Supermix (Bio-Rad), and analyzed on a Stratagene Mx4000 multiplex quantitative PCR system with three technical replicates.

GeneChip hybridization and statistical analysis of GeneChip data

Purified total RNA for three biological replicates of each genotype were sent to Iowa State University GeneChip Facility for labeling and hybridization to new maize 100 format whole transcriptome arrays according to the instructions of GeneChip WT terminal labeling and hybridization user manual (http://www.plexdb.org/modules/PD_general/Maize100WT_description.php). Raw GeneChip
data were normalized using the Robust Multi-array Average (RMA) approach which is implemented in the PLEXdb Pipeline (Dash et al., 2012). Differentially expressed genes were identified by linear model analyses carried out with the R package limma (Smyth, 2005). Each gene-specific linear model included effects for genotypes and replications. P-values from the genotype tests were converted to q-values (Nettleton, 2006). Genes with FDR less than 0.05 were identified as differentially expressed.

**mRNA and small RNA sequencing and data processing of Illumina reads**

cDNA sequencing libraries were constructed by the Iowa State University DNA Facility, and then sequenced on an Illumina HiSeq 2000 to generate 100 bp single-end reads, also at the Iowa State University DNA Facility. Random barcodes were used for each sample and three biological replicates of each genotype were performed and run on three independent lanes. Sequencing reads were trimmed by FASTX toolkit (http://hannonlab.cshl.edu/fastx_toolkit/) and then aligned to the maize B73 reference genome (RefGen version 2; www.maizesequence.org) by Tophat (Trapnell et al., 2009) using default parameters except a maximum intron size of 50,000 bp. Raw read count of each gene model (ZmB73_5b_FGS) (http://ftp.maizesequence.org/current/filtered-set/) was computed by HTSeq package (Anders et al., 2015) for uniquely mapped reads. Only genes having at least 45 mapped reads were considered to be expressed in the sampled tissue. Differentially-expressed genes with FDR less than 0.05 were identified based on the raw read counts by using QuasiSeq package (Lund et al., 2012).

For small RNA-Seq, libraries were constructed and sequenced at Beijing Genome Institute (BGI). Small RNA reads were trimmed by FASTX toolkit and then mapped to maize B73 reference genome (RefGen_V2) by Bowtie (Langmead et al., 2009). The genome was
subdivided into 100-kb windows across the 10 maize chromosomes, and the expression of small RNAs was represented by the number of mapped reads per 100-kb window.

**Statistical tests of transcript regulation models**

The experiment included three replications. Within each replication, one RNA sample was sequenced from each of the three genotypes. For each gene, we compared three generalized linear models, henceforth referred to as M0, M1, and M2 in increasing order of complexity. In each model, the log of the mean read count was assumed to be a linear function of gene-specific parameters plus a sample-specific normalization factor, determined for each sample by the log of the 0.75 quantile of sample-specific read counts (Bullard et al., 2010). The linear function for model M0 included an intercept and replication effects but no genotype effects. The linear function for model M1 included an intercept, replication effects, and a slope coefficient on the log of the number of copies of the duplicated region (2, 3, or 4 depending on the genotype). The linear function for the most general model (M2) included an intercept as well as replication and unrestricted genotype effects.

Genes for which model M0 is adequate are not differentially expressed across genotypes. Genes that require a more complicated model (either M1 or M2) are differentially expressed across genotypes. When model M1 is adequate relative to model M2, and when the slope coefficient in model M1 is not significantly different from 1, a gene's expression pattern is consistent with a dosage effect that implies mean expression level is proportional to the number of copies of the duplicated region. To determine which scenario holds for each gene, we performed the following analyses.

Genes with an average of at least one uniquely mapped read across samples were analyzed using the R package QuasiSeq (http://cran.r-project.org/web/packages/QuasiSeq). The
negative binomial QLShrink method implemented in the QuasiSeq package and described by Lund et al. (2012) (Lund et al., 2012) was used to compute a p-value for each gene and each model comparison (M0 vs. M1, M0 vs. M2, and M1 vs. M2). Using the p-values for each comparison, the approach of Nettleton and associates (2006) (Nettleton et al., 2006) was used to estimate the number of genes with true null hypotheses among all genes tested. Using this estimate, q-values were computed from p-values according to the method of Storey (2002) (Storey, 2002). To obtain approximate control of the false discovery rate at 5%, the null hypothesis was rejected for all tests with q-values no larger than 0.05. To assess the plausibility of effects proportional to dosage of the duplicated region, approximate 95% confidence intervals were constructed for the slope coefficient in model M1. Genes with intervals containing 1 are consistent with effects proportional to dosage.

Data access

All MIAME-compliant GeneChip profiling data are available as accession ZM50 at the PLEXdb expression resource for plants and plant pathogens (www.plexdb.org). RNA-Seq data is deposited as accession GSE71448 at NCBI-GEO and accession SRP061705 in the NCBI’s Sequence Read Archive (SRA); Small RNA-Seq data is deposited as accession GSE71959 at NCBI-GEO and accession SRP062285 in the SRA.

Acknowledgements

We thank Lisa Coffey and Patrick Schnable for assistance in growing plants in their growth chambers, Dafang Wang for preparing RNA samples for small RNA sequencing, and James Birchler for helpful comments on the manuscript. We thank the ISU Genome
Technologies Facility for providing the qRT-PCR service. This research was funded by NSF award 0923826 to TP, DN, DFW and JZ.

**Author Contributions**

T.Z., J.Z., and T.P. conceived and designed the research; T.Z., J.Z., and D.F.W. performed the experiments; T.P., J.Z., D.N. R.W., S.D. contributed new reagents/materials/analysis tools; T.Z., J.Z., A.L., S.D., R.W., D.N., and T.P. analyzed the data; and T.Z., A.L., D.N. and T.P. wrote the paper.

**References**


Proceedings of the National Academy of Sciences of the United States of America 110, 5241-5246.


FIGURES and LEGENDS

Figure 1. Effects of copy-number alterations on plant stature and ear size

In three BC$_5$F3 families, homozygous $p1$-ww714 plants have shorter stature, are delayed in development (A and B), and have shorter ears (C and D), than the standard maize inbred B73. Plants heterozygous for the duplication allele ($p1$-ww714/B73) are intermediate in ear length (C and D).
Figure 2. Differentially expressed genes (DEGs) and dosage-dependent expression pattern revealed by RNA-Seq.

(A) The distribution of differentially expressed genes (DEGs) between homozygous $p1$-ww714 and B73 plants identified by RNA-Seq. Bars indicate the numbers of over-expressed (red) and under-expressed (blue) DEGs identified in the 14.6 Mb duplicated region (first bar), the remainder of chromosome 1 (second bar), and other chromosomes.

(B) Ratios of chromosome 1 gene expression among three genotypes as determined by RNA-Seq. Log2-fold changes in transcript levels were plotted for pairwise comparisons among $p1$-ww714/$p1$-ww714, B73/B73 and $p1$-ww714/B73 sibling plants. Expression ratios are shown for all genes on chromosome 1. The gene copy-number ratios (log2) of $p1$-ww714/$p1$-ww714 vs B73/B73, $p1$-ww714/B73 vs B73/B73, and $p1$-ww714/$p1$-ww714 vs $p1$-ww714/B73 are 1, 0.585 and 0.415, respectively; these values are indicated as the dotted red lines in panel (B). The X-axis indicates the position of each gene in Mb on maize chromosome 1; the Y-axis indicates the transcript level ratios (log2) among the three tested genotypes. The segment duplicated in $p1$-ww714 (48.1 Mb to 62.7 Mb) is indicated by the two green vertical lines; expression ratios of genes within this segment are indicated in blue.

(C) Trans effects on expression of genes located outside the duplicated region. Heat map of transcript ratios (log2) of 128 differentially expressed genes located outside the duplicated region: 42 genes over-expressed (red) and 86 genes under-expressed (green) in $p1$-ww714/$p1$-ww714 vs. B73/B73. The transcript ratios were compared among B73/B73, $p1$-ww714/B73, and $p1$-ww714/$p1$-ww714 sibling plants as determined from mRNA-Seq data.
Genes in the duplicated region

Genes not in the duplicated region

95% CI for genes in the duplicated region

95% CI for significant genes not in the duplicated region
Figure 3. Statistical test of cis and trans-acting dosage effects.

A and B. Distributions of estimated values of $\beta_1$ for levels of gene transcripts in $p1$-ww714/$p1$-ww714 vs. B73/B73.

(A) A log-linear model was applied to test the hypothesis that gene transcript level is directly proportional to gene dosage. The histograms show the numbers of genes (y-axis) with the indicated estimates of $\beta_1$ (x-axis). The dosage effect hypothesis predicts that $\beta_1 = 1$ for genes within the duplicated segment (A), and $\beta_1 = 0$ for genes not in the duplicated region (B). The predicted values of $\beta_1$ are indicated by the red lines.

C and D. 95% Confidence Interval (CI) for each estimated $\beta_1$ value.

(C) 95% CI of $\beta_1$ value for each gene within the duplicated region. Red represents significantly Differentially-Expressed (DE) genes, and black indicates non-DE genes.

(D) 95% CI of $\beta_1$ value for each significantly DE gene located outside the duplicated region.
Figure 4. Dosage dependent expression of small RNAs. Log2-fold changes in small RNA levels per 100-kb window were plotted for comparisons between p1-ww714/p1-ww714 and B73/B73 (A), p1-ww714/p1-ww714 and p1-wwid1/p1-wwid1 (B), p1-ww714/p1-ww714 and p1-wwid1/p1-wwid1 (C) and p1-wwid1/p1-wwid1 and p1-wwid4/p1-wwid4 (D). The X-axis indicates the position of each 100-kb window in Mb on maize chromosome 1; the Y-axis indicates the ratios (log2) of small RNA level between the tested genotypes. Vertical lines (green) indicate the duplication borders. The numbers within two vertical lines represent the size of duplications. The thicker red lines indicate the 2 fold copy-number ratio between genotypes.
(A) Expression difference between \( p_{1-ww714}/p_{1-ww714} \) and \( B73/B73 \). \( p_{1-ww714} \) carries a 14.6 Mb segmental duplication (48.1 Mb to 62.7 Mb), indicated by the two green vertical lines; expression ratios of genes within this segment are indicated in blue.

(B) Expression difference between \( p_{1-ww714}/p_{1-ww714} \) and \( p_{1-wwid1}/p_{1-wwid1} \). \( p_{1-wwid1} \) has a 3.3 Mb segmental duplication which is overlapped by the 14.6 Mb duplication of \( p_{1-ww714} \).

(C) Expression difference between \( p_{1-ww714}/p_{1-ww714} \) and \( p_{1-wwid4}/p_{1-wwid4} \). \( p_{1-wwid4} \) has a 400 kb segmental duplication which is overlapped by the duplication in \( p_{1-ww714} \). The 400 kb overlapped duplication is represented by the first two green lines.

(D) Expression difference between \( p_{1-wwid1}/p_{1-wwid1} \) and \( p_{1-wwid4}/p_{1-wwid4} \). \( p_{1-wwid4} \) has a 400 kb segmental duplication which is overlapped by the 3.3 Mb duplication of \( p_{1-wwid1} \). The 400 kb overlapped duplication is represented by the first two green lines.
Table S1. Phenotypic differences observed between copy-number variant genotypes.

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Ear length - Summer</th>
<th>Ear length - Winter</th>
<th>Plant height - Summer</th>
<th>Plant height -- Winter</th>
<th>Flowering time (days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>B73/B73</td>
<td>16.0</td>
<td>14.7</td>
<td>210.8</td>
<td>215.2</td>
<td>66</td>
</tr>
<tr>
<td>714/B73</td>
<td>13.6**</td>
<td>13.4*</td>
<td>207.3</td>
<td>229.5</td>
<td>73**</td>
</tr>
<tr>
<td>714/714</td>
<td>11.9**</td>
<td>11.1**</td>
<td>164.2**</td>
<td>184.2*</td>
<td>77**</td>
</tr>
</tbody>
</table>

*: P <0.05 when compared with B73/B73.

**: P <0.01 when compared with B73/B73.

Table S2. Number of raw reads and uniquely mapped reads.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Raw Reads</th>
<th>Uniquely mapped reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>714/714-1</td>
<td>18,982,686</td>
<td>15,248,095</td>
</tr>
<tr>
<td>714/714-2</td>
<td>14,497,045</td>
<td>11,705,340</td>
</tr>
<tr>
<td>714/714-3</td>
<td>20,309,215</td>
<td>15,963,344</td>
</tr>
<tr>
<td>714/B73-1</td>
<td>14,611,587</td>
<td>11,504,948</td>
</tr>
<tr>
<td>714/B73-2</td>
<td>20,266,165</td>
<td>16,189,684</td>
</tr>
<tr>
<td>714/B73-3</td>
<td>11,874,525</td>
<td>9,486,044</td>
</tr>
<tr>
<td>B73/B73-1</td>
<td>15,458,862</td>
<td>12,201,558</td>
</tr>
<tr>
<td>B73/B73-2</td>
<td>15,571,597</td>
<td>12,114,433</td>
</tr>
<tr>
<td>B73/B73-3</td>
<td>18,142,503</td>
<td>9,668,982</td>
</tr>
</tbody>
</table>
Figure S1. Schematic structure and PCR analysis of inverted duplication allele *p1-ww714*.

(A) Lines indicate segments of maize chromosome 1S in progenitor allele *p1-vv-9D9A* (upper), and duplication allele *p1-ww-714* (lower). The solid boxes are exons 1, 2, and 3 (right to left) of the *p1* gene. Red lines with arrowheads indicate Ac and fractured Ac elements. Blue regions represent the 14.6 Mb duplicated segment. Short horizontal numbered arrows represent PCR primers. The breakpoint sequence is located between a and b.

(B) PCR was performed to analyze the structure of *p1-vv9D9A* (Lane 2, 4, 6 and 8) and *p1-ww714* (Lane 3, 5, 7 and 9) by using primers 1+2 (lane 2 and 3), primers 2+3 (lane 4 and 5), primers 1+4 (lane 6 and 7) and primers 4+5 (lane 8 and 9). Lane 1: DNA ladder.
Figure S2. Ratios distribution of all genes between p1-ww714/p1-ww714 and B73/B73 across all ten chromosomes. The numbers in x-axis indicate the genomic positions of maize chromosome. Red line represents 2-fold differentiation. Blue spots reflect the expressional difference of genes located in 14.6 Mb duplicated region.
Figure S3. Ratios (log2) distribution of the genes located in chromosome 1 from GeneChip.

The same RNA samples used for RNA-Seq were hybridized with GeneChip microarray.

Transcriptional changes among three genotypes were compared. The numbers in x-axis indicate the genomic positions of maize chromosome 1. The altered chromosome region (48.1 Mb to 62.7 Mb) locates within two green vertical lines. The red horizontal lines indicate the expected ratio changes for genes in altered segment.
Figure S4. qRT-PCR validation of DEGs via a 3.3 Mbp inverted duplication allele (*p1-wwid1*). 5 DEGs and 1 non-DEG previously identified between B73/B73 and *p1-ww714/p1-ww714* were analyzed by qRT-PCR as described in text. Gene1: GRMZM2G124532. Gene2: GRMZM2G394212. Gene3: GRMZM2G013639. Gene4: GRMZM2G033491. Gene5: GRMZM2G01823. Gene6: GRMZM2G070047. Genes 1 to 5 are DEGs. Gene 6 is non-DEG.
Figure S5. Ratios distribution of small RNA levels between \( p1-ww714/p1-ww714 \) and \( B73/B73 \) across all ten chromosomes. The numbers in x-axis indicate the genomic position of maize chromosome. Red line represents 2-fold differentiation. Two green vertical lines highlight the 14.6 Mb duplicated region in chr1. Each blue spot reflects the 100-kb window within 14.6 Mb duplicated region.
Figure S6. Ratios distribution of small RNA levels between *p1-ww714/p1-ww714* and *p1-ww714/B73* (upper), and *p1-ww714/B73* and *B73/B73* (lower) in maize chromosome 1 from 14-day old seedling tissue. The 14.6 Mb region is located between two green vertical lines. The expressional difference between two genotypes in a 100-kb window is represented by a spot; blue spots are within the 14.6 Mb region and black means outside the region. Red line represents the copy-number ratios (log2) of the 14.6 Mb segment; 0.415 for *p1-ww714/p1-ww714* vs. *P1-ww714/B73*, and 0.585 for *p1-ww714/B73* vs. *B73/B73*. 
Figure S7. Ratios distribution of small RNA levels between p1-vv9D9A/p1-vv9D9A and B73/B73 (upper), and p1-wwid1/p1-wwid1 and B73/B73 (lower) in maize chromosome 1 from 14-day old seedling tissue. The numbers in x-axis indicate the genomic position of maize chromosome. Red line represents 2-fold differentiation. Two green vertical lines highlight the 3.3 Mb duplicated region in chr1. Each blue spot reflects the 100-kb window within 3.3 Mb duplicated region.
Figure S8. Ratios distribution of small RNA levels in maize chromosome 1 from 14-day old root tissue. The comparisons were made between p1-wwD9A/p1-wwD9A and B73/B73 (upper), p1-wwid1/p1-wwid1 and B73/B73 (middle), and p1-wwid1/p1-wwid1 and p1-wwid4/p1-wwid4 (lower). Red line represents 2-fold differentiation. Two green vertical lines highlight the 3.3 Mb duplicated region in Chr1. The blue spot in the middle comparison reflects the 100-kb window within 3.3 Mb duplicated region.
CHAPTER 6. GENERAL CONCLUSIONS

The application of modern comparative genomic hybridization and sequencing techniques into biological studies has revealed a great many common features shared between eukaryotic genomes. Two of them are especially important and fundamental. Firstly, a significant proportion of eukaryotic genomes are derived from various families of transposable elements (TE); e.g. about 45% of our human genome is made of TE (Lander et al. 2001), and TE accounts for 85% of the maize genome (Schnable et al. 2009). Secondly, structural variations, especially copy number variations (CNVs), are commonly present in the genomes of two individuals within the same species. For instance, approximately 10% of the human genome is composed of CNVs (Iafrate et al. 2004; Sebat et al. 2004; Redon et al. 2006; Stankiewicz and Lupski 2010). It is conceivable that the enrichment of transposable elements and the widespread occurrence of structural variations might be causally related.

Barbara McClintock first discovered the Dissociation (Ds) element in the 1940’s. She found that the Ds locus is highly unstable; it could frequently induce chromosome breakage, and also lead to a variety of chromosome rearrangements such as deletion, duplications, inversions and translocations (McClintock 1948; McClintock 1950; McClintock 1956). Subsequent studies from various research groups revealed the underlying mechanisms that generate chromosome breakage and rearrangements: pairs of Ac/Ds elements in opposite orientation could induce chromosome breakage and certain types of sequence rearrangements including deletions and inverted duplications (Ralston et al. 1989; Dooner and Belachew 1991; MartinezFerez and Dooner 1997) via an alternative transposition mechanism termed sister chromatid transposition (SCT) involving sister chromatids (Weil and Wessler 1993; English et al. 1995). Additionally, a pair of Ac/Ds elements in reversed ends configuration could also induce chromosome breakage,
as well deletions, inversions and translocations via another alternative transposition mechanism named reversed-ends transposition (RET) (Zhang and Peterson 2004; Zhang et al. 2009b; Yu et al. 2010; Yu et al. 2011). These results show that Ac/Ds elements have the capability to produce structural variations.

To generate structural variations, SCT involves two sister chromatids, whereas the previously-characterized RET events act on only one sister chromatid. What if RET involves two sister chromatids? In CHAPTER 2, we extensively tested the idea of RET occurring on different chromatids. The allele we used for this study is p1-ovov454, which carries Ac and fAc insertions in intron 2 of the p1 gene. The ends of Ac and fAc in p1-ovov454 are reversely oriented, and the distance between them is about 800 bp. We characterized two reciprocal duplication/deletion alleles at the sequence level, and concluded that RET can generate reciprocal deletions and duplications when the paired termini from one chromatid insert into its sister chromatid. The identified 8 bp target site duplications, one at the breakpoint of the deletion allele and the other at the breakpoint of the duplication allele, supports the hypothesis that the reciprocal deletion and duplication were derived from a single RET event. In total, we isolated 9 duplications from approximately 2000 ears screened, suggesting a relatively high frequency of generating tandem direct duplications when paired Ac/Ds elements are less than 1 kb apart.

Previous research showed that Ac/Ds transposition usually occurs during S phase of the cell cycle (Chen et al. 1992). Accordingly, it is reasonable to suggest that paired TE ends could excise from a replicated DNA sequence, and then reinsert into an un-replicated target. We proposed a model of RET that involves an un-replicated target, and tested the model in CHAPTER 3. We found that Ac/Ds-mediated RET could produce reciprocal deletions and duplications, but also activate the re-replication of Ac/Ds and their adjacent sequences when they
insert into an unreplicated target site. The two re-replicated segments (one from Ac side and the other from fAc side) will become double strand breaks after the abortion of re-replication, and the two resulting double strand breaks will be joined by DNA repair machinery to form a composite insertion. Using a powerful visual genetic screen, we isolated three reciprocal deletion and composite insertion alleles. The expected 8 bp target site duplications were identified in the breakpoint junctions of reciprocal duplication and deletion alleles, and flanking the composite insertions, confirming that these 3 twin alleles studied originated by RET. We further characterized the internal structures of the composite insertion alleles; 3 are from twin alleles and the other 3 are whole ear event. The two double strand breaks in 3 out of 6 composite insertions were repaired by Homologous Recombination, whereas in the other three alleles the two re-replicated segments were joined by Non-Homologous End Joining. The composite insertions range in size from 4.5 kb to about 23 kb, and their distances to the initial Ac/fAc position varies from 13 kb to 1.7 Mb.

In CHAPTER 4, we investigated the evolutionary impact of RET mechanism by studying the prevalence of tandem direct duplications produced by RET. To identify tandem direct duplications in a target genome, we developed a new algorithm and bioinformatics pipeline (STRAND: Search for Transposon-Induced Tandem Direct Duplications) that can quickly and efficiently locate transposon-associated tandem direct duplications. The STRAND pipeline was applied to search for transposon-associated duplications in 22 plant genomes for which both genome sequences and DNA transposon (MITE) databases are available (Chen et al. 2014). We identified in total 62 transposon-associated two-copy duplications from 11 plant genomes. Because Non-Allelic Homologous Recombination (NAHR) can produce similar duplication structures (Hastings et al. 2009; Zhang et al. 2009a), we manually checked each duplication for
the presence of TSDs and the position and orientation of TE sequences in order to determine the mechanistic origin of the duplications. In fact, about half of the 62 duplications (29 cases) were derived from RET, and only 1/5 of them (11 cases) were produced by NAHR, indicating a higher frequency of RET in inducing transposon-associated tandem direct duplications. Previously research showed that other types of DNA transposon could undergo alternative transposon to generate chromosome rearrangements (Gray et al. 1996; Goryshin et al. 2003; Parks et al. 2004). Our results further strengthen this idea as we found RET-mediated tandem direct duplications associated with TEs from the *Tc1/Mariner, Mutator* and *PIF/Harbinger* families.

The last question addressed in this thesis concerns the immediate impact of recent generated duplications. Previous research has shown that duplications could cause dramatic phenotypic differences in plants and animals, but most of the studied duplications originated many years ago. The immediate gene regulatory response to duplication is not clear. In CHAPTER 5, we performed phenotypic and transcriptional analysis on a recent 14.6 Mb duplication which was generated by *Ac/Ds*-mediated alternative transposition. The major findings are: 1) Immediate phenotype changes were observed from plants carrying additional copy of the 14.6 Mb segment. Interestingly, the 14.6 Mb segment exerts a dosage-dependent effect on some tested traits such as ear length and flowering time; 2) Expression studies show that genes within the duplicated region are overrepresented among all detected differentially express genes; 3) Rigorous statistical tests show that genes within the duplicated region tend to exhibit dosage-dependent expression pattern; 4) The expression of genes outside the duplicated region could be affected by a trans-acting dosage effect; 5) Besides genes, small RNA transcripts also exhibit dosage-dependent expression. Our results extend our understanding of the immediate
influence of segmental duplications in causing phenotypic changes, and advance our knowledge of the immediate response of genes and small RNA transcripts to dosage changes.

Transposable elements are prevalent in eukaryotic genomes, and play a critical and unique role in creating mutations, influencing flanking gene expressions, and reshaping chromosome structures. My dissertation is focused on understanding the role of transposons in generating genome rearrangements. Our study extends the knowledge of the mechanism and impact of transposon-mediated duplications in the following aspects: 1) RET could induce reciprocal tandem direct duplication and deletion when the paired ends excise from one chromatid and then reinsert into the other sister chromatid; 2) RET could mediate a second round of replication of the transposon and their flanking sequences to form a composite insertion. This happens during DNA replication when the target site undergoes replication; 3) RET has played an important role in genome evolution, and the frequency of RET-induced tandem direct duplication is greater than that of NAHR; 4) Dosage-dependent expression likely plays a significant role in segmental duplications to regulate genes and small RNA transcripts thereby resulting in phenotypic differences.

References


MartinezFerez I, and Dooner H. 1997. Sesqui-Ds, the chromosome-breaking insertion at bz-m1, links double Ds to the original Ds element. *Molecular General Genetics* 255: 580-586.


ACKNOWLEDGEMENTS

First of all, I would like to thank Dr. Thomas Peterson for giving me the opportunity to study in his group. I am very grateful for his exceptional guidance, constantly encouragement, and great patience throughout my entire graduate study. What I learned from him is not just about being a good scientific researcher, but also the attitude of being a better person. I appreciate Dr. Dan Nettleton’s consistent support and guidance in the statistics and computational side. I would like to thank all my POS committee members Dr. Erik Vollbrecht, Dr. Roger Wise, and Dr. Nick Lauter for their support, guidance, and helpful suggestions.

Also, I would like to take this opportunity to thank the former and current members from Dr. Peterson’s group, including Dr. Jianbo Zhang, Dr. Chuanhe Yu, Dr. Dafang Wang, Dr. Jian Li, Weijia Su, Terry Olson, and Jeremy Schuster for their helpful discussion, suggestions, and comments. Many thanks are given to my good friends at ISU, especially Hsien-chao Chou, Josh Strable, Kokulapalan Wimalanathan, and Erica Unger-Wallace for their comments and suggestions. Finally, I would like to express my gratitude to my family members for their love, understanding, selfless support, and persistent encouragement.