The role of transposable elements and DNA damage repair mechanisms in gene amplification and protein domain shuffling in plant genomes

Ksenia V Krasileva*

1 Department of Plant and Microbial Biology, University of California, Berkeley CA 94720

* correspondence: kseniak@berkeley.edu

Short title: plant gene duplications and gene fusions

Highlights

- Plant genomes are dynamic, and gene-size insertions are prominent source of structural variation evident even within species.
- Gene insertions, tandem duplications and gene fusions can be created by different genomic processes, including transposons, as well as recombination and DNA repair machinery.
- Tandemly duplicated genes themselves can serve as template for new structural variation.
- Extrachromosomal circular DNA is untapped source of gene amplification and variation that can propagate outside of main genome and have adaptive potential.

Abstract

Plant genomes are shaped by structural variation. Gene-size insertions and among most prominent events and can have significant effects on amplification of gene families as well as facilitate new gene fusions. Transposable elements as well as plant DNA repair machinery have overlapping contributions to these events, and often work in synergy. Activity of transposable elements is often lineage specific and can preferentially affect specific gene families, such as disease resistance genes. Once duplicated, genes themselves can serve templates for additional variation that can arise from non-allelic homologous recombination. Non-homologous DNA repair mechanisms contribute to additional variation and diversify the mechanisms of gene movement, such as through ligation of extra-chromosomal DNA fragments. Genomic processes that generate structural variation can be induced by stress and therefore can provide adaptive potential. This review describes mechanisms that contribute to gene-size structural variation in plants, specifically gene duplication and generation of new plant genes through gene fusion.

Introduction

Plant genomes evolve through mutation, recombination, as well as structural variation (SV) that ranges from large scale chromosome re-assortment to gene-size insertions. The global gene order in plant genomes is conserved in large syntenic blocks and large-scale re-arrangements are relatively rare. In contrast, gene-size insertions are more common. Genes or gene fragments can duplicate, move around the genome, and form gene fusions [1–3]. Genome plasticity allows plants to adapt to new environments and played a large role in crop domestication [4].
Contiguous assemblies of complex plant genomes and pan-genome projects are now revealing the extent of SV in plants [5–8].

The SV in plant genomes has been decoded using two main approaches. First, a comparison of syntenic blocks or orthologous genes can identify genes and regions that disrupt synteny. It is often used in whole genome projects and has been applied to understand variation in gene content within and across species. Second approach involves analysis of a particular gene family or class of transposable elements. There are several plant gene families that showcase SV. Tandem duplication played significant role in lineage specific expansions of genes that determine agronomic traits, such as lectins [9], papains [37] and gluten [11,12]. The plant immune receptors of Nucleotide Binding Leucine Rich repeat domain (NLRs) vary in gene number from 30 in cucumber to 2,500 in wheat [13]. NLRs are often found in gene clusters with both tandem and inverted duplications (Fig 1) and participate in gene conversions, as well as fusions with other plant genes that become attached to NLR as integrated domains (IDs) [14–16]. Although the exact mechanism of NLR-ID formation is unknown, it involves gene duplication and inter-chromosomal translocation of either NLRs or IDs [2,3]. Whether comparing whole genomes or particular gene families, studies often reveal that different mechanism have contributed to the genomic landscapes that we see today.

Transposable elements (TE) are major drivers of plant genome evolution. TEs have been shown to facilitate duplication or deletion of plant genes, affect their expression or function, and combine genes from different locations into new fusions [17]. TEs come in two major classes, retrotransposons that move through an RNA intermediate, and DNA transposons that move directly as DNA [18]. TEs can be autonomous, encoding all enzymes needed for transposition or non-autonomous encoding the recognition sequences but relying on enzymes from elsewhere in the genome. Amplification of active elements contributes to genome size. The content of TEs in plant genomes varies from 3% in the genome of aquatic plant U. gibba [19] to over 85% in complex polyploid plants, such as wheat [20]. Long Terminal Repeat (LTR) retrotransposons contribute most to the expansion of TE content [21], constituting over 65% of 17 Gb wheat genome [20]. Other TEs, such as PacMULEs and Helitrons contribute relatively small percentage to the whole genome content (from 0.1% in wheat [20] to 2.2% in maize[22]), but have a profound effect on gene composition due to their ability to trap genes and move them around the genome and facilitate novel domain combinations through gene fusions [23–25]. Noteworthy, active TEs are also often associated with loci controlling adaptive responses, such as abiotic and biotic stress, suggesting their beneficial role in genome evolution [26].

Gene-size insertions in plants can also be a result of DNA repair mechanisms [1,27,28]. Repetitive DNA as well as duplicated genes provide templates for non-allelic homologous recombination (NAHR) that occurs during the loss of recombination control [29]. Another DNA repair process, non-homologous end joining (NHEJ) can ligate double stranded breaks in genome without homologous template [30]. NHEJ is a predominant process that repairs double stranded DNA breaks in plants and can lead to either deletions during the recess of broken DNA or insertions if a broken strand is filled in [1]. Plant TEs and DNA repair machinery often work together and can produce similar SV outcomes. This is well exemplified in recent discovery of extra-chromosomal circular DNA molecules (eccDNA), that can arise from looping out of tandemly duplicated elements in the genome or direct ligation of retrotranscribed TEs by NHEJ [31].
This review summarizes some of the mechanisms that contribute to evolution of gene content in plant genomes, with a focus on gene-size SV, such as insertions, duplications, translocations and fusions of genes and gene fragments.

Main text of the review

Different scales of genome re-arrangements

The SV can be divided into two groups based on the length of affected genomic segments. Large-scale SV, such as chromosomal translocations are relatively rare in plants, however they are well documented in model species as well as crops. Large chromosomal translocations are often associated with initiation of DNA breaks that occurred during natural history of plant genomes [32] or induced by mutagenesis and T-DNA insertions [33]. Large scale translocations can affect gene expression, and genes directly at the breaks. Large-scale translocation events are often implied from disrupted synten of large genomic blocks.

SV on a smaller, kilobase scale (0.1 kb to 1Mb) influences tandem gene duplication, translocation and new gene fusions. Synteny analysis between Arabidopsis thaliana and A. lyrata estimated 420 individual genes to be affected by insertions/deletions [28]. Analysis of 453 rice accessions revealed a total of nearly 90,000 SV of 5-500 kb in size with several thousands of SV affecting genes [6]. In wheat as many as 27% of its nearly 100,000 genes are found as tandem duplicates [20].

Such extensive SV is exciting because it allows to evaluate the plasticity of different genes and their genomic context with potential of targeted manipulation. To present a more detailed example of gene-size SV, we examined a 35 kb locus in Brachypodium pan-genome [5] (chr4:9,395,000-9,430,000) that contains one of validated NLR-WRKY gene fusions [2,3] (Figure 1). Our analysis of this region showed that only six genes were maintained in synteny, while thirteen genes have been inserted from elsewhere in the genome. Analysis of this region in Brachypodium confirms rapid evolution of the same region in rice [2]. Such extensive SV likely due to combined activity of different genomic processes, that include active TEs, replication, recombination and DNA damage repair machinery.

Gene duplication and movement facilitated by retro-transposons

Autonomous and non-autonomous retrotransposons constitute major part of plant genomes. [18]. Among the most prominent and active TEs that have been shown to capture genes are LTR retrotransposons with characteristic long terminal repeats [34,35]. The gene capture by LTRs can occur during the reverse transcription when a template switches to a messenger RNA of another plant gene. Upon insertion back into the genome, the gene will be duplicated without its introns and will be now surrounded by LTR sequences (Figure 2A). Identification of retroposed genes affected by LTRs therefore is relatively straight forward. An example of gene family expansion through retro-transposition is sub-clade specific expansion of NLR immune receptors in peppers [34] (Table 1). Several next generation sequencing based methods have been developed to isolate and sequence active LTRs [31,36]. Global analyses of LTR activity show that they can be induced by stress [31,36], so the genes trapped by LTRs might have an adaptive advantage.
Tandem repeats of retrotransposons also facilitate tandem gene duplications that does not involve retroposition. In this case, terminal repeats can align to induce gene excision or gene duplication that is mediated by host NAHR. NAHR can occur due to the loss of replication control as have been well documented in other eukaryotes, including yeast [37,38] and animals [29]. Because NAHR breakpoints occur at homologous locations, they are often undetected and finding them requires specialized algorithms [29].

**Gene capture and novel gene fusions inside DNA transposons**

The two main DNA elements that can trap multiple gene fragments facilitating new domain architectures are Pac-MULEs and Helitrons. Pac-MULEs were first identified in rice and have been shown to contain gene fragments as well as chimeric gene fusions [24] (Table 1). Pac-MULEs preferentially acquire GC-rich fragments and insert into 5’ ends of genes, facilitating formation of new exons and gene fusions [25]. Pac-MULEs are Mutator family Class II DNA elements that leave terminal inverted repeats (TIRs) surrounding their insertions (Figure 2B). Identification of Pac-MULEs therefore is based on TIRs. Genome-wide identification of Pac-MULEs often relies on RepeatMasker [39] to identify Mutator families of TEs and subsequent custom analyses of captured genes and gene fragments.

Helitrons have been discovered relatively recently, and they do not have repeats on their ends [40] (Table 1). Instead, they begin with CTRR (most sequence start with CTAG) and end with a short palindromic repeat [40]. Notably, Helitrons can capture multiple genes and produce chimeric transcripts [41]. Upon alternative splicing, such transcripts result in novel chimeric protein fusions [42,43]. Several mechanisms of gene capture by Helitrons have been proposed [40,41,44]. One mode by which Helitrons can capture genes involves a readthrough during transposition where the first the terminal element is either lost or not-functional and the machinery continues to read through the neighboring gene until the next terminal element is encountered (Figure 2C). The second proposed mode of gene capture is double-stranded break repair with an exogenous gene providing repair template usually through a very short matching sequence (Figure 2C). The two modes can be clearly distinguished only if a gene capture occurred very recently (e.g. the original Helitron is present upstream of parental gene). Helitrons are challenging to identify because they do not have terminal inverted repeat and do not replicate host sequence upon insertion. Genome-wide identification tools of Helitrons rely on DNA structure prediction or include local combinational variable analyses [45].

DNA transposons can also participate in generation of tandemly duplicated genes through reserve end transposition as documented for maize Ac elements [46]. This occurs when TEs are positioned next to each other in correct orientation, which triggers an alternative form of transposition that deploying the terminal repeats of two different elements. A detailed overview of reverse end transposition is depicted in [46].

**Gene duplication and conversion through non-allelic homologous recombination**

Plant TEs often work in synergy with host recombination and DNA damage machinery. Several studies have concluded that most gene movements in plants cannot be explained by transposable element activity alone [1,3,28] (Table 1). A pair of directly repeated genes even without other TEs nearby can lead to gene excision or further duplications through NAHR. Early experiments with reporter genes in Arabidopsis demonstrated that even in the absence of transposable
elements, crossing over between repeated genic sequences can either increase or decrease copy number [47]. Recent studies showed that NAHR can affect more than one gene in a region [48]. Besides gene duplication, NAHR can lead to new chimeric combinations of repeated copies [49–52] (Table 1). In addition, NAHR can also induce large scale translocations and chromosome fusions when repeated sequences are located on different chromosomes. For schematic examples of outcomes of NAHR, see [29].

It is yet unclear if inter-chromosomal NAHR might be also guided by proximity of non-allelic homologs in 3D as could be in future determined by techniques such as Hi-C [53,54]. NAHR can often go undetected because it involves exchange of closely related sequences and requires probabilistic models [29]. Multiplex ligation-dependent probe amplification has been instrumental in identifying gene copy number in absence of contiguous genome assemblies [55].

**De novo repeat generation and non-homologous DNA repair**

NHEJ is one of the main DNA repair mechanisms that participates in many gene movement and gene fusions processes, either initiated by TEs, such as discussed above or during the DNA damage induced by cellular enzymes, UV and other mutagens. The outcomes of NHEJ can be large scale translocations when ends of different chromosomal segments are joined together or small SV when repair is local. NHEJ plays a critical role in gene translocations and chimeric gene fusions when the whole chromosome is undergoing catastrophe. Catastrophic reconstruction of chromosomes is critical during genome elimination and can also create new gene fusions [56]. Intriguingly, analyses of rice genomes implicated NHEJ in most cases of gene transpositions [27] (Table 1).

NHEJ is also central to generation of tandem duplications without activity of TEs. Two main mechanisms have previously been implicated in this process: replication slippage and long-patch base excision repair (BER) mechanism. Replication slippage depends on microhomology between the end of duplicated region and beginning of the donor site. Long-patch BER depends on double stranded DNA breaks with complementary overhangs that are filled in using complementary strand, and end products are ligated with NHEJ [57]. A study of natural gene insertions in rice showed that long-patch base excision repair mechanism is prevalent and accounts for most recent insertions [57] (Table 1).

**Gene movement and amplification on eccDNA**

Multiple roles of NHEJ are exemplified by generation of eccDNA. The eccDNA molecules are present in both plants and animals and vary in size from 2000 to over 20,000 base pairs [58]. The eccDNA can arise by NHEJ of reverse transcription products of retroelements (Figure 2D) [31] or directly from genome when during DNA repair, NHEJ can also create circular DNA products from any tandemly duplicated sequences, either TEs (Figure 2E) [46], other repeats or duplicated genes. The resulting eccDNA can replicate multiplying copies of trapped sequences and providing platform for mutation. Traditionally, targets on eccDNA can be isolated by rolling circle amplification. With advancements of sequencing technologies, the whole eccDNA repertoire can be sequenced. The circular eccDNA are protected from exonuclease and therefore can be separated from genomic DNA by digestion and amplified through rolling circle amplification using random primers [31].
While most of eccDNA contains non-coding repetitive elements, further research is needed to examine genic content of eccDNA on the genome scale. Interestingly, among the genes reported to be propagated on eccDNA is 5-ENOLPYRUVLISHIKIMATE-3-PHOSPHATE SYNTHASE (EPSPS), the molecular target of glyphosate and the gene responsible for transmission of herbicide resistance in crops [59]. The mutated EPSPS has been found in high copy number on autonomously replicating eccDNA, suggesting that eccDNA can also serve as an important reservoir of producing rapid somatic variation that can then be transmitted to germ line.

Comparative genomics of Arabidopsis thaliana and A. lyrata proposed that recombination across tandem gene copies can lead to formation of eccDNA containing one of the repeated elements as well as a deletion in the source (Figure 2E). The circular DNA can then move to another location and recombine with a similar repeat elsewhere in the genome, including other genes. This implies that in principle, genes from eccDNA could integrate back to genome. The result could be a gene insertion or generation of novel gene fusion [5], however this process has yet to be demonstrated. Advances of single molecule sequencing technologies should allow now to investigate generation and evolution of eccDNA, their content and ability to integrate back to genome.

Conclusions:

Gene-size insertions are one of the main drivers of small-scale SV and gene content variation in plant genomes. The same outcomes can be generated through a variety of mechanisms, involving TEs and DNA repair machinery. While we have a solid understanding of each individual process that can generate SV, a comparative analysis of their contributions across plant lineages is only becoming possible now due to large number of high-quality genomes and pan-genome projects. Tools that can process large scale dataset and integrate different types of SV signatures are becoming available [60]. Even within a lineage, specific TE elements can be dominant [61], have specificity towards a particular gene family or even sub-family [34] or show activity under specific stress conditions [36,59]. Understanding when and how each process is activated, and its specificity can inform not only natural history of plant genomes but also more efficient genome editing tools.

Glossary

Long terminal repeats (LTR) – identical sequence of DNA that can be repeated at the ends of retrotransposons.
Retroposition – DNA movement through RNA intermediate which is converted back to DNA prior to insertion.
Terminal inverted repeats (TIRs) – palindromic sequence often found at the ends of Class II transposable elements
PacMULEs – Mutator-like DNA transposons that often capture exogenous genes or gene fragments within their TIR.
Helitrons – rolling circle replicating DNA transposons without TIRs which also are capable of capturing genes or gene fragments and inserting them preferentially in GC-rich 5-ends of genes. Helitrons can also induce alternative splicing.
Extra-chromosomal circular DNA (eccDNA) – circular DNA from 200 bp to 20 kb and more that is derived from and accessory to main chromosomes.
Long patch base excision repair (BER)
**Chromosome catastrophe** – double-stranded break induced digestion of whole chromosome, often followed by re-arrangement or fusion of fragments.

**Non-homologous end joining (NHEJ)** – ligation of double stranded breaks in genome that does not require homologous template for repair.

**Non-allelic homologous recombination (NAHR)** – ligation of double stranded breaks in genome that required homologous template for repair from non-allelic location.

**Nucleotide Binding Leucine Rich Repeat proteins (NLR)** – a large family of plant immune receptors known for rapid evolution, gene movement, and forming new gene fusions.

**Integrated domains (IDs)** – genes found in fusions to NLRs.
Acknowledgements

I would like to thank Erin L Baggs and Michael Freeling for the discussions of mechanisms underlying plant structural variation. I am grateful to Daniil Prigozhin for additional comments and manuscript editing.

Disclosure

The author declares no conflict of interest

Literature of interest and special interest

1. **Wicker T, Buchmann JP, Keller B:** Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res* 2010, doi:10.1101/gr.107284.110. This study examines gene-size insertions in grasses, one of the few studies that systematically document different mechanisms of gene movement and their relative contributions. While some of the gene insertions are bordered by transposable elements, others are likely to arise from template slippage or unequal crossing over.


The paper shows that inraspecific erosion of gene colinearity in maize can be due to movement of 4 genes by Helitron. The ends of Helitron that are often elusive were reconstructed from the Helitron-minuses maize lines.


This is a first systematic analysis of Pac-MULEs based on rice genome that documents it ability to capture gene fragments from different chromosomes creating chimeric gene fusions inside the element.

This paper documents preference of Pac-MULEs to capture GC-rich genes and gene fragments and preferentially insert into 5'-ends of genes, with possibility of creating new gene fusions. The Pac-MULEs can be in part responsible for the negative GC gradient observed in grasses.


This study provides a genome wide scan for gene insertions that happened since the divergence of Arabidopsis thaliana and Arabidopsis lyrata, with papaya as an outgroup. The results indicate the most of gene movement occurs independent of transposons. Certain gene categories, including NLRs are more prone to move. The proposed model involves circular DNA intermediate that loops out from the donor site and inserts into the destination site through non-allelic homologous recombination.


This study sequences extra chromosomal circular DNA from Arabidopsis and rice to characterize the mobile elements. The methods of isolating eccDNA from plants are well described. Results suggest inducible activity of selected retrotransposable elements upon stress and destabilization of epigenetic regulation.


   http://www.repeatmasker.org 2017,


   A whole genome comparison across two maize lines identified that helitrons are responsible for eight out of nine observed gene sized insertions. Some of the transcripts contained fragments of different genes solidifying the role of helitrons in novel gene fusions.


   This is a well documented analysis of reverse-end transposition and the role of DNA transposons in generation of tandem repeats.


This paper reports a critical example of adaptive gene evolution on eccDNA that allowed wild grass species to acquire herbicide resistance.


This paper presents a detailed analysis of LTR family dynamics across Pennisetum, Saccharum, Sorghum and Zea lineages. Their analyses indicate that independent activation of distinct LTR elements occurred in different lineages and these elements can become responsible for genome amplification. The results also suggest random activation of LTRs, with some families being more prone to re-activation and amplification.
### Table 1: Known gene insertion mechanisms

<table>
<thead>
<tr>
<th>Name</th>
<th>Consequence</th>
<th>Footprint</th>
<th>References</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transposable elements</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Host machinery</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BER + NHEJ</td>
<td>Tandem duplication</td>
<td>Tandem repeats</td>
<td>Vaughn and Bennetzen, PNAS 2014</td>
</tr>
<tr>
<td>NHEJ</td>
<td>Large scale translocations, chromosome fusions</td>
<td>Nucleotide recession (loss), nucleotide addition (filler)</td>
<td>Ma et al PNAS 2004, Wicker et al Genome Res 2010</td>
</tr>
</tbody>
</table>
Fig 1: Gene duplications, insertions and deletions in a genomic locus across 14 Brachypodium accessions

The overall relatedness of 14 Brachypodium accessions based on whole genome analysis [5] is displayed on the right with accession names next to the branches. The analyzed 35 kb region surrounding NLR-WRKY fusion (Bradi4g09886) is located on chromosome 4 (chr4:9,395,000-9,430,000) in the reference Bd21 accession. Gene numbers correspond to orthologs/paralogs in following categories: 1-6 syntenic genes (purple, yellow, salmon-green displaying a fusion, blue and black), 7-11 exogenous NLR insertions (salmon), 12-18 exogenous non-NLR insertions (grey). Filled arrows represent annotated proteins, empty arrows are genes and gene fragments confirmed to be present in genomic assembly by BLAST, but absent in protein annotation.
Fig 2. Schematic diagram of the mechanisms that lead to gene duplications and gene fusions. A) Gene capture by LTR during the reverse transcription stage: both LTR and other plant genes are first transcribed into mRNA; during reverse transcription of LTR, there is a...
template switch and mRNA of another gene is used which creates DNA chimera. Re-integration back into the genome traps a gene or its part within an LTR. B) Sequential gene capture by Pac-MULE which creates new gene fusion: during DS repair, priming to another gene is used as template which leads to integration of a gene within MULE, if a process continues, another gene gets trapped, the element then can propagate new fusion. C) Gene capture by Helitron: capture can occur either by readthrough during element replication or during DS repair, the result can trap one or more genes inside the element and create new fusions. D) Generation of eccDNA from TEs using an example of a gene trapped within LTR from A): After transcription and reverse transcription of and LTR with a gene trapped inside, the end of the linear DNA can be ligated by NHEJ creating circular product. eccDNA can replicate amplifying TE and any trapped gene, providing platform for mutation, and subsequently re-integrate into the genome. E) Generation of eccDNA from tandemly duplicated gene: similar to any repeat, tandemly duplicated gene can loop out and through NAHR form a circular product leaving a deletion in the genome. The eccDNA can amplify and possibly re-integrate elsewhere in the genome, forming gene insertions and gene fusions.