CHROMOANAGENESIS IN PLANTS AND THE EFFECTS OF STRUCTURAL VARIATION IN POPLAR

By

WEIER GUO
DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

PLANT BIOLOGY

in the

OFFICE OF GRADUATE STUDIES

of the

UNIVERSITY OF CALIFORNIA

DAVIS

Approved:

Luca Comai, Chair

Julin Maloof

Andrew Groover

Committee in Charge

2023

ABSTRACT

Chromoanagenesis in Plants and The Effects of Structural Variation in Poplar

By

Weier Guo

Structural variation plays an important role in plant genome architecture and phenotypes. It is also suggested to be a new type of DNA marker for genomic selection in plant breeding. While many studies have characterized the properties and effects of structural variation on crops, its origin and the interplay with other genetic factors still remain unclear. This study investigated the origin of complex genomic structural variation, as well as the interplay between induced structural variation and natural nucleotide polymorphism on plant phenotypes. One novel type of structural variation characterized recently is chromoanagenesis. Chromoanagenesis is described as a catastrophic event resulting in chromosomal restructuring on a localized region, mostly involving one single chromosome. Although chromoanagenesis has been largely characterized in animal cells, its presence and origin in plants have not been determined. Sequencing of the genomes of a gamma irradiation-mediated *Populus* hybrid population detected 2 F1 lines carrying shattered chromosomes. One line exhibited shattered patterns on chromosome 1 and the other on chromosome 2. Novel DNA junctions were identified and validated in these 2 lines, and the results confirmed that the reorganized segments were consistent with what is expected as the product of chromoanagenesis. Genomic features enrichment analysis indicated that breakpoints were likely to occur in gene rich regions. Chromoanagenesis-like patterns were also observed in a hybrid *Arabidopsis thaliana* line carrying the *asy1* mutation. Short-read sequencing revealed that the shattered pattern was on chromosome 1. 249 novel DNA junctions were identified with both ends associated with shattered regions. As in the *Populus* case, breakpoints were significantly enriched in genic regions. SNP frequency analysis revealed that the restructured chromosome resulted from mis-segregation at Meiosis I in the maternal parent. These two cases suggested that chromoanagenesis can originate from mutagenesis in plants. To further characterize structural variation and document their effect on plant phenotypes, we decided to investigate the effects of structural variation in forest trees, whose structural variation studies are relatively new. Highly polymorphic forest trees are expected to carry high levels of allelic and dosage variation, and the interaction of these two types of variation and their combined effect on phenotype is unclear. In a *Populus* hybrid population, QTL analysis was performed to document the effect of two types of variation on traits - natural allelic variation and induced dosage variation. Results suggested that QTLs from allelic and dosage variation were independent. Integrating the QTLs from both allelic and dosage variation exhibited significant improvement on phenotypic variance explanation compared to only allelic or dosage QTLs. These findings provide a snapshot of the relationship between allelic and structural variation and their effects on plant phenotypes.

Dedicated to dad and mom.

ACKNOWLEDGEMENTS

I thank Dr. Luca Comai for his mentorship during my research.

I thank Dr. Isabelle Henry for her mentorship, guidance, insightful comments, and suggestions at every stage of the research work.

I thank my committee, Drs. Luca Comai, Julin Maloof, and Andrew Groover for their expertise and guidance.

I thank my collaborators, Dr. Héloïse Bastiaanse, Meric Lieberman, and Helen Tsai, for their contributions and technical support, and Dr. Kirk Amundson, Benny Ordoñez, and other members and alumni of the Comai laboratory for lending their expertise.

I thank Dr. Yuehui Chao, my undergraduate research mentor, for his guidance and support. His encouragement aroused my enthusiasm to pursue research.

I thank the Plant Biology Graduate Group members for their friendship.

I thank my dad Ping, who is always patient in listening to my ideas, complaints, successes, failures, happiness, and concerns and encouraging me with his life experiences and humor. I want to thank my mom Huiling, who greatly supports my education and well-being. I want to thank my grandma Xiue, who never stops reading and learning; her spirit encourages me a lot. I would also like to thank other family members for their unwavering support.

I thank my husband, Dr. Xian Xiao, for his patience, encouragement, and belief in me. His support helps me get through the difficult times.

I thank my cat Baibey for her companionship.

TABLE OF CONTENTS

List of Tables

List of Figures

Chapter 1

Introduction

Introduction

Structural variation in plants

Structural variation (SV) is defined as a DNA region that undergoes changes in sequence length, copy number, orientation and/or genomic location between individuals [1]. SV can cause gene loss, gene duplication and novel gene production, which in turn affect plant phenotypes. In general, SV is categorized into deletion, insertion (large-scale indels), copy number variations (CNVs), inversion and translocation. SV usually refers to the rearrangement of DNA sequences larger than 50 bp and are considered to have greater effects on gene expression compared to single nucleotide polymorphisms [2].

SV is prevalent in many plant species, and a large number of SV have been shown to affect protein-coding regions, especially for unbalanced SV (large-scale indels and CNVs) [3–14]. A significant fraction of genes within SV encode for hypothetical or unknown proteins [15,16], while those annotated genes in SV regions are enriched for large gene families, and are often associated with disease resistance and biotic stress response [15–21]. Although some SV have qualitative effects on traits [6,8,9,11,13,14,22,23], it is suggested that most SV tend to have quantitative effects, making them more challenging to detect [24]. As a result, the effects of SV on traits tend to be underestimated. This is especially true for complex traits such as growth and adaptation. Therefore, SV may represent a new type of DNA marker for plant phenotypic prediction and genomic selection [25–27].

A novel type of structural variation: Chromoanagenesis

Many novel types of genomic structural variation have been recently discovered thanks to vast improvements in sequencing technologies [28–30]. One of the novel genomic structural variations is chromoanagenesis, which is described as a catastrophic event that involves localized chromosomal restructuring [31]. Chromoanagenesis was firstly identified in human tumor cells and later was found prevalent among many cancer types [32–35]. There are several features that make chromoanagenesis unique from other types of SV: (a) Chromanagenesis occurs on one or a few (< 4) chromosomes, while the other SV types are usually randomly distributed along the genome. (b) Chromoanagenesis can result in tens to hundreds of double strand breaks on a localized region, which is much more intensive than other SV types. (c) The products of chromoanagenesis - extremely rearranged chromosomal segments - all occur simultaneously, or within a few cell division cycles. This indicates that chromoanagenesis is a catastrophic event rather than the accumulation of multiple modest structural variations.

Studies in cancer suggest that the origin of chromoanagenesis is always associated with DNA damage [31,36]. Plants may experience chromoanagenesis following naturally or artificially induced mutagenesis [30,37–39]. Genetic engineering is commonly used on plants, and many genetic engineering techniques are demonstrated to induce genome instability [40–43]. Recent studies have reported chromoanagenesis-like events occurring in genetic engineered plants [41,44]. Studying chromoanagenesis on plants can improve our understanding of the advantages and limitations of genetic engineering techniques on plant breeding.

Many chromoanagenesis-like events were identified in plant species recently (Details in Chapter 2). The current challenges to studying chromoanagenesis in plants includes (a) developing a standard method for predicting, identifying and validating chromoanagenesis in plants (b) understanding the origin and potential mechanism underlying this process (c) investigating the effects of chromoanagenesis on plant traits.

Challenges of *Populus* breeding

Populus (poplars, aspens and cottonwoods) is a model genus among forest trees [45]. It firstly became domesticated in Europe, and later spread and became cultivated in four continents [46–49]. *Populus* has a wide range of practical applications, including providing bioenergy feedstock, pulping and timber production [50]. However, traditional *Populus* breeding approaches are expensive and time consuming. Additionally, different traits (such as bioenergy feedstock and timber production) required very different *Populus* ideotypes. So it becomes very important to develop improved *Populus* cultivars with the increasing demands on forest tree products.

Developing better *Populus* cultivars starts with an improved understanding of many aspects of poplar physiology and genetics. Many *Populus* crucial traits such as biomass yield are under complex genetic regulatory mechanisms. Identifying genetic components regulating these mechanisms can help predict *Populus* phenotypes and select desirable cultivars. Since *Populus* is dioecious and its genome is highly polymorphic, it provides abundant DNA sequence polymorphisms. Early studies have mostly focused on natural allelic variation (SNPs) to identify genetic components for interest traits. However, many complex traits, including biomass yield, cannot be fully explained by these observed genetic components. It is possibly because: (a) Some of the genetic variation has not been taken into account; (b) Some genes have subtle effects on traits, which are non-detectable through genomic selection; (c) The recent whole genome duplication event occurred in *Populus* created redundancy in gene contents, which may help buffer the effects of variation. New genomics-based approaches need to be developed for addressing these problems.

Opportunities and limitations of structural variations in *Populus* breeding

SV can serve as good DNA markers for identifying genetic regulators of complex traits. Many studies on domesticated crops have shown the phenotypic contributions of SV [3,5,6,10,14,17,51,52]. The study of SV in forest trees is relatively new [53–55]. As a model plant of forest trees, *Populus* becomes an attractive system to study SV and their effects on phenotypes, especially because of its modest genome size and ease of vegetative propagation. Pan-genomic analysis of *Populus* observed naturally occurring SV were mostly located in intergenic regions, and genes affected by SV were kept at low expression levels [56]. Among different SV types, deletion tends to be the most prevalent type in gene coding sequences [57–59]. Adaptive traits, including phenology and physiology were reported to be associated with naturally occurring gene copy number changes [57]. A *Populus* hybrid population (~ 600 individuals) carrying artificially induced SV was established at UC Davis [60]. Large-scale deletions and insertions (indels) were characterized in this system, and these indels were used to investigate the gene dosage effects on quantitative traits in *Populus* [61–63]. Studies based on this *Populus* system found many novel genomic loci associated with various traits including phenology, biomass, vessel development and leaf morphology [61–63].

Although SV provides novel variation for genomic selection on many traits in tree breeding, they are also associated with some limitations. First, focusing on the effects of SV may overlook the influence of allelic variation. As a dioecious tree species, *Populus* has a genome that is highly heterozygous. Many studies on *Populus* have shown the contribution of allelic variation in traits including phenology [64,65], disease resistance [66], biomass [67] and leaf shape [68,69]. Current studies using SV as a marker to identify regulators beneficial for *Populus* traits mostly focus on the unbalanced SV such as CNVs, deletions and insertions [57,61–63]. Although these studies identified novel potential genomic regions for traits and increased explanation percentage on observed phenotypic variance, there was still missing heritability for quantitative traits. Second, SV contributes to the phenotypes with a larger magnitude compared to SNPs [55]. Therefore, many genes with subtle contributions on phenotypes may be masked by the large effects of SV. Integrating the different types of genomic variation may be a potential method to solve these two problems [70]. First, integrating the effects of allelic and structural variation may better predict observed phenotypic variation. Second, the interactions between allelic and structural variation can help interpret the mechanisms underlying complex traits, which leads to the observation of genes with subtle contributions on traits.

Problem definition

Structural variation (SV) is demonstrated to be widespread in plants and can affect plant phenotypes. However, methods and approaches for identifying SV and characterizing their effects on plants are not well defined [71]. It is crucial to build common criteria for SV investigation in plants. For example, an increased number of novel types of genome instability events were detected in plants, which do not belong to any traditional type of SV. Chromoanagenesis is one example. Chromoanagenesis was initially observed in human tumor cells [32]. The first chromoanagenesis event in plants was found in *Arabidopsis* associated with defective centromere [30]. It usually results in one extremely rearranged chromosome, with the rest of the genome kept intact. Despite one observation of chromoanagenesis in plants, other potential triggering processes and the occurring mechanisms of chromoanagenesis in plants remain unknown. On a related note, the effects of SV on plant phenotypes can be modified by many factors such as DNA sequence polymorphism. For example, *Populus*, carries a highly heterozygous genome and bears naturally occurring SV [56,72]. Many commercially important *Populus* traits are very complex and lack robust genetic regulators for precise genomic selection. Focusing exclusively on either DNA sequence polymorphism or SV is not sufficient to systematically understand these complex traits. Integrating the effects of multiple types of variation on *Populus* is needed.

Objectives

1. Characterize novel types of structural variation in plants: the cases of chromoanagenesis triggered by gamma-irradiation in poplar or in the *A. thaliana* meiotic mutant *asy1*.
2. Dissect the interaction between natural allelic variation and induced dosage variation on quantitative traits in poplar.

Dissertation outline

Chapter 1 presents an introduction to chromoanagenesis in plants and describes the opportunities and limitations of structural variation on *Populus* breeding.

Chapter 2 is a review of the currently described chromoanagenesis events in plants.

Chapter 3 is a study using deep whole genome sequencing to characterize chromoanagenesis events in 2 poplar F1 lines carrying radiation-induced genome damage.

Chapter 4 is a study using deep whole genome sequencing to characterize a chromoanagenesis event in an *Arabidopsis* hybrid line carrying the *asy1* mutation.

Chapter 5 is a study investigating the quantitative effects of natural allelic variation and induced dosage variation on phenology, biomass and leaf morphology traits in a poplar F1 population.

Chapter 6 presents a summary of the findings of this work and concluding remarks.

Chapter 2

Chromoanagenesis in plants: triggers, mechanisms, and potential impact

[Published in: Trends in Genetics]

Weier Guo, Luca Comai, Isabelle M. Henry*

Genome Center and Dept. Plant Biology, University of California Davis, Davis, California, United States of America

*Corresponding author: imhenry@ucdavis.edu

Abstract

Chromoanagenesis is a single catastrophic event that involves, in most cases, localized chromosomal shattering and reorganization, resulting in a dramatically restructured chromosome. First discovered in cancer cells, it has since been observed in various other systems, including plants. In this review, we discuss the origin, characteristics and potential mechanisms underlying chromoanagenesis in plants. We report that multiple processes, including mutagenesis and genetic engineering, can trigger chromoanagenesis via a variety of mechanisms such as micronucleation, breakage-fusion-bridge cycles or chain-like translocations. The resulting rearranged chromosomes can be preserved during subsequent plant growth, and sometimes inherited to the next generation. Because of their high tolerance to genome restructuring, plants offer a unique system for investigating the evolutionary consequences and potential practical applications of chromoanagenesis.

Chromoanagenesis: the rebirth of chromosomes

Chromosome rearrangements can be frequent, and have been described extensively in many systems [3,18,73–76]. More recently, the development of advanced sequencing technologies has facilitated the discovery of novel complex chromosomal rearrangements. Unlike previous examples of complex chromosomal rearrangements, which carry a relatively small number of translocations spread over multiple chromosomes [77], these newly observed events are more extreme [32,33,78]. Specifically, they exhibit multiple copy number variations (CNVs), which are typically clustered on a single chromosome, and most often originate from a single catastrophic event. Based on these properties, the process leading to these rearrangements is now referred to as chromoanagenesis (See Glossary), to signify the "rebirth" of the chromosome, following sudden shattering and reassembly [79,80].

Chromoanagenesis was originally described in association with human cancer [79]. It is now sub-divided into three distinct processes - chromothripsis, chromoanasynthesis and chromoplexy. All three cases involve early DNA breakage, which can result from a variety of processes, ranging from mis-segregation to replication fork arrest. In all three cases, the resulting genome displays chaotic chromosomal rearrangements, but the mechanisms differ, and the genomic outcomes are slightly different as well.

Chromothripsis

During chromothripsis one, or occasionally a few chromosomes (< 4) undergo catastrophic pulverization, whereby dsDNA breaks form tens to hundreds of chromosomal fragments that subsequently randomly religate into a new chromosome [32,81] (Fig. 2.1A). These newly formed chromosomes carry regions with clustered copy number variation, oscillating between two copy number states (occasionally three [32]), corresponding to deleted and retained DNA fragments. The retained fragments join in random order and orientation [82] through non-homologous end joining (NHEJ) and occasionally microhomology-mediated end joining [30,32,83].

Chromoanasynthesis

During chromoanasynthesis, clustered chromosomal rearrangements associated with wide copy number variation (1-4 copies or more) form on a single chromosome [33]. This outcome is difficult to explain by a simple fragmentation and re-ligation process such as what is observed for chromothripsis. Instead, error-prone DNA replication mechanisms including replication fork stalling and template switching (FoSTes), and microhomology-mediated break induced replication (MMBIR), are possibly at play [84,85]. During DNA replication, DNA breaks arrest the replication fork. Next, the single stranded portion of the broken DNA can switch to a nearby replication fork and find a template through micro-homology. Multiple cycles of template switching of the growing DNA strand can result in a mosaic chromosome with higher copy number states (Fig. 2.1B).

Chromoplexy

In contrast to chromothripsis and chromoanasynthesis, where rearrangements are highly clustered and mostly affect a single chromosome, chromoplexy involves fewer rearranged fragments and involves multiple chromosomes [34] (Fig. 2.1C). Chromoplexy results in extensive intra- and inter-chromosomal translocations, and can occur through several events that happen either all at once, or sequentially. This process results in a “closed-chain” reorganization of the involved chromosome fragments, where large pieces from several chromosomes are reattached to each other in a novel order. This sometimes produces derivative chromosomes, usually with few copy number changes and breakpoints [86]. Chromoplexy has been shown to be associated with transcriptional disruption as well [87,88].

Mechanisms underlying chromoanagenesis

The mechanisms underlying chromoanagenesis have been most extensively described in human cancer cells [79,80] and can result from various, and frequently interconnected processes. They are briefly reviewed below.

One potential model involves micronucleation. For instance, a chromosome mis-segregates, and fails to reach the pole during cell division [89]. The rest of the chromosomes are incorporated into the major nuclei, while the lagging chromosome(s) frequently and preferentially becomes isolated and captured by defective nuclear envelopes [90]. The resulting small nucleus, called a micronucleus [91], provides poor conditions for DNA replication and repair [92], and has a tendency to rupture for various reasons, including insufficient nuclear pore density, nuclear stretching and physical compression [93–96]. Rupture exposes chromatin to the cytoplasm, triggering transient and possibly localized exposure of DNA to agents that cause dsDNA breaks [93], such as cytoplasmic nucleases [97]. Micronuclei have been specifically shown to be frequently associated with chromoanagenesis, consistent with the start of cascading genome instability [98,99]. It is probable that the chromosomes inside the micronuclei undergo poor transcription, which result in single strand DNA breaks on DNA-RNA hybrids [100]. These single strand breaks can be converted into double strand breaks through aberrant DNA replication or nearby breaks on the opposite DNA strand, which results in chromosome fragmentation.

Alternatively, bridge breakage can also lead to chromoanagenesis. For instance, a dicentric chromosome, which may be formed through telomere fusion, can develop into a chromatin bridge during the mitotic division [97]. The stretching of the bridge leads to nuclear envelope rupture, subsequently resulting in bridge breakage [97,101]. The broken chromosomes in the daughter cells undergo defective DNA replication during the interphase, which can result in complex chromosomal rearrangements, including local DNA fragmentation (chromothripsis) and repeated insertions of fragments with microhomology (chromoanasynthesis) [36].

Besides telomere fusion and dicentric chromosome breakage, other processes that induce DNA breaks can also lead to chromoanagenesis. For example, ionizing radiation can induce DNA damage and further form multiple breaks on chromosomes, which resembles the consequence of chromothripsis [102]. Similarly, recent studies demonstrated that CRISPR-Cas9-induced DNA double strand breaks are able to generate chromothripsis on the targeted chromosome in human blood cells [103].

While chromothripsis and chromoanasynthesis are often associated with segregation errors and micronucleation, the molecular mechanisms underlying chromoplexy remain unclear. It was recently proposed that chromoplexy can be initiated when DNA damage occurs specifically at a location where several chromosomes aggregate, such as a transcription hub. This induces multiple dsDNA breaks on different chromosomes that are physically close to each other. The repair of these breaks may result in translocations of fragments among the broken chromosomes, resulting in a new "chain" of chromosomal segments that were previously unlinked [34,104] (Fig. 2.2).

Cases of chromoanagenesis in plants

Here, we review the various documented cases of chromoanagenesis observed in plants, and discuss the putative mechanisms leading to these events, as well as their potential consequences, both in terms of plant evolution and bioengineering (Table 2.1). To better introduce these cases, we classify them into two groups: 1. Extreme rearrangements clustered on a single chromosome; 2. Extreme rearrangements involving multiple chromosomes.

Extreme rearrangements clustered on a single chromosome

Haploid induction

The first case of chromoanagenesis in plants was observed in *Arabidopsis thaliana* plants generated from haploid induction crosses [30]. Haploid induction is a powerful plant breeding tool, resulting in the production of a haploid progeny from a cross between two diploid parents, either via the rapid loss of one parental genome after fertilization, or the production of a viable haploid offspring without fertilization. Specifically, the cross between two *Arabidopsis* lines - one carrying a mutated form of the centromeric specifier CENH3 and the other wild type -

results in frequent maternal genome elimination, and produces offspring of different types in similar numbers: paternal haploids, aneuploids, and diploids. In approximately 11% of these aneuploid individuals, the additional chromosome displays many instances of copy number variation. These patterns are fully consistent with chromoanagenesis [30]. Remarkably, the range of rearrangements also include the production of minichromosomes, i.e. chromosomal segments containing the centromere [105] but lacking most of both chromosome arms.

In all cases observed, the shattered chromosomes were always derived from the parent containing the mutated *cenh3*. One fertile F1 line was chosen for in depth characterization. Illumina short-read sequencing analysis demonstrated the presence of 38 novel DNA junctions within chromosome 1. These junctions were independently validated by PCR. They occurred on the shattered segment of the chromosome, and were localized in tight association with segments displaying CNVs. Sequence junctions fell into two categories: junctions between segments present in one and two copies (duplicated segments), and junctions between segments present in two or three copies (triplicated segments). Analysis of the sequence context surrounding the breakpoints from these novel junctions indicated that breakpoints associated with duplicated segments were significantly enriched in gene-space, while breakpoints associated with triplicated segments were significantly closer to origins of replication than expected. This suggested that the breakpoints associated with duplicated fragments more often resulted from cuts in or near genes, while the breakpoints associated with triplicated fragments were potentially associated with replicative activity. Studies have shown that breakpoints in chromothripsis result from base excision repair on DNA-RNA hybrids during poor transcription [100], confirming that they are likely to occur in genic regions. Chromoanasynthesis, on the other hand, is generally associated with error-prone DNA replication, where the growing DNA strand continuously switches among replication templates with microhomology [85]. Taken together, Tan's results are consistent with the possibility that both chromothripsis and chromoanasynthesis participated in forming these highly recombined chromosomes.

Subsequent reports on this system hint at the potential mechanism underlying these cases. The centromeres contributed by the parent carrying the mutant *cenh3* are defective compared with the wild-type one [106]. Hybridization between wild-type *Arabidopsis* and centromere-modified lines can cause chromosome mis-segregation during embryogenesis, leading to micronuclei formation [30]. Further, cytological characterization of the early embryos resulting from haploid induction crosses documented formation of micronuclei around the mis-segregating chromosomes [107]. It is thus likely that the defective centromere is unable to perform as well as the wild-type centromere, resulting in lagging of the defective chromosomes during the first mitotic divisions following fertilization [107]. These laggards are incorporated by a defective nuclear envelope, forming micronuclei [107]. We hypothesize that genome instability is triggered in these micronuclei. We further posit that, in the cases observed here, the micronucleus was partitioned in the daughter cell that had already inherited intact copies of both parental genomes, producing a trisomic embryo. During the following mitosis, the chromosome in the micronucleus underwent chromoanagenesis, resulting in the formation of a partially trisomic chromosome with severely altered organization [98]. The rupture of the micronucleus released this reassembled shattered chromosome into the major nucleus and was preserved through plant development (Fig. 2.3A, D, E).

Gamma Irradiation

Chromoanagenesis events have also been observed in mutant *Populus* individuals [38]. These *Populus* hybrids originated from an interspecific cross between wild-type egg cells from *Populus deltoides* and gamma irradiated pollen grains from *Populus nigra* [60]. The genomic constitution of these hybrids was characterized using low-pass illumina sequencing. Approximately half of the resulting hybrid genomes carried one or a few large indels [60]. This population exhibited tremendous phenotypic variation in all traits observed, and was used as a functional genomics resource to investigate the effect of dosage variation on gene function [61–63].

In two of the hybrid lines characterized (N = 592), multiple CNVs (21 and 11 CNVs on chromosome 1 and chromosome 2 of two *Populus* lines, respectively) were observed and were clustered on a single chromosome. All identified CNVs occurred on the chromosome inherited from the irradiated *P. nigra* parent, confirming the role of gamma irradiation in this localized structural variation. Deeper illumina short-read sequencing data, combined with computational analysis and PCR confirmation indicated that multiple DNA breaks and reassembly occurred on the restructured regions. These two lines exhibited significantly higher numbers of breakpoints (24 and 28 breakpoints) compared to their siblings, which showed an average of 2.5 CNVs per individual [60]. Combined with the fact that these breakpoints were clustered on a single chromosome, these results suggested that these two lines might have experienced a chromoanagenesis-like process. Finally, one of the two *Populus* lines displayed multiple copy number states, ranging from 1 to 5, indicating the occurrence of fragment deletions, duplication, triplication and even quadruplication. The high copy number states of the rearranged chromosome suggest chromoanasynthesis. Chromosome dosage variation in the second *Populus* line only exhibited three copy number states (1, 2 and 3), corresponding to deleted, neutral and duplicated states. This is consistent with chromothripsis.

The study in *Populus* reinforces the notion that ionizing radiation can cause extreme genomic damage and lead to unanticipated chromosomal rearrangement. The possibility of radiation-induced chromoanagenesis-like rearrangement has been reported in tumor cells as well. For example, Morishita's research [102] indicated that targeting nuclei with ionizing irradiation can induce chromoanagenesis-like chromosomal rearrangements.

Based on these data, we propose that mis-segregation happened during pollen development, since gamma radiation was applied on mature binucleate pollen [60]. Specifically, we propose that the irradiation treatment initiated double strand breaks in the generative cell, where the broken chromosome mis-segregated during the second pollen mitosis, and micronuclei formed in sperm cells. Later, fertilization between wild-type egg cells and irradiated sperm cells brought micronuclei into the hybrid zygotes. The shattered chromosome was produced by catastrophic restructuring that occurred within micronuclei, and the degradation of micronuclei brought it back into the major nucleus (Fig. 2.3B, D, F). Genomic structural variation has also been demonstrated to occur during pollen mitosis in maize haploid inducer lines [108]. Although the structural variations observed in that case were not as extreme, these observations provide further evidence that gametophyte development is a critical developmental stage in terms of genomic stability.

Defective meiosis

The latest identified example of chromoanagenesis in plants comes from a study aimed at understanding the role of ASY1 in *A. thaliana*. ASY1, the *Arabidopsis* homolog of the yeast chromosome axis component HOP1, plays an important role in crossover assurance and interference [109,110]. A recent study [111] investigated the outcome of a cross between Col-0/L*er*-1 hybrid *asy1* mutants (*asy1*Col-0 x *asy1*L*er*-1, female) and a wild-type Col-0 (male), and showed that one individual, out of the 176 individuals characterized, carried drastic genomic rearrangement. These rearrangements resembled the consequence of chromoanagenesis: this line exhibited multiple CNVs, all clustered within the first half of the chromosome (from 1 to 16.1Mb). In silico analysis of Illumina short-read sequencing data from this individual identified 520 novel breakpoints compared to its siblings' genomes, forming 260 novel DNA junctions [39]. These novel DNA junctions exhibit several characteristics of chromothripsis [82]. First, the rearranged region displays oscillation between two copy number states (2 and 3). Second, the novel DNA junctions involve the joining of two fragments in random orientation. Third, unbalanced chromosome segregation and micronuclei have been observed during microsporogenesis through cytological experiments in this system [111]. This case is particularly interesting because of the extremely high density of rearrangements observed, which is reminiscent of the more extreme cases observed previously in cancer cells [32].

While it is possible that this event is not associated with the presence of the *asy1* mutation, several observations are consistent with the potential for chromothripsis in the *asy1* mutant background. First, micronuclei have been observed during male sporogenesis in *asy1* mutants, but not in the control WT background [111]. While micronuclei formation has been shown to be associated with mitosis previously [96], this example indicates that micronuclei can be produced during meiosis as well. ASY1 is involved in crossover assurance and interference in *Arabidopsis*, and *asy1* mutants exhibit altered recombination patterns and unbalanced chromosome segregation during meiosis. Here, we propose that a megaspore carrying a micronucleus was present at the end of sporogenesis. During the following female mitosis, the chromosome trapped in the micronucleus underwent chromoanagenesis. When the shattered chromosome was partitioned into the egg cell, it was inherited by the zygote (Fig. 2.3C, D, G).

The three cases above are consistent with a common mechanism in which chromoanagenesis can be triggered by different events, but all result in chromosome mis-segregation and micronucleation. When the affected cell enters the next cell division, the chromosome in the micronucleus can sometimes undergo abnormal or delayed DNA replication, resulting in chromosome shattering, when the rest of the chromosomes experience regular segregation. The shattered chromosome subsequently re-integrates the main nucleus in its altered form.

Extreme rearrangements involving multiple chromosomes

Biolistic transformation

Transformation is known to trigger complex chromosomal rearrangement [112–115]. Some of these rearrangements also show features consistent with chromoanagenesis. For example, extensive genomic disruption was detected in transformed rice and maize plants [44]. Using biolistics, the authors transformed linear and circular DNA into regenerable calli of rice (*Oryza sativa*) and maize (*Zea mays*) and characterized the transformed plants using whole genome sequencing. Three rice transgenic lines exhibited intra- and inter-chromosomal translocations

(14, 28, 107 breakpoints in each line, respectively), as well as hundreds of broken chromosomal segments interlinked with transformed DNA. Multiple chromosomes (≥ 6) were involved in these complex rearrangements. Similarly, three maize transgenic lines also exhibited genomic deletions and duplications, although with lower number of breaks and reassembly events compared to the rice individuals.

Identification of large numbers of breakpoints in this study implies its possible correlation with chromoanagenesis. The involvement of multiple chromosomes suggests it may not be the consequence of chromothripsis or chromoanasynthesis but the presence of multiple copy number states (1-4 states) is not consistent with chromoplexy either. It is possible that the restructuring observed in these individuals comes from a combination of transformed DNA insertion and catastrophic chromosome rearrangements.

The authors proposed that the founding of targeted cells by metal particles severely damaged the nuclear envelope and caused exposure of nuclear DNA to cytoplasmic components, which can induce DNA breaks. Damaged DNA may have undergone imperfect repair, resulting in ligation of transformed DNA with multiple genomic fragments. Many other studies have reported that T-DNA insertion can induce genomic rearrangements, including translocation, inversion and deletion [42,112,116,117]. The rearrangements of the affected chromosomes are not usually as severe as those reported for biolistic transformation [44], but may result from a similar mechanism. Alternatively, T-DNA integration may be more likely in cells that are experiencing increased instances of dsDNA breaks that are being repaired [118].

Protoplast regeneration

Another chromoanagenesis-like event was reported by Fossi et al, where extreme chromosomal rearrangements were observed in the genome of potato (*Solanum tuberosum*) plants regenerated from protoplasts [41]. In Fossi's report, 3/15 potato plants regenerated from protoplasts displayed multiple deletions and duplications affecting single chromosomes [41], consistent with chromoanagenesis. In-depth analyses of these individuals will be required to investigate the mechanism underlying these events (Dr. Kirk Amundson, personal communication).

Natural somatic variations

Catastrophic chromosomal rearrangements can also result from spontaneous somatic variation in plants. Carbonell-Bejerano et al. (2017) reported a chromothripsis-like pattern in a somaclonal variant of grapevine (*Vitis vinifera*) [37]. The study compared a somatic variant (Tempranillo Blanc, TB) with its ancestor (Tempranillo Tinto, TT), and demonstrated that TB harbors complex chromosomal rearrangement including 6 novel junctions spread over 3 chromosomes. The rearranged regions in the TB variant are composed of alternating monosomic and disomic fragments. The authors proposed that this resulted from chromothripsis based on statistical analyses.

The author of this study proposed that this event was induced by breakage-fusion-bridge (BFB) **cycles** because one of the rearranged chromosomes exhibited pseudodicentric characteristics [119], that are expected to induce DNA breaks during cell division. However, another group of studies have shown that breakage-fusion-bridge cycles do not always result in extreme chromosomal rearrangement [120–123]. It is possible that the rearranged segments in this case were initialized by one BFB cycle followed by secondary rearrangements, such as micronucleation [36]. Alternatively, considering the small number of novel junctions, and the fact that several chromosomes were involved, these rearranged segments may be the results of chromoplexy, or may not be associated with chromoanagenesis at all (Fig. 2.2).

Plant genome evolution

Several studies have shown footprints of chromosomal rearrangements, and proposed extreme rearrangements as one of the mechanisms contributing to plant chromosome evolution. For example, Mandakova et al (2019) reported that chromosome shattering was involved in the evolution of the *Camelina* genome [124]. Specifically, the authors proposed that the three chromosomes present in the ancestral *Camelina* genome were reorganized into a single mosaic chromosome in the *Camelina* diploid variety. Similarly, translocation, inversion and centromere repositioning all contributed to the emergence of the *Cucumis* genome from multiple ancestral chromosomes [125]. In these examples, species-specific karyotype variations are initiated by whole genome duplication, and followed with multiple chromosomes merging and breaking events. These papers propose that these rearrangements that occurred during plant genome evolution exhibit chromothripsis-like patterns, although they seem closer to the outcomes of chromoplexy, with interspersed distribution of rearrangements among multiple chromosomes (Fig. 2.1C).

Another process mentioned in these publications is descending dysploidy, which is an important diploidization process ultimately resulting in lower base chromosome number [124–126]. Specifically, chromosomal rearrangements, including reciprocal translocations, inversions and centromere inactivation/elimination are suggested to occur step-wise, and result in descending dysploidy [126]. The plant genomes that underwent descending dysploidy display restructuring footprints similar to those found in the genomes of tumors that underwent chromoplexy [81,126], suggesting that the two processes may be related and involve similar mechanisms. On the other hand, several other studies describing descending dysploidy report cases of reciprocal large-scale translocations and chromosomes fusions [127–129], but not the more complex chromoplexy-like process which undergoes “chain-like” translocations on more than three chromosomes [104]. Overall, it is possible that chromoplexy contributes to genome evolution, but if so, it is probably a low-frequency event.

Properties of chromoanagenesis in plants

Besides the common features of chromoanagenesis, including multiple copy number variation clustered in localized regions, additional properties were uncovered by comparing the events described above.

Sequence context

CNV breaks or novel sequence junctions are statistically preferentially localized in gene-rich regions. This was characterized in details in at least three cases: in the events triggered by haploid induction in *A. thaliana* [30], in the *Populus* individuals produced from gamma-irradiated pollen [38], and in the progeny of the *asy1* mutant in *A. thaliana [39]*. Interestingly, both in the haploid induced *Arabidopsis* and *Populus* examples, 50% of novel breakpoints directly affected a gene coding sequence [30,38]. This may induce loss of function in multiple genes, or create potential novel genes (see Outstanding Questions). Moreover, novel DNA junctions from the progeny of *asy1* mutated *Arabidopsis* were significantly enriched in chromatin states associated with transcription [39]. This result was reminiscent of the micronucleus-related chromosome fragmentation mechanism in tumor cells, where chromosome breaks were induced from poor transcription in micronuclei, typically on excessive accumulated DNA-RNA hybrids [100].

Aneuploidy

In several of the cases described above, the crosses or systems that triggered the events involved the formation of aneuploid individuals and, specifically trisomic (or disomic) plants, in which the additional chromosome was shattered. Specifically, in *A. thaliana*, haploid induction crosses produce up to ⅓ aneuploid individuals [30]. Similarly, the *asy1* mutation results in altered meiosis, the presence of laggards, and production of aneuploid individuals [111]. This begs the question of whether chromoanagenesis can be triggered merely by the presence of extranumerary chromosomes, or the presence of laggards at meiosis.

In both cases, the chromoanagenetic lines exhibited copy number oscillations between 2 and 3 (or between 1 and 2 in the case of disomic individuals in a haploid background). This suggests the presence of two (or one) intact chromosomes, and one additional chromosome that had undergone shattering and subsequently lost seemingly random fragments. The addition of an extra chromosome or some portion of the chromosome have been observed in other genome instability events as well, which did not lead to chromoanagenesis [105].

The case of *Populus* is slightly different, with some pieces of the shattered chromosome present in only one copy in a trisomic (2n + 1) background. In both cases, sequence analysis using parent-specific SNPs confirmed that the modified chromosomes originated from the gamma-irradiated pollen, and not the untreated female parent. There are at least two possible explanations for this second observation. One possibility is that the two sister chromatids of this chromosome, both originating from the pollen parent, participated in the chromoanagenesis event. Alternatively, it is possible that one of the paternally inherited chromosomes carried a deletion formed independently of the chromoanagenesis event. In this population of mutants, approximately half of the individuals carry at least one large-scale deletion or insertion [60]. It is therefore plausible that this is an independent event.

Inheritance

Structurally rearranged chromosomes may cause synapsis failure during meiosis, and are typically poorly inherited in animal systems [130,131]. The inheritance of chromoanagenesis events is rare but has been demonstrated in humans [132]. Similarly, in at least one of the chromoanagenesis events observed after haploid induction in *A. thaliana,* the shattered chromosome was transmitted sexually to the offspring [30]. These results demonstrate that meiosis can proceed despite the presence of extensively rearranged chromosomes, and that the rearranged chromosomes can be transmitted under some circumstances. These events may have potential applications to identifying novel traits for plant breeding, or for investigating human diseases.

Concluding Remarks and Future Perspectives

Plants provide an attractive system for studying extreme chromosomal rearrangements. Aberrant chromosomes can persist both in tissue culture and in mature plants, and, in some cases, be transmitted sexually to the new generation [30]. This provides a unique system to investigate the phenotypic consequences and the meiotic behavior of the novel chromosomes. Additionally, relatively complex rearrangements, which would likely be lethal in animal systems, can be explored more easily in plants because they are generally more tolerant of copy number variation and aneuploidy [133,134]. This is particularly attractive in systems with abundant genetic resources, such as *A. thaliana*, where characterized mutant collections and rich diversity, facile transformation, and other sophisticated tools are available to investigate the effect of specific pathways and genes on the processes leading to chromoanagenesis.

Chromoanagenesis has been associated with multiple biotechnological manipulations, such as irradiation [38], biolistic transformation [44], or protoplast regeneration [41], all of which are common approaches used for genetic analysis and engineering. These findings highlight the fact that severe chromosomal rearrangements, including chromoanagenesis may be unintentionally triggered during plant breeding processes and may alter the genome of the resulting plants significantly. The bioengineering potential of chromoanagenesis is worth investigating further (see Outstanding Questions). For example, gene amplification has been shown to result in resistance to the herbicide glyphosate [22]. Gene amplification is likely mediated by extensive chromosome instability, and could be explored at the organismal level [135]. Depending on the case, minichromosomes, a by-product of instability, and which can be triggered by various processes including haploid induction [105], pollen irradiation [136] and artificial engineering

[28], could potentially result from chromoanagenesis as well.

Acknowledgements

This research was supported by the US Department of Energy Office of Science, Office of Biological and Environmental Research, Grant DESC0007183 to LC.

Tables and Figures

Table 2.1. Summary of chromoanagenesis-like events in plants

Groups	**Events**	**Types of chromoanagenesis**	**Literatures**
Extreme rearrangements clustered on a single chromosome	Haploid induction cross in *Arabidopsis*	Chromothripsis, chromoanasynthesis	Tan et al. 2015
	Gamma irradiation in *Populus*	Chromothripsis, chromoanasynthesis	Guo et al. 2021
	asy1 mutation in *Arabidopsis*	Chromothripsis	Guo et al. 2022
Extreme rearrangement involving multiple chromosomes	Biolistic transformation in *Oryza sativa* and *Zea mays*	Undetermined[1]	Fossi et al. 2019
	Protoplast regeneration in potato (*Solanum tuberosum*)	Undetermined[2]	Liu et al. 2019
	Natural somatic variation in grapevine (*Vitis vinifera*)	Chromoplexy[3]	Carbonell-Bejerano et al. 2017
	Genome evolution in *Camelina* and *Cucumis*	Chromoplexy[4]	Mandáková et al. 2019; Zhao et al. 2021

[1]: The authors classified the event as chromothripsis. Based on the results and proposed mechanism, it is unclear which chromoanagenesis type this event belongs to.

[2]: The chromoanagenesis-like events were not characterized in detail.

[3]: The authors classified the events as chromothripsis based on statistical analysis. Considering the number of rearrangements and underlying mechanism, this case may result from chromoplexy instead.

[4]: The authors classified these events as chromothripsis. After comparing with the characteristics of three processes in chromoanagenesis, the authors of this review propose that these events might be more consistent with chromoplexy.

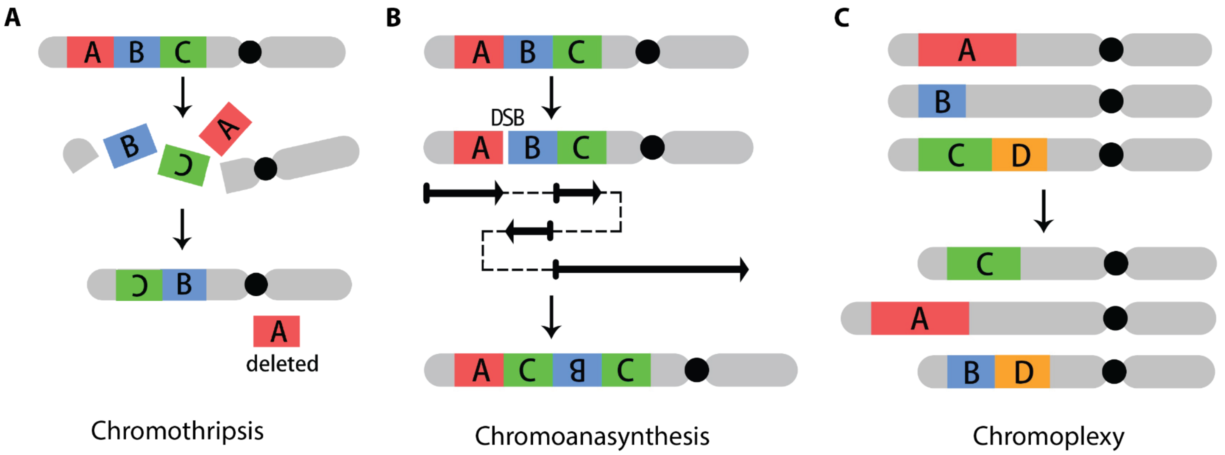


Figure 2.1. Schematic diagrams of chromothripsis, chromoanasynthesis and chromoplexy. (A) Chromothripsis is a catastrophic event where a single chromosome is pulverized into tens to hundreds of fragments and a subset of the pieces are randomly reassembled together. (B) Chromoanasynthesis is associated with aberrant DNA replication, where the replication fork gets arrested by a double strand break, and subsequently leads to continuous switching of the templates with microhomology. It produces a single rearranged chromosome with wide copy number variation (shown here as one copy of A and B and two copies of C). (C) Chromoplexy describes translocations involving multiple chromosomes, with few copy number changes.

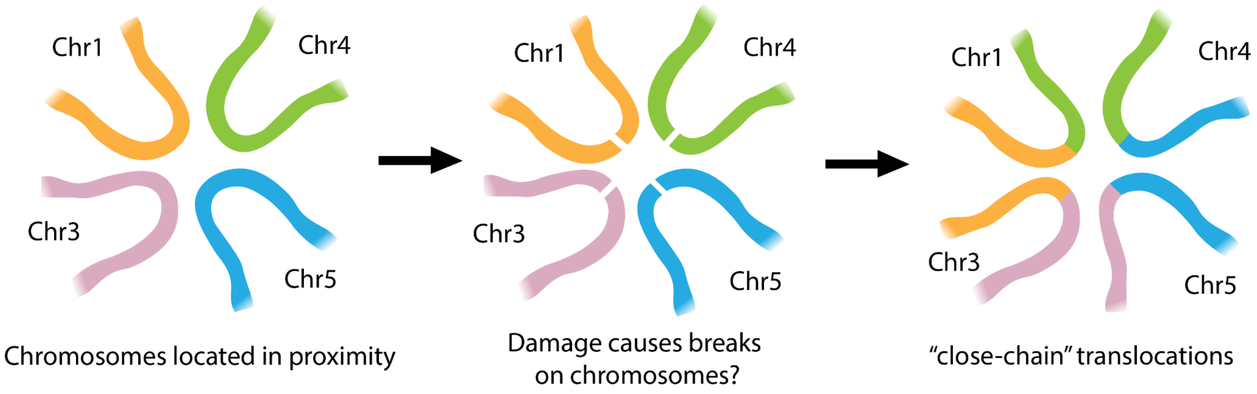

Chr1
Chr4
Chr3
Chr5
Chromosomes located in proximity
Chr1
Chr4
Chr3
Chr5
Damage causes breaks on chromosomes?
Chr1
Chr4
Chr3
Chr5
"close-chain" translocations

Figure 2.2. Potential mechanism of chromoplexy. Chromoplexy may be initialized when unknown damage causing double strand breaks is applied to chromosomes located in proximity to each other. DNA repair results in "closed-chain" translocations between these chromosomes.

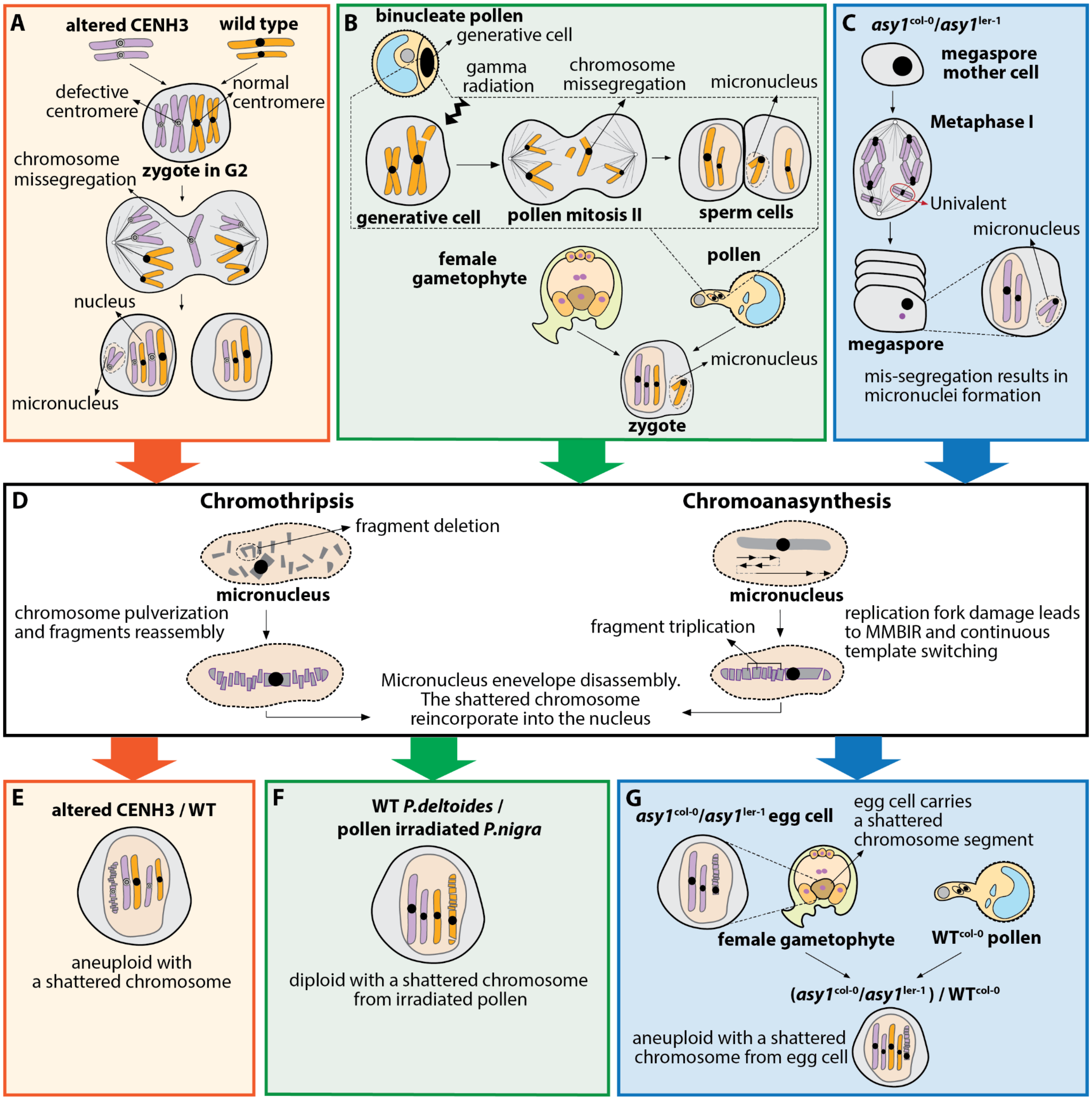

A
altered CENH3
wild type
defective centromere
normal centromere
zygote in G2
chromosome missegregation
nucleus
micronucleus
B
binucleate pollen
generative cell
gamma radiation
chromosome missegregation
micronucleus
generative cell
pollen mitosis II
sperm cells
female gametophyte
pollen
micronucleus
zygote
C
$asy1^{col-0}/asy1^{ler-1}$
megaspore mother cell
Metaphase I
Univalent
micronucleus
megaspore
mis-segregation results in micronuclei formation
D
Chromothripsis
fragment deletion
micronucleus
chromosome pulverization and fragments reassembly
Chromoanasynthesis
micronucleus
fragment triplication
replication fork damage leads to MMBIR and continuous template switching
Micronucleus envelope disassembly. The shattered chromosome reincorporate into the nucleus
E
altered CENH3 / WT
aneuploid with a shattered chromosome
F
WT P.deltoides / pollen irradiated P.nigra
diploid with a shattered chromosome from irradiated pollen
G
$asy1^{col-0}/asy1^{ler-1}$ egg cell
egg cell carries a shattered chromosome segment
female gametophyte
WT^{col-0} pollen
$(asy1^{col-0}/asy1^{ler-1}) / WT^{col-0}$
aneuploid with a shattered chromosome from egg cell

Figure 2.3. Model of the different triggers and potential mechanisms leading to chromoanagenesis in plants. (A-C) Origin and triggers of chromoanagenesis, including haploid induction crosses in *A. thaliana* (A), pollen-irradiation in *Populus* hybrids (B) and meiotic abnormalities in the *asy1* mutant in *A. thaliana* (C). (D) The micronucleus-incorporated chromosome undergoes extreme rearrangements, typically undergoing chromothripsis and/or chromoanasynthesis. (E-G) Events occurring after micronucleus envelope disassembly. (A) In the *A. thaliana* haploid induction crosses, chromosomes from the maternal genome carry altered CENH3 proteins, which leads to defective centromeres. These chromosomes can lag and are sometimes enclosed in a micronucleus. If the micronucleus is partitioned into the daughter cell that includes the intact genome, it creates a trisomy. After extreme chromosomal rearrangement (D), it results in an aneuploid zygote with one intact genome and an extra shattered chromosome (E). (B) DNA damage in generative cells is induced from gamma-radiation of binucleate

pollen. It can result in chromosome lagging or bridge formation during pollen mitosis 2. A micronucleus is sometimes produced within the sperm cells, and later brought into the zygote through fertilization. The chromosome in the micronucleus undergoes severe reorganization, and the collapse of the micronucleus brings the shattered chromosome into the major nucleus (D). It finally produces a diploid with one paternally-inherited chromosome exhibiting severe rearrangements (F). (C) *asy1* mutants experienced unbalanced chromosome segregation during meiosis, which resulted in the formation of micronuclei in megaspores. Chromosomal rearrangements occur during megagametogenesis (D). The shattered chromosome can be preserved into the offspring if it is kept by the egg cell after micronucleus disassembly (G).

Chapter 3

Chromoanagenesis from radiation-induced genome damage in *Populus*

[Published in: PLOS Genetics]

Weier Guo, Luca Comai, Isabelle M. Henry*

Genome Center and Dept. Plant Biology, University of California Davis, Davis, California, United States of America

*Corresponding author: imhenry@ucdavis.edu

Abstract

Chromoanagenesis is a genomic catastrophe that results in chromosomal shattering and reassembly. These extreme single chromosome events were first identified in cancer, and have since been observed in other systems, but have so far only been formally documented in plants in the context of haploid induction crosses. The frequency, origins, consequences, and evolutionary impact of such major chromosomal remodeling in other situations remain obscure. Here, we demonstrate the occurrence of chromoanagenesis in poplar (*Populus sp.)* trees produced from gamma-irradiated pollen. Specifically, in this population of siblings carrying indel mutations, two individuals exhibited highly frequent copy number variation (CNV) clustered on a single chromosome, one of the hallmarks of chromoanagenesis. Using short-read sequencing, we confirmed the presence of clustered segmental rearrangement. Independently, we identified and validated novel DNA junctions and confirmed that they were clustered and corresponded to these rearrangements. Our reconstruction of the novel sequences suggests that the chromosomal segments have reorganized randomly to produce a novel rearranged chromosome but that two different mechanisms might be at play. Our results indicate that gamma irradiation can trigger chromoanagenesis, suggesting that this may also occur when natural or induced mutagens cause DNA breaks. We further demonstrate that such events can be tolerated in poplar, and even replicated clonally, providing an attractive system for more in-depth investigations of their consequences.

Author summary

Plant breeders often use radiation treatment to produce variation, with the goal of identifying new varieties with superior traits. We studied a population of poplar trees produced by gamma irradiation of pollen, and asked what kind of DNA changes were associated with this variation. We found many changes, most often in the form of added (insertions) or removed (deletions) pieces of DNA. We also found two lines with much more drastic changes. In those lines, we observed massive reorganization. We characterized these two lines in detail and found that catastrophic pulverization and random reassembly only occurred on a single chromosome. Looking closely at how the pieces were put back together suggest that the rearrangements in these two lines may have resulted from two slightly different mechanisms. This type of rearrangement is commonly observed in human cancer cells, but has rarely been observed in plants. We demonstrated here that they can be induced by gamma irradiation, indicating this type of event might be more widespread than we expected. Characterizing such genome restructuring instances helps to understand how genome instability can remodel chromosomes and affect genome function.

Introduction

Genomic structural variation (SV) includes various types of chromosomal rearrangements, such as insertion, deletion (INDEL), copy number variation (CNV), inversion and translocation. Structural variation can produce evolutionary significant variation, because it can affect large regions of the genome, and influence multiple traits at once. In one extreme scenario, restructuring of the genome results in clustered CNV affecting a single or a few chromosomes, a syndrome called chromoanagenesis. Chromoanagenesis results from a single triggering event that leads to highly complex segmental rearrangements [79,137]. The extreme restructuring of a single chromosome (or rarely two or more) results from two distinct processes: (i) in chromothripsis dsDNA breaks and Non Homologous End Joining rearrange tens to hundreds segments, with oscillations between two copy number states (occasionally three) [32,82], and (ii) in chromoanasynthesis, replication forks stalled at DNA breaks switch templates, resulting in segmental duplication and triplication events combined with complex chromosomal rearrangement of the implicated and intervening segments [138]. Chromothripsis and chromoanasynthesis are associated with missegregation of chromosomes, followed by micronucleus formation around a single chromosome, leading to a single, catastrophic pulverization event [91]. A third type of restructuring classified under chromoanagenesis differs in mechanism and outcome: during chromoplexy, chromosomes are broken in pieces, shuffled together and reassembled, resulting in rearranged chromosomes. Chromoplexy always affects more than one chromosome [78]. Chromoplexy can occur sequentially and may be originally related to DNA breaks caused by transcription factor binding [34]. In plants, chromoplexy-like events have been observed in natural variants in camelina [37], and also as a consequence of plant transformation in *Arabidopsis*, rice and maize [44].

Chromothripsis and chromoanasynthesis were originally identified in human cancerous cells [79]. To distinguish them from indels, precise criteria are applied [36,139]. In plants, there are multiple cases of extensive genomic rearrangements [44,124], but when applying the important criterion of highly frequent and clustered (at least 10) rearrangements within a single chromosome, only haploid induction crosses in *Arabidopsis thaliana* display catastrophic chromosomal reconstructing patterns [30]. In these haploid induction crosses, both chromothripsis and chromoanasynthesis were detected, and the early zygotic divisions are often also accompanied by the formation of micronuclei [140], another diagnostic feature of chromothripsis [36,79].

A critical step in the plant life cycle is pollen production and fertilization. Pollen is prone to natural mechanisms that break DNA [108,141] and it is also a classical target for chemical and radiation mutagenesis [142]. While traditional chromosomal rearrangements have been described, the range of variation resulting from these mechanisms, however, has not been determined. We decided to address this question in a poplar F1 population that we previously developed from an interspecific cross using gamma-irradiated pollen. This population was characterized genetically and harbors >650 unique large-scale insertions and deletions, ranging from a few hundred kbp to entire chromosomes. Cumulatively, these indels cover the genome multiple times [60]. To investigate whether gamma irradiation could have also resulted in more severe genome reorganization events, we screened this population for signs of clustered copy number variation patterns. We identified two individuals with genomic patterns reminiscent of chromoanagenesis, which we characterized in detail. Our results indicate that DNA breaks induced by irradiation triggered single chromosome fragmentation and restructuring patterns consistent with chromoanagenesis. These results suggest that pollen DNA breaks, either natural or induced, can produce extreme structural variations that may provide evolutionary innovation and, in perennial plants such as poplar, where we were able to preserve the chromoanagenesis outcomes by vegetative propagation, provide an attractive system for long-term investigation of the outcome of chromoanagenesis.

Results

Gamma irradiation can result in chromoanagenesis in poplar

A poplar *P. deltoides* x *P. nigra* F1 hybrid population was developed previously [60], and characterized using low-coverage illumina genome sequencing. In this population, ~58% of the lines carried large-scale genomic insertions and deletions (indels) [61], induced by gamma-irradiation of pollen grains before fertilization. Each F1 line was characterized by a unique set of indels randomly distributed along the 19 chromosomes of the poplar genome [60,61]. Interestingly, two of these lines exhibited dosage variation consistent with chromoanagenesis. Specifically, they displayed multiple clustered CNVs on a single chromosome. To investigate the mechanisms that resulted in these extreme genomic rearrangements, we selected 9 lines for further analysis: the 2 lines exhibiting extreme rearrangements (Shattering Group, Fig. 3.1C), 4 lines with limited number of indels (Lesion Group, Fig. 3.1B), and 3 lines with no apparent dosage variation (No-lesion Group, Fig. 3.1A). Genomic DNAs from these 9 lines were sent for higher coverage Illumina genomic sequencing (coverage 25-50), with the goal of characterizing dosage variation in detail, especially those lines with shattered chromosomes.

The dosage variation patterns obtained using the deep-sequencing reads were consistent with their corresponding low-coverage data. Also consistent with previous results [60], parental allele frequencies from our high-coverage data indicated that all indels in the genome of the 9 selected individuals originated from loss or gain of the paternal *P. nigra* copy (Fig. 3.2), confirming that the irradiated *P. nigra* pollen caused dosage variation.

Both lines in the Shattering Group fit our definition of clustered changes (>10 events per chromosome arm) (Fig. 3.2A and 3.2B and Table S3.2). POP33_31 exhibited 21 CNVs on Chromosome 1, including 2 deletions and 19 insertions, of sizes ranging from 10kb to 5.7Mb (Fig. 3.2A and Table S3.2). Among the clustered CNVs, we observed multiple copy number states, ranging from 1 to 5 (Fig. 3.2A and Table S3.2), suggesting that some fragments had been lost, while others went through duplication, triplication or even quadruplication. The second individual in the Shattering Group, POP30_88, only exhibited single-copy dosage variation. Specifically, 11 CNVs were found in this line, including 3 deletions and 8 duplications, all localized on Chromosome 2. These fragments ranged in size, from 80kb to 10.7Mb (Fig. 3.2 and Table S3.1 and S3.2). Taken together, these results suggested that chromoanagenesis is a possible outcome of gamma-irradiation. The different copy number variation patterns observed in these two lines further suggest that these two rearranged genomes might have been shaped by different rearrangement mechanisms.

Novel DNA junctions can be detected using high-coverage short-read sequencing

To further confirm the hypothesis that these two lines underwent chromoanagenesis, we aimed to characterize their genome structure in detail (Fig. 3.3). Specifically, we sought to characterize the patterns of genome restructuring by searching for novel DNA junctions created with the observed rearrangements. To identify these novel junctions, we searched for sequencing reads with ends that mapped to two different positions within the genome, suggesting that these two sequences are now adjacent in the reconstructed genome. Because these rearrangements are expected to occur randomly, these junctions should be unique to each line. The boundaries of the indels described above provide prime candidates for the localization of novel junctions, but other locations in the genome are possible as well. Once potential junctions were identified, the exact position of the breakpoints were determined through *de novo* assembly of the corresponding sequencing reads.

Consistent with our expectations, multiple potential junctions were identified from both of the lines exhibiting shattered chromosomes, but overall fewer were identified for the other lines (Table 3.1). We next validated the presence of these junctions using PCR amplification followed by Sanger sequencing, and using sibling F1 lines as negative controls. For the two lines in the Shattering Group, 26/33 assembled potential junctions were validated by PCR. On the other hand, none of the potential junctions from the Lesion Group (0/22) and No-lesion Group (0/11) were validated. Junctions were determined as invalid if they could be amplified from the genome of other sibling lines as well, or if the Sanger sequencing results were not consistent with the expectation. In total, we identified 26 novel DNA junctions, all of which originated from the two shattered lines (Table S3.3).

Extreme genomic rearrangements are associated with intra-chromosomal junctions

By using the junction detection approach mentioned above, we observed multiple novel DNA junctions in the lines containing a shattered chromosome (Fig. 3.4A). We next seeked to characterize them further and attempted to reconstruct the rearranged sequences.

First, we documented the genomic localization of the validated junctions in each line. Junctions and dosage variation data were plotted together on Circos Plots (Fig. 3.5). For both of the lines exhibiting shattering, all of the junctions were located on a single chromosome, whether they corresponded to a shift in dosage variation or not (Fig. 3.5A and 3.5B). In POP33_31, only 2 breakpoints (each junction consisted of two breakpoints) occurred on regions with no detected dosage variation, while the other 22 overlapped with CNV boundaries (Fig. 3.5C and Table S3.3). However, in POP30_88, only 12/28 breakpoints corresponded to CNV regions (Fig. 3.5D and Table S3.3). This suggests that the mechanisms underlying the rearrangements in these two lines might differ. Based on the orientation of two junction ends, we observed that 17% and 36% of the junctions involved an inverted fragment in POP33_31 and POP30_88, respectively (Table S3.3). Finally, we observed three different types of junctions based on the sequence structure of each junction: microhomology, perfect joining, and insertion (Fig. 3.4B, C, D). Both shattered lines exhibited all three junction types (Fig. 3.4A).

With the exact genomic position and orientation of two breakpoints in each junction, we were able to partially reconstruct the structure of the rearranged sequences in the two shattered lines. Specifically, we were able to construct 9 and 12 rearranged chromosomal pieces for POP33_31 (Chromosome 1) and POP30_88 (Chromosome 2), respectively (Figs 3.6 and S3.1). In both cases, our results suggest that the restructured region underwent extreme fragmentation, with chromosomal fragments joined together in a seemingly random order, some fragments lost altogether, and others copied multiple times (Figs 3.6 and S3.1).

The novel DNA junctions are enriched in gene-rich regions

To investigate the DNA context around the novel junctions identified in the shattered lines, we asked whether the junctions occurred more often in genic or repeated regions of the genome. Every validated novel junction contained two breakpoints. For each breakpoint, we calculated the enrichment ratio (see Material and Methods) of genomic features. We used two different window sizes, 10kb and 100kb, for investigating gene contents and repeated elements. For both of the shattered lines, breakpoints occurred significantly more often in gene rich regions and significantly less often near repeated elements (Fig. 3.7 and Table S3.6). These results were consistent with previous studies of aneuploid *Arabidopsis thaliana* individuals carrying shattered chromosomes, which also indicated that breakpoints were more likely to occur in gene-rich regions [30]. Additionally, 26 breakpoints (50%) occurred within a gene coding sequence (11/24 for POP33_31; 15/28 for POP30_88, Table S3.4), and 14 of these involved gene to gene fusion (Table S3.5). Breakpoints were found on different genic elements, including coding region, introns and untranslated regions. Genes of various functions were affected by these breakpoints (meiosis-specific proteins, dynamin, etc, see Table S3.4). These results indicate that novel DNA junctions induced by irradiation had the potential to dramatically influence the function of multiple genes at once.

Discussion

We identified and characterized two instances of highly clustered CNVs on a single chromosome in poplar F1 hybrids that resulted from interspecific crosses using gamma-irradiated pollen. To investigate the structure of these extreme genome rearrangements, we characterized the candidate chromosomes from two individuals, and identified localized shattering and rejoining of DNA in each. Specifically, we identified and characterized 12 and 14 novel DNA junctions in these two lines, which were clustered on a single chromosome, and always appeared in the shattered genomic region. These observations are consistent with the characteristics of chromoanagenesis, which is a catastrophic event creating large numbers of complex rearrangements on one or a few chromosomes [79]. They also suggest that gamma-irradiation of pollen can result in chromoanagenesis-like patterns in poplar. In our population, we observed shattered chromosomes in 2/592 individuals. The two poplar lines carrying the shattered chromosomes did not exhibit significant phenotypic differences compared to their siblings. One of the two was sufficiently robust to be selected amongst the F1 individuals that were clonally propagated and transferred to a field for a population-wide phenotyping experiment [61,62], and did not exhibit extreme phenotypic behaviors in the traits analyzed.

To date, extreme chromosomal rearrangement have only been reported in a few plant species, including in aneuploid *Arabidopsis thaliana* individuals originating from haploid induction crosses [30], in maize and rice individuals that have undergone biolistic transformation [44], and in somatic variants of grape [37]. But, except for Tan's reports in *Arabidopsis,* which reported the observation of extreme DNA damage on a single chromosome, other reports described genomic restructuring involving multiple chromosomes and thus fitting chromoplexy [78]. Our study and Tan's study are the only two that showed evidence of clustered, single chromosomal rearrangement in plants, thus fitting the definition of chromothripsis and chromoanasynthesis [82,138].

The two shattered poplar lines both carry highly clustered breakpoints but differ in other ways, suggesting that the mechanisms underlying these events might be different. The line carrying a shattered Chromosome 1 (POP33_31), exhibits a wide variation in copy number states, ranging from 1 to 5, which indicates segmental duplication and triplication during the genomic remodeling. This is consistent with the replication-based complex rearrangements of chromoanasynthesis [138]. During chromoanasynthesis, the replication fork stops, and the polymerizing strand switches to a proximate template with micro-homologous sequences, and finally causes the formation of a complex chromosomal rearrangement involving multiple copy number states [33]. On the other hand, several features of the shattered chromosome of POP30_88 suggest that it is more likely to be the result of chromothripsis, the fragmentation and random reorganization of one or a few chromosomes [143]. First, the shattered chromosome of POP30_88 only exhibits three copy number states (1, 2 or 3), which is consistent with the limited copy number states observed in chromothripsis. Chromothripsis usually exhibits two copy number states: the lower one represents fragment deletion, and the higher one represents fragment retention [32]. Occasionally, it can also carry three copy number states. This can be caused by the partial duplication of the rearranged chromosome after experiencing chromothripsis [32]. The oscillation of three copy number states in POP33_88 Chromosome 2 suggests that it may have undergone chromothripsis, followed by a segmental duplication. Second, the novel DNA junctions observed in POP30_88 cover all four types of the orientations (H-T, T-H, H-H, T-T), and the rearranged fragments order appears random. This feature can also be potential evidence for chromothripsis, since the randomness of fragments orientation and order is a representative property for this type of catastrophic event as well [82]. Altogether, our results suggest that the two chromosomal rearrangements observed might have originated from two different mechanisms: chromoanasynthesis for POP33_31 and chromothripsis for POP30_88.

Ionizing radiation has a long-standing role in plant mutation breeding [144]. The genomic consequences of ionizing mutations depend on tissue type [145], radiation dosage, and type of ionizing mutations, and can produce many different types of mutations [60,146–148], including the creation of variants exhibiting potentially favorable characteristics [61,62]. Ionizing radiation has also been proposed as a potential trigger of chromoanagenesis [32,149,150]. Finally, localized ionizing radiation targeting the nuclei of tumor cells was shown to induce chromoanagenesis-like patterns in those cells [102]. In this experiment, the authors used a microbeam system to precisely target the nuclei and induce double strand breaks in some chromosomes. Their study reported 14 *de novo* junctions involving four chromosomes, and proposed that targeted irradiation induced chromothripsis on a few chromosomes. Based on the features of the three types of chromoanagenesis events, Morishita's results suggest that their lines underwent chromoplexy, since the novel junctions are sparsely distributed on several chromosomes. Yet, it is also possible that, if the beam only targets a portion of the nuclei, only the chromosomes located in the affected area underwent rearrangement.

In contrast, our study reported highly clustered novel DNA junctions in a limited genomic region, while the initial irradiation treatment targeted whole desicated pollen grains [60]. Formation of extreme rearranged chromosomes by chromoanagenesis occurs over at least two mitotic divisions: during the first mitosis, a broken chromosome lags during anaphase and is incorporated into a micronucleus. During the following interphase, DNA replication of the micronucleus chromosome is delayed compared with the chromosomes in the major nucleus. During the second mitotic divisions, the replicating micronucleus chromosome pulverizes and reassembles randomly, forming a shattered chromosome, which is then incorporated into the normal set [79].

In poplar, mature pollen is binucleate, and must undergo the second pollen mitosis, in which the generative cell divides into two sperm cells, just before fertilizing the egg cell. It is thus possible that the radiation-induced DNA breaks remained unrepaired, causing chromosome missegregation during the generative cell division, possibly resulting in the formation of a micronucleus in one of the sperm cells (Fig. 3.8). After fertilization of the egg cell by the micronucleus-carrying sperm, during the first zygotic mitotic division, damage, such as incomplete replication, results in catastrophic DNA pulverization of the chromosome in the micronucleus [79,91]. The rearranged chromosome is reincorporated in the normal set during the subsequent mitosis. If the centromere is present, the shattered chromosome can segregate normally in the main nucleus, fixing the rearrangement.

Our study shows that novel DNA junctions were significantly enriched in gene-rich regions, which is consistent with Tan's results in *Arabidopsis* [30]. Similar outcomes have also been demonstrated in human breast cancer, where high density of DSBs occurred on chromosome 17, one of the human chromosomes with high gene content [151]. Further, open chromatin may be more available for recombination. In our analysis, 14/26 breakpoints formed junctions between genes, suggesting the potential of these events for genic innovation.

Our analysis used Illumina short reads to identify and assemble novel DNA junctions. In the shattered lines, this approach was successful as 78% of novel DNA junctions could be validated *in vitro*. Based on the number of copy number variation boundaries found in these two lines, and the number of novel breakpoints (each junction contains two breakpoints) that match these boundaries, we can estimate that we successfully identified 68% of the novel breakpoints. The false negative breakpoints could be caused by poor read coverage across those genomic regions, by the presence of repeated sequences complicating the read mapping process, or by differences between the reference genome used for read mapping (*P. trichocarpa)* and that of the male parent (*P. nigra).* When applying this approach to sibling lines containing sparse indels along the genome, we did not identify any novel breakpoints despite the presence of seven large-scale insertions in these lines, which indicates that at least 7 new breakpoints should be present. As a result, it is still unclear where the duplicated fragments detected in these lines are located. Given that the probability of identifying real breakpoints in the two lines displaying shattering was 0.68 (34 breakpoints / 50 copy number shifts), our failure to find any real breakpoint out of 7 copy number shifts for the normal indels is surprising (p-value of Bootstrap hypothesis testing = 0.0081). It is thus possible that breaks giving rise to indels may result from a different DNA damage mechanism. For instance, unlike the junctions detected in the shattered lines, those present in the other lines may not be located within gene space and might therefore be more difficult to detect using short reads. Nevertheless, our analysis using short reads was successful at identifying enough novel junctions to confirm the randomly reorganized state of the shattered chromosomes.

In conclusion, our study demonstrated that chromoanagenesis can be induced in plants by ionizing radiation of pollen, indicating that extreme chromosomal rearrangements can be more widespread, and more tolerated than expected. Notably, natural mechanisms can also produce dsDNA breaks in pollen [108,141]. This type of cataclysmic outcomes is thus possible in a natural setting and can contribute evolutionary innovations, similarly to chromosomal inversions [152]. They may also mediate gene amplification [153], which has been detected in glyphosate-resistant weeds [154]. Because poplar is vegetatively propagated, we were able to produce several clones from each chromoanagenetic line and maintain some of these extreme chromosomal rearrangements in the field for at least five years. Finally, our results show that the observed chromosomal rearrangements directly affected the sequence of multiple genes and, in some cases, have the potential to produce new chimeric proteins. While most of these random events will probably result in non- or dys-functional proteins, it is an interesting avenue for the creation of new gene functions.

Materials and Methods

Genomic sequencing and dosage variation analysis

Genomic DNA was extracted from leaf samples and prepared for deep-sequencing using Illumina technology, as previously described [60]. Sequencing reads (150 PE) were demultiplexed into individual libraries based on their barcodes, using a custom Python script (http://comailab.genomecenter.ucdavis.edu/index.php/Barcoded_data_preparation_tools), as described in previous studies [60]. Next, reads were aligned to the poplar reference *P. trichocarpa* v3.0 [72], using a custom Python script based on mapping using BWA [155] (http://comailab.genomecenter.ucdavis.edu/index.php/Bwa-doall). This generated an alignment file (sam file) for each line, which was used for further analysis.

To detect dosage variation, we calculated relative read coverage values across the genome for each line, as described previously [156]. Specifically, the genome was divided into a series of non-overlapping consecutive bins of 100kb or 10kb, depending on sequence coverage (see results). Next, for each bin, relative read coverage was calculated by taking the fraction of aligned reads in a particular bin for that line, and dividing it by the mean fraction of reads aligning to the same bin in all lines, and multiplying by 2, the background ploidy of poplar. A custom Python script was used to achieve these calculations (http://comailab.genomecenter.ucdavis.edu/index.php/Bin-by-sam). The relative coverage values obtained were then plotted according to the corresponded genomic region of their belonging bins. Values around 2 indicate the expected two copies, while values closer to 3 and 1 suggest the presence of insertions, or deletions, respectively.

Detection of novel genomic junctions

To detect novel genomic junctions, we first searched for indels boundaries, which represented the potential breakpoints of reorganized genomic fragments. Based on the dosage variation plots obtained using 10kb bins, we recorded potential junctions using the following criteria: bins where relative read coverage decreased or increased by >0.7 compared to their adjacent forward bin, and instances where this trend was true for at least three consecutive bins. Additionally, potential breakpoints were only retained if they were unique to a single line. These potential breakpoints became the most likely locations for forming novel DNA junctions. To characterize novel junctions in more detail, we next searched for reads mapping to two distant genomic locations, and therefore crossed the targeted junctions. A custom Python script (https://github.com/guoweier/Poplar_Chromoanagenesis) was used. Specifically, the script divided the genome into non-overlapping consecutive 10kb bins and, for each combination of two non-consecutive bins that were at least 2,000 bp apart, the number of reads mapping to both bins was recorded for each line. Numbers were then compared between lines to identify pairs of bins with high coverage in a single line compared to the others, suggesting the presence of a novel junction. In order to set a minimum threshold of coverage to eliminate false positives, we needed to calculate the expected average coverage over each junction. To do so, we created artificial non-overlapping 5 kb bins throughout the genome, considered the boundary between two consecutive bins as pseudo-junctions and recorded the average reads coverage at these pseudo-junctions. These values were then divided by 2, to account for the fact that these pseudo-junctions are expected to be present in two copies in the diploid poplar genome, while the indels and other novel junctions are expected to only affect one copy of the genome. We used these line-specific thresholds as minimum coverage thresholds for the identification of potential novel junctions. Second, to ensure that junctions were specific to a single line, we discarded bin-pairs that were positive in more than one line. Specifically, a potential in-pair was only retained if none of the other lines exhibited reads that mapped to those two bins.

Novel junction validation

To assemble potential novel junctions, we searched the alignment file (sam file) of each line and extracted the cross-junction reads identified at the selected bins, using a custom Python script (https://github.com/guoweier/Poplar_Chromoanagenesis). Next, the PRICE genome assembler was used to assemble the cross-junction reads into contigs [157]. The assembly parameters and input data can be found in our github repository (https://github.com/guoweier/Poplar_Chromoanagenesis). To confirm the junction genomic composition, we aligned the output contigs to the *P. trichocarpa* genome by using blast+ package [158] by using a custom bash script (https://github.com/guoweier/Poplar_Chromoanagenesis). When the two ends of the contig aligned to the expected regions, we considered that the novel junction was confirmed *in silico*.

To validate these potential junctions *in vitro*, PCR primers were designed using Primer3 [155] (Table S3.1). PCR were run using the GoTaq Green Mastermix (Promega Corporation, Madison, WI) with 1ng sample gDNA. The obtained PCR products were purified using gel extraction (QIAquick Gel Extraction Kit, Qiagen) and sent for Sanger sequencing.

SNP frequency analysis

We used parental SNP allelic percentage to identify the parental origin of the lesions. Single nucleotide polymorphism (SNP) between *P. deltoides* (female) and *P. nigra* (male) were identified previously [60]. We genotyped each line as described previously [60]. In short, to calculate the percentage of *P. nigra* and *P. deltoides* alleles at each position, we created an mpileup file containing every base allele and coverage for all examined lines, using a custom Python package based on Samtools [159] and built-in mpileup function (http://comailab.genomecenter.ucdavis.edu/index.php/Mpileup). The mpileup file was then simplified by converting a parsed-mpileup file, using the custom Python package described above. Next, the parsed-mpileup file was used to search for the preselected SNPs position. Finally, to obtain robust allele percentages, SNP allele calls were pooled within consecutive bins, and the percentage of *P. nigra* parental alleles were calculated for each bin. According to this approach, a diploid chromosome exhibited 50% *P. nigra* alleles. A deletion on one chromosome is expected to exhibit 0% or 100% *P. nigra* alleles, depending on which parental chromosome was lost. An increase of copy number states is expected to exhibit allelic ratio bias between two parents, with 1:2 represented DNA fragment duplication, 1:3 represented DNA fragment triplication, and so on.

Genome restructuring analysis

To reconstruct each mutant genome based on the identified validated junctions, we searched for fragments with the same breakpoints and strung them together manually, with the expectation that each breakpoint should be involved in two junctions, one on each side of the breakpoint. With this logic, we manually looked for paired fragment end locations among the junctions, and arranged them into longer pieces. We then built the rearranged chromosomes, while taking junction orientation and fragments copy number into account.

Enrichment ratio analysis

The poplar genome annotation file, including the genomic positions of gene and repeatmasked (GFF-Version3.0) was downloaded from Phytozome (http://phytozome.jgi.doe.gov/pz/portal.html). Next, we used a custom python script to calculate genes/repeats density around each of the novel breakpoints (https://github.com/guoweier/Poplar_Chromoanagenesis). Specifically, each potential breakpoint was set as the center of a 10kb or 100kb window, and the nucleotide number of typical genomic features within these windows was recorded. Next, to provide a random set of junctions, we used the previously constructed pseudo-junction pool, and randomly selected 1,000 of these pseudo breaks for genomic feature density calculation. For each line, this type of random pseudo-break datasets were established 1,000 times for every examined genomic feature. Enrichment ratios were calculated by taking the means of genomic feature density at real breakpoints, divided by the means of the corresponded features density at random pseudo breakpoints datasets. Significance was assessed by comparing the density of real breakpoints and 1,000 randomized datasets using one sample t-test.

Acknowledgements

We acknowledge Meric Lieberman for assistance on bioinformatics.

Tables and Figures

Table 3.1. Summary of DNA junction validation frequencies.

Type	Lines	In silico Assembled Junction	PCR validated Junction	Validation Frequency	Assembled Junction in Group	Validated Junction in Group	Validation Frequency in Group
Shattering Group	POP33_31	16	12	0.75	33	26	0.788
	POP30_88	17	14	0.824			
Lesion Group	POP25_72	5	0	0	22	0	0
	POP26_54	1	0	0			
	POP27_88	14	0	0			
	POP28_86	2	0	0			
No-lesion Group	POP27_32	10	0	0	11	0	0
	POP27_77	1	0	0			
	POP31_79	0	0	0			

The validation frequencies represent the percentage of PCR-validated junctions out of the total number of in silico predicted junctions.

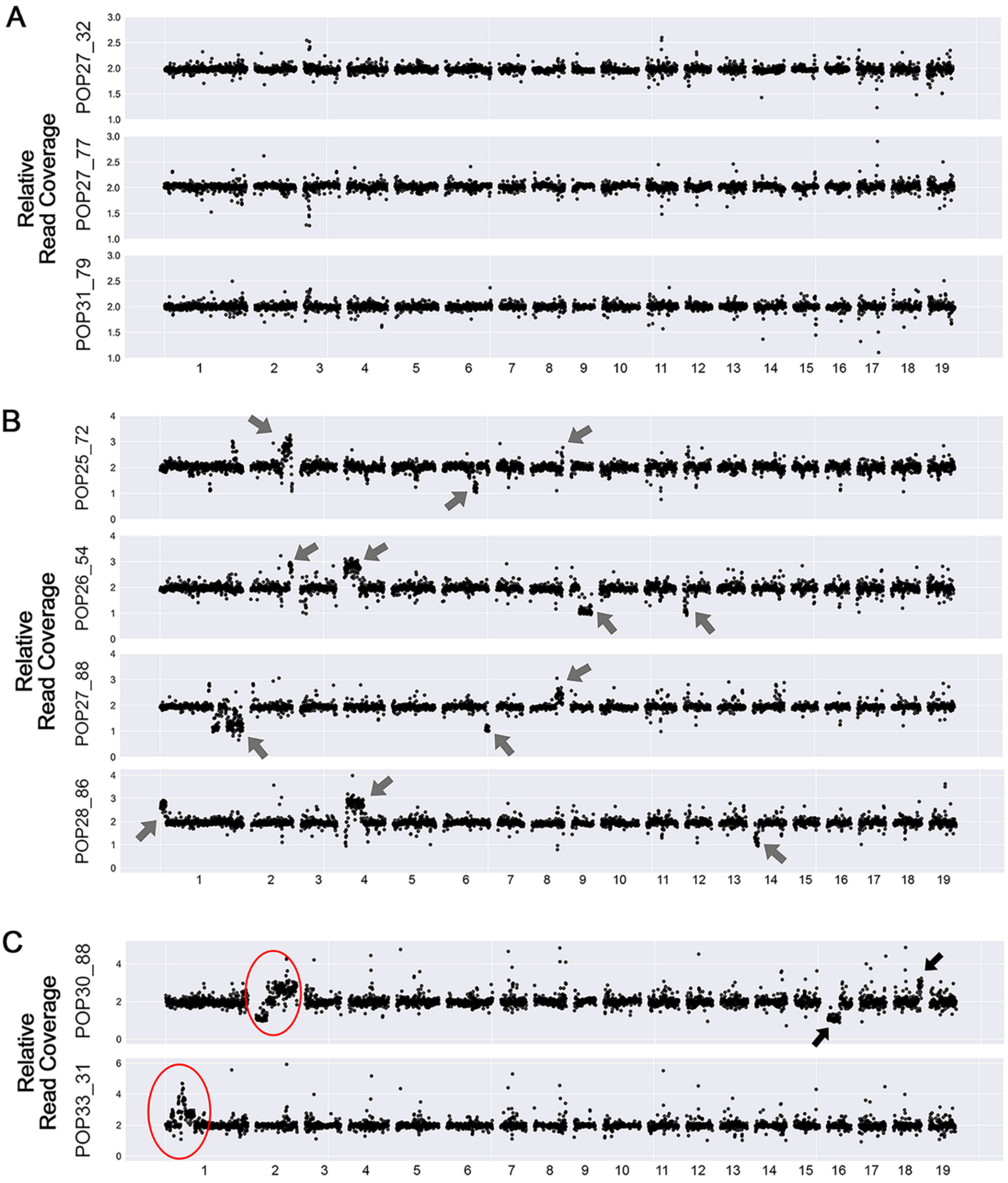


Figure 3.1. See next page for caption.

Figure 3.1. Dosage variation detection. Dosage variation was detected by displaying relative read coverage. Each data point represents the mean read coverage in non-overlapping 100kb bins, standardized to the mean read depth across all 9 lines. The expected value for a diploid line is a relative read coverage of 2. Values around 1 suggest deletions and values around 3 suggest insertions. (A) Dosage plots for 3 lines exhibiting no obvious instances of dosage variation. (B) Dosage plots for 4 lines containing a small number of indels. The arrows point to the randomly distributed indels identified. (C) Dosage plots for the 2 lines exhibiting shattering patterns. The red circles represent the regions displaying highly clustered copy number variation.

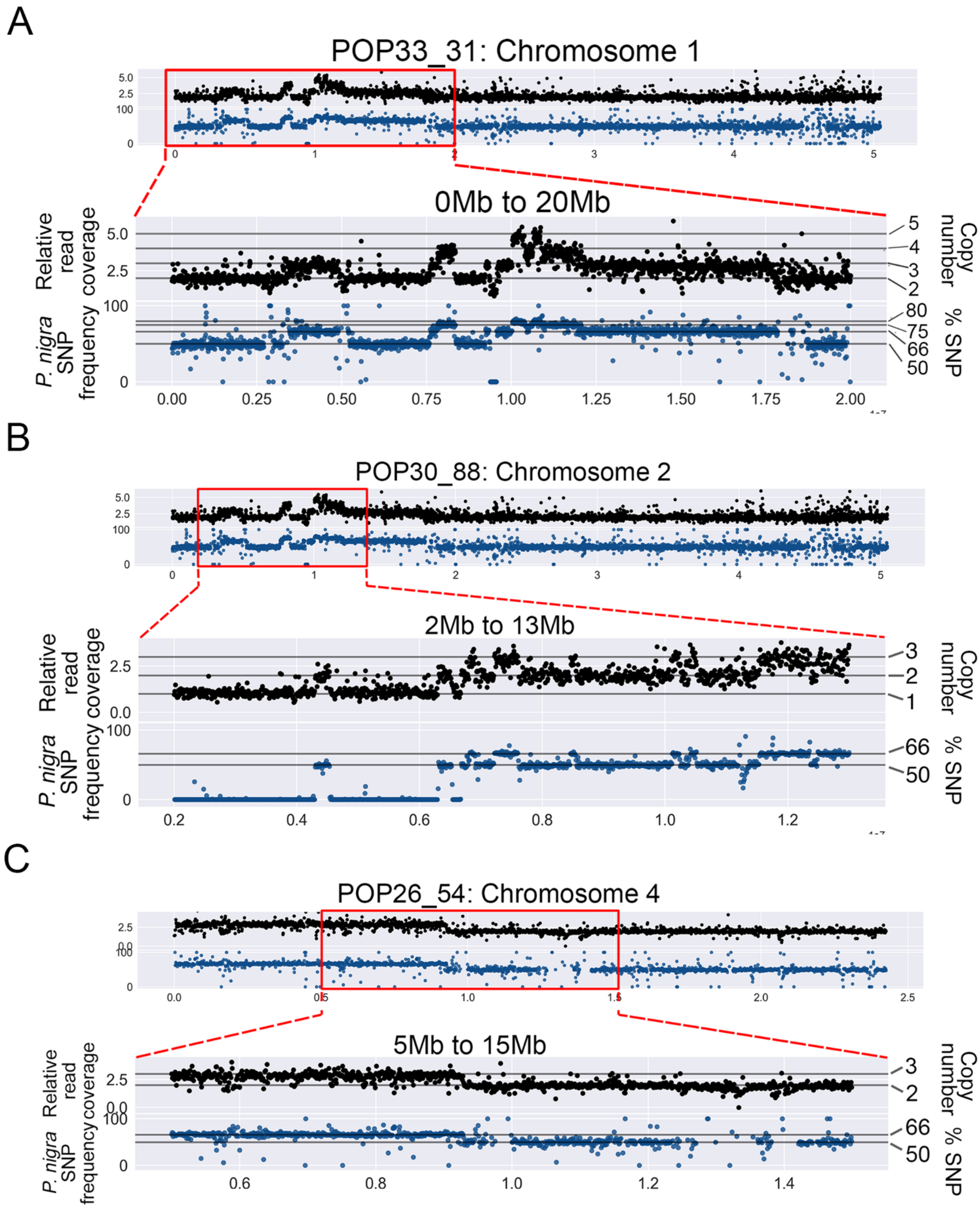


Figure 3.2. See next page for caption.

Figure 3.2. Association of dosage variation patterns with SNP frequency. To obtain a detailed view of the shattered regions, the genome was divided into narrower bins (10kb bins). Additionally, to confirm the origin of indels, *P. nigra* (male) SNPs frequencies were calculated for 10kb bins. Black scatterplots: each black dot represents the relative read coverage for a 10kb bin. Blue scatterplots: each blue dot represents the average *P. nigra* SNP frequency for a 10kb bin. Horizontal lines indicate the expected SNP frequency for different copy number states, as indicated on the right. (A) Chromosome 1 of POP33_3 displayed extremely clustered dosage variation within the first 20Mb, and all variation patterns were associated with *P. nigra* SNP frequency shifts. (B) Chromosome 2 of POP30_88 displayed extremely clustered dosage variation in the region between 3Mb and 13Mb, and all CNVs were associated with expected *P. nigra* SNP frequency shifts. (C) One of the large-scale lesions on the POP26_54 genome is shown, providing a comparison between larger randomly distributed indels and the observed shattering patterns in the other two lines.

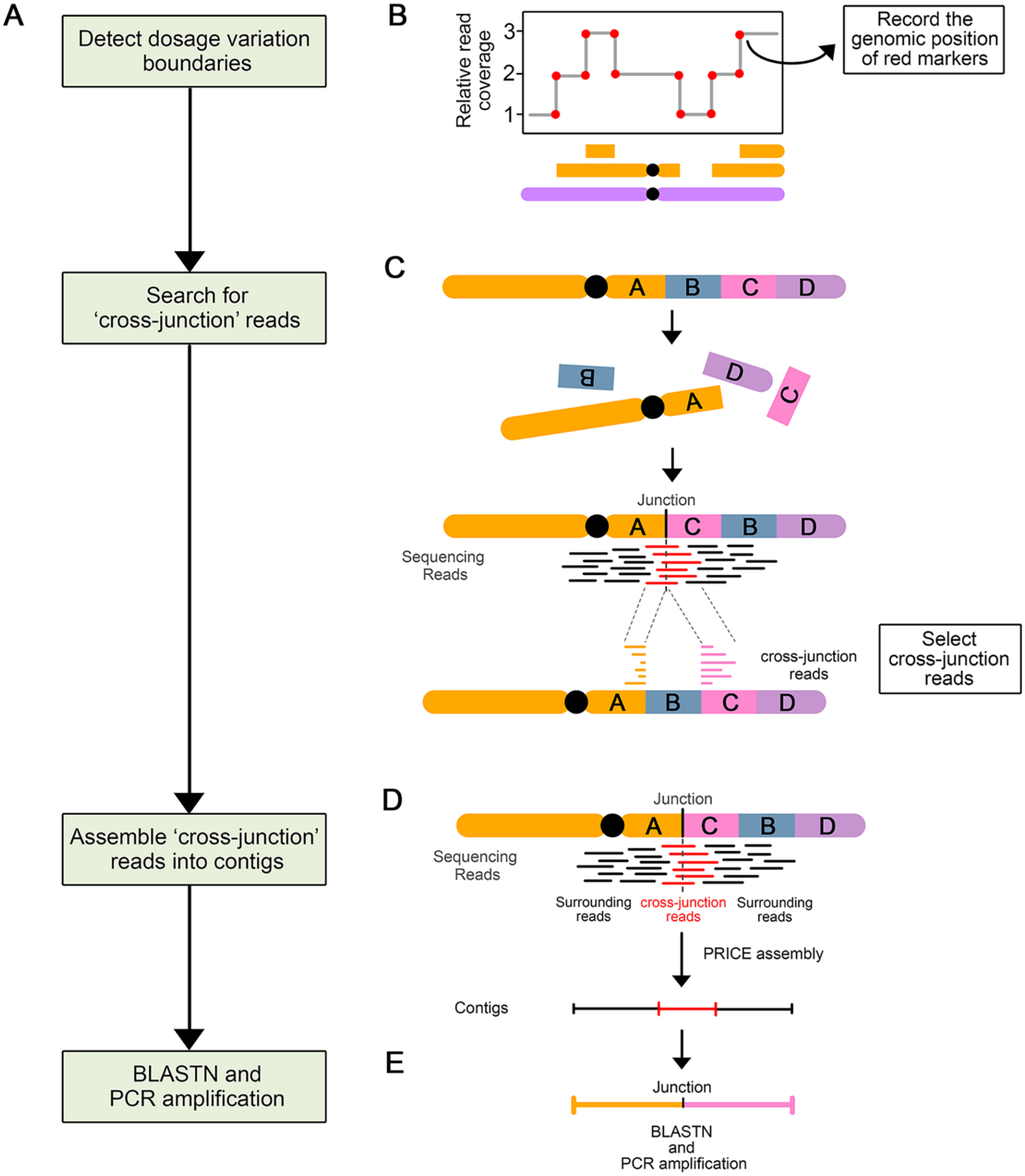

A
Detect dosage variation boundaries
Search for 'cross-junction' reads
Assemble 'cross-junction' reads into contigs
BLASTN and PCR amplification
B
Relative read coverage
1
2
3
Record the genomic position of red markers
C
A
B
C
D
Junction
Sequencing Reads
cross-junction reads
Select cross-junction reads
D
Junction
Sequencing Reads
Surrounding reads
cross-junction reads
Surrounding reads
PRICE assembly
Contigs
E
Junction
BLASTN and PCR amplification

Figure 3.3. See next page for caption.

Figure 3.3. Process of novel DNA junctions selection and validation. (A) Flow chart illustrating the steps involved in novel DNA junction detection, selection, and validation. (B-E) Diagram illustrating the approach involved in each step. (B) A schematic dosage plot showing a genomic region containing many instances of dosage variations. The red dots highlight the boundaries of every indel and constituting potential breakpoint positions. (C) Schematic diagram illustrating the origin and mapping behavior of cross-junction reads. After chromosomal rearrangement, fragment A and C joined together and formed a novel DNA junction. The sequencing reads (in red) that crossed this novel DNA junction are called cross-junction reads. These cross-junction reads map onto two different locations on the reference genome. (D) Assembly of the novel DNA junctions. Cross-junction reads are assembled into one contig using the PRICE assembler. (E) Each newly assembled scaffold is compared to the reference genome using BLASTN to: (i) find out the exact alignment positions of two breakpoints of the novel junction; (ii) confirm the uniqueness of contigs.

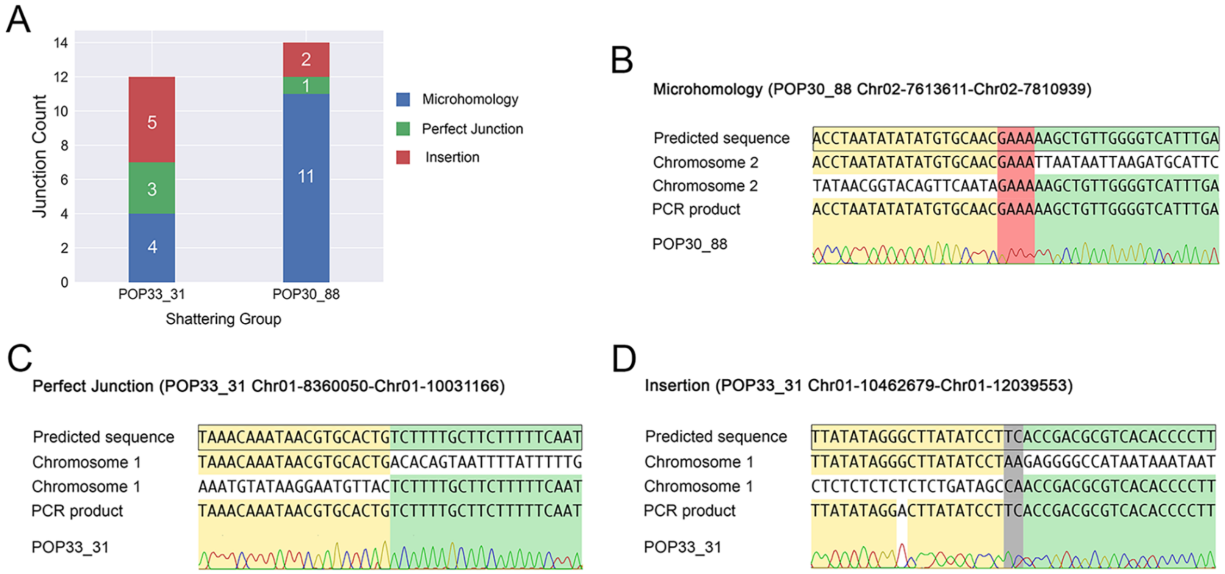


Figure 3.4. Types of novel DNA junctions. (A) Number and types of validated DNA junction identified in each line. Different colors represent the three junction types. (B) Microhomology: presence of 1-11 bp of overlap between the two reconstructed fragment ends. (C) Perfect junction: the two fragment ends are perfectly joined together, with neither overlapping bps nor inserted bps. (D) Insertion: 1-18 bp of novel nucleotide sequence is inserted between two fragment ends.

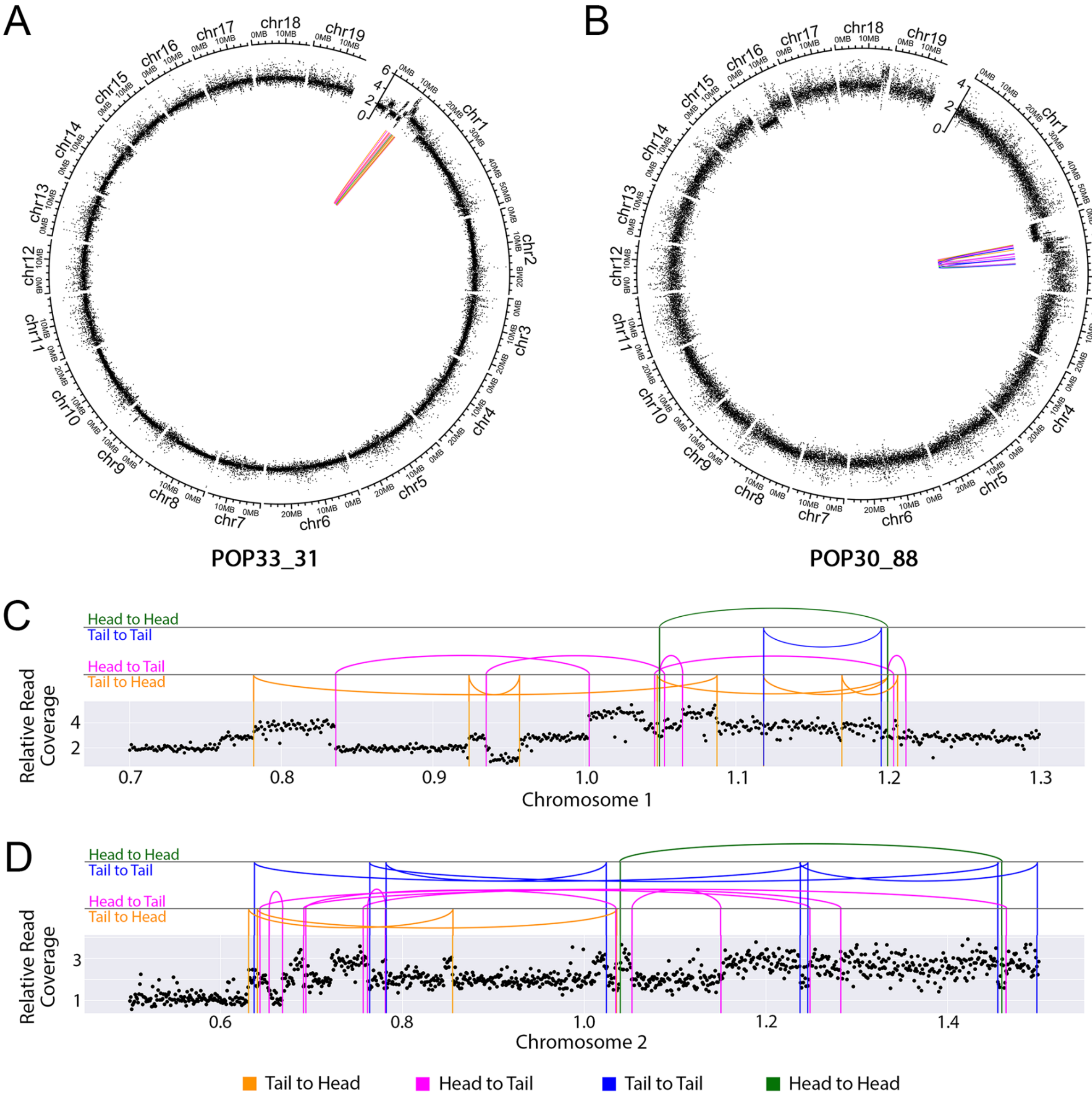


Figure 3.5. Distribution of the genomic location of the validated DNA junctions. (A-B) DNA junctions in the two shattered lines (POP33_31 (A) and POP30_88 (B)). The outermost layer displays each chromosome. The next layer displays relative reads coverage, averaged over 10kb non-overlapping bins. In the center, colored lines connect the original genomic locations of each pair of sequences found in novel DNA junctions. (C-D) Close-up view of DNA junctions distribution on the shattered regions of chromosome 1 in POP33_31 (C) and chromosome 2 in POP30_88 (D). The scatter plots show average relative read coverage per 10kb bins, and the colored vertical lines represent exact breakpoints. The arc connecting two vertical lines illustrate the novel junctions connecting vertical lines that represent the breakpoints. All panels: Magenta and orange lines represent sequences that connect in the same direction (Head to Tail in magenta and Tail to Head in orange). Blue and green lines represent sequences that connect in opposite directions (Tail to Tail in blue and Head to Head in green).

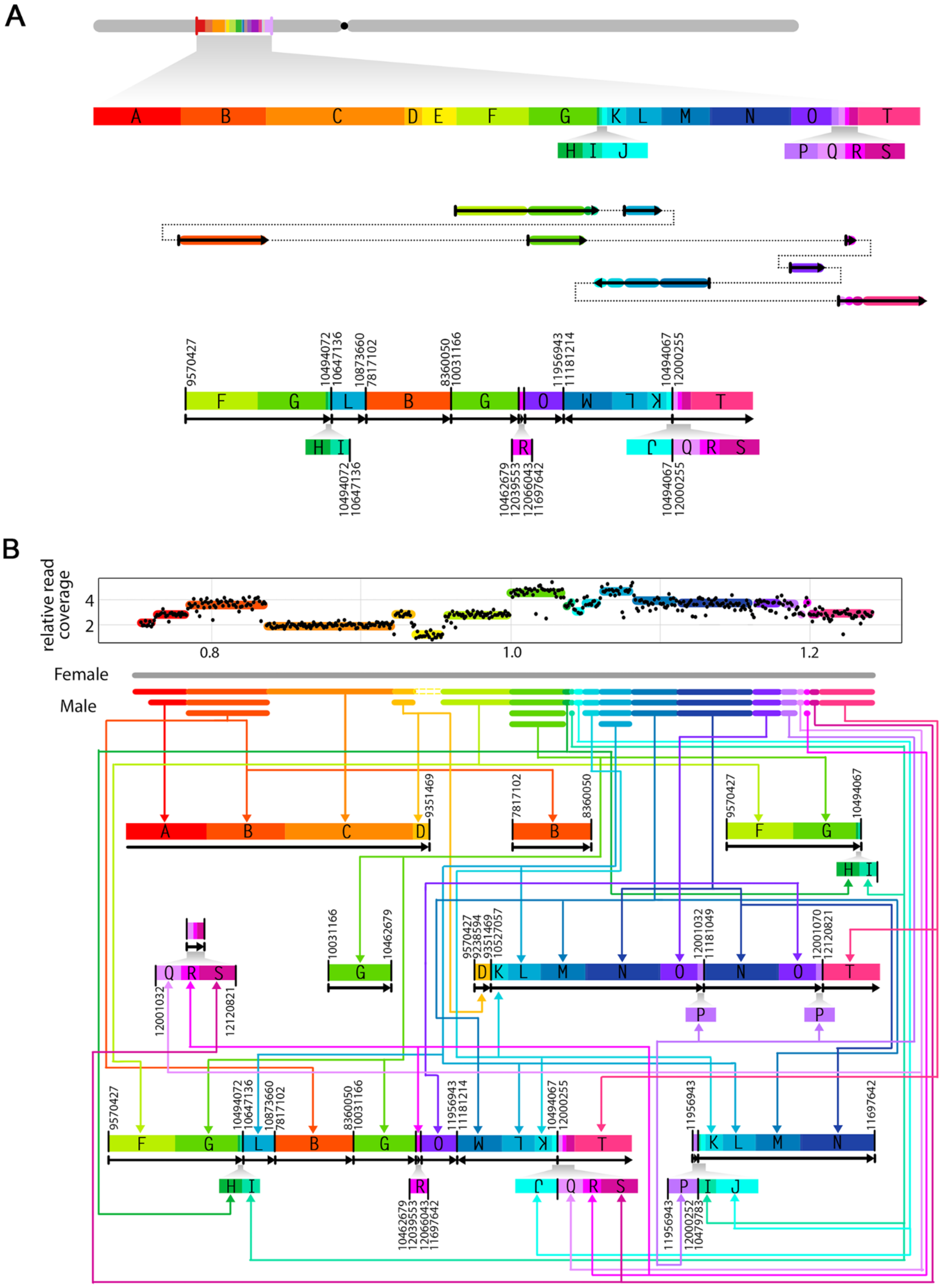

A
A B C D E F G K L M N O T
H I J
P Q R S
9570427
10494072
10647136
10873660
7817102
8360050
10031166
11956943
11181214
10494067
12000255
F G L B G O M L K T
H I
R
J Q R S
10494072
10647136
10462679
12039553
12066043
11697642
10494067
12000255
B
relative read coverage
4
2
0.8
1.0
1.2
Female
Male
A B C D
9351469
7817102
B
8360050
9570427
F G
10494067
H I
Q R S
12001032
12120821
10031166
G
10462679
9570427
9238594
9351469
10527057
D K L M N O
12001032
11181049
N O
12001070
12120821
T
P
P
9570427
F G L B G O M L K T
10494072
10647136
10873660
7817102
8360050
10031166
11956943
11181214
10494067
12000255
H I
R
10462679
12039553
12066043
11697642
J Q R S
11956943
K L M N
11697642
P I J
11956943
12000252
10479783

Figure 3.6. Unraveling the structure of the shattered chromosomes. Schematic diagrams illustrating the breakpoints rearrangement in one of the genome shattered lines (POP33_31). (A) The reference chromosome 1 is shown in grey and the regions engaged in rearrangement are labeled in alphabetical order. Each labeled block has a unique color and represents a genomic fragment with validated breakpoints on its flanking ends. The small blocks are enlarged below. All block sizes are proportional to genomic coordinates. In the middle is the potential junctions creating one of the rearranged fragments. Solid arrows with colors represent the corresponding blocks, and the dashed lines illustrate the order of blocks reconstruction. At the bottom is the new structure of that same fragment. Novel junctions are highlighted with bold vertical lines, and are labeled with their original genomic positions on two sides. Black arrows below blocks indicate the orientation of reconstructed fragments. Small blocks are enlarged below proportionally. Fragment duplications are linked and pointed out with the same color. (B) Summary of the fragments reconstructed based on the data obtained from line POP33_31. The dosage plot on top displays relative read coverage of the shattering region in chromosome 1. Each DNA block is labeled with the same color used in (A). A schematic representation of chromosome is shown below the dosage plot, with female (*P. deltoides,* WT) inherited chromosome colored in grey, and male (*P. nigra*, pollen irradiated) inherited chromosome illustrated in their corresponded colors and copy number states. For each DNA block, the same color arrows guide to the corresponding fragments on the reconstructed chromosome pieces.

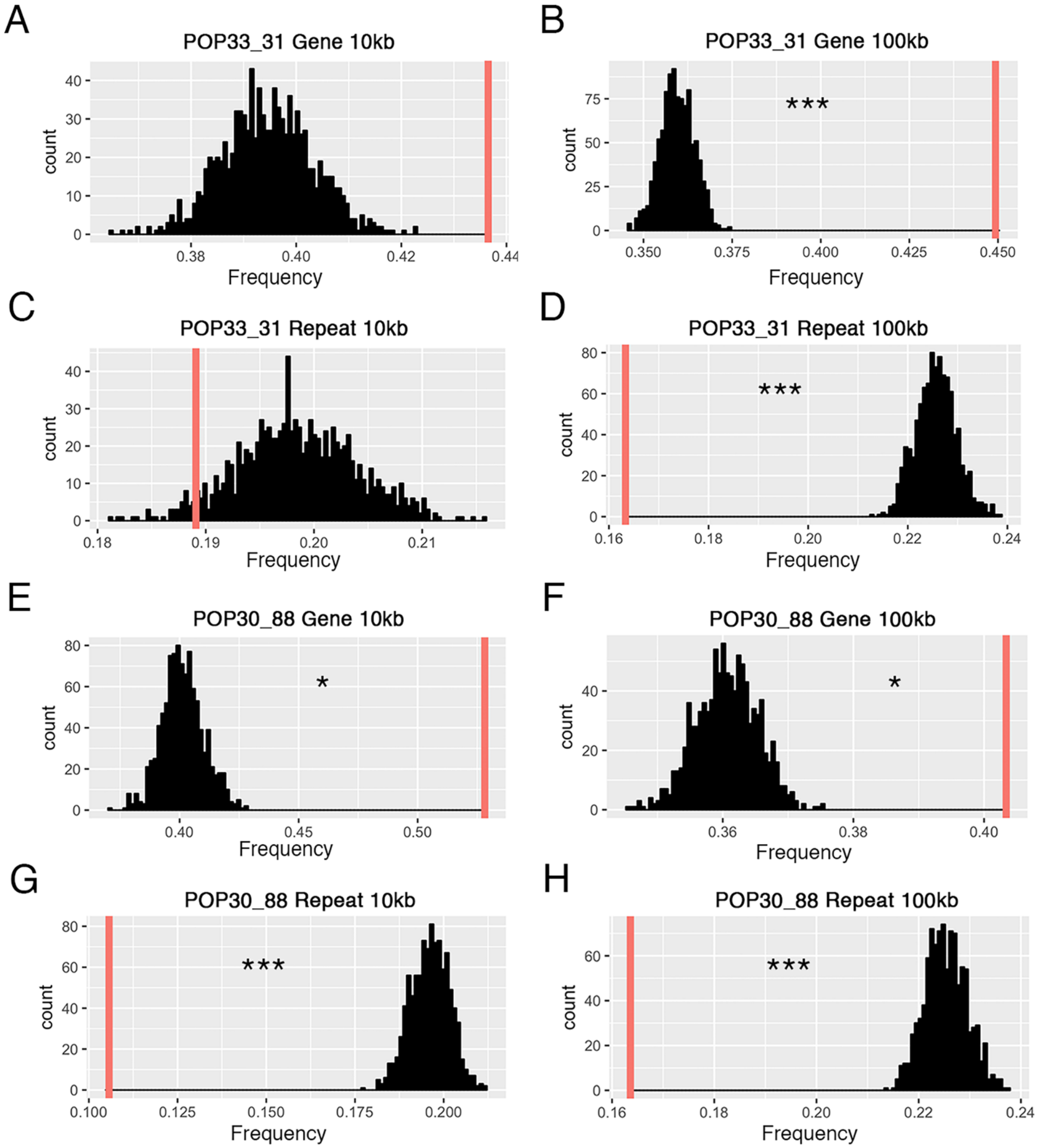


Figure 3.7. See next page for caption.

Figure 3.7. Sequence context surrounding the breakpoints of novel DNA junctions. The frequency of genes and repeated elements surrounding novel junctions is compared to the corresponding frequencies in randomly selected pseudo junctions. For each panel: 1,000 pseudo-junctions were selected at random and the mean percentage of gene or TE space in these 1,000 junctions was calculated. This process was repeated 1,000 times and the distribution of these means are represented in black. The red vertical line represents the mean of enriched frequency for the observed validated novel junctions. Breakpoints of novel junctions in POP33_31 (A-D) occur significantly in gene-rich, repeats-deficient regions under 100kb window size (p-value < 0.001), but do not show statistical significance in 10kb window size. Results were similar for POP30_88 (E-H). The observed junctions are significantly enriched with genes (p-value < 0.05), and have the lack of repeated elements (p-value < 0.001) regardless of window size.

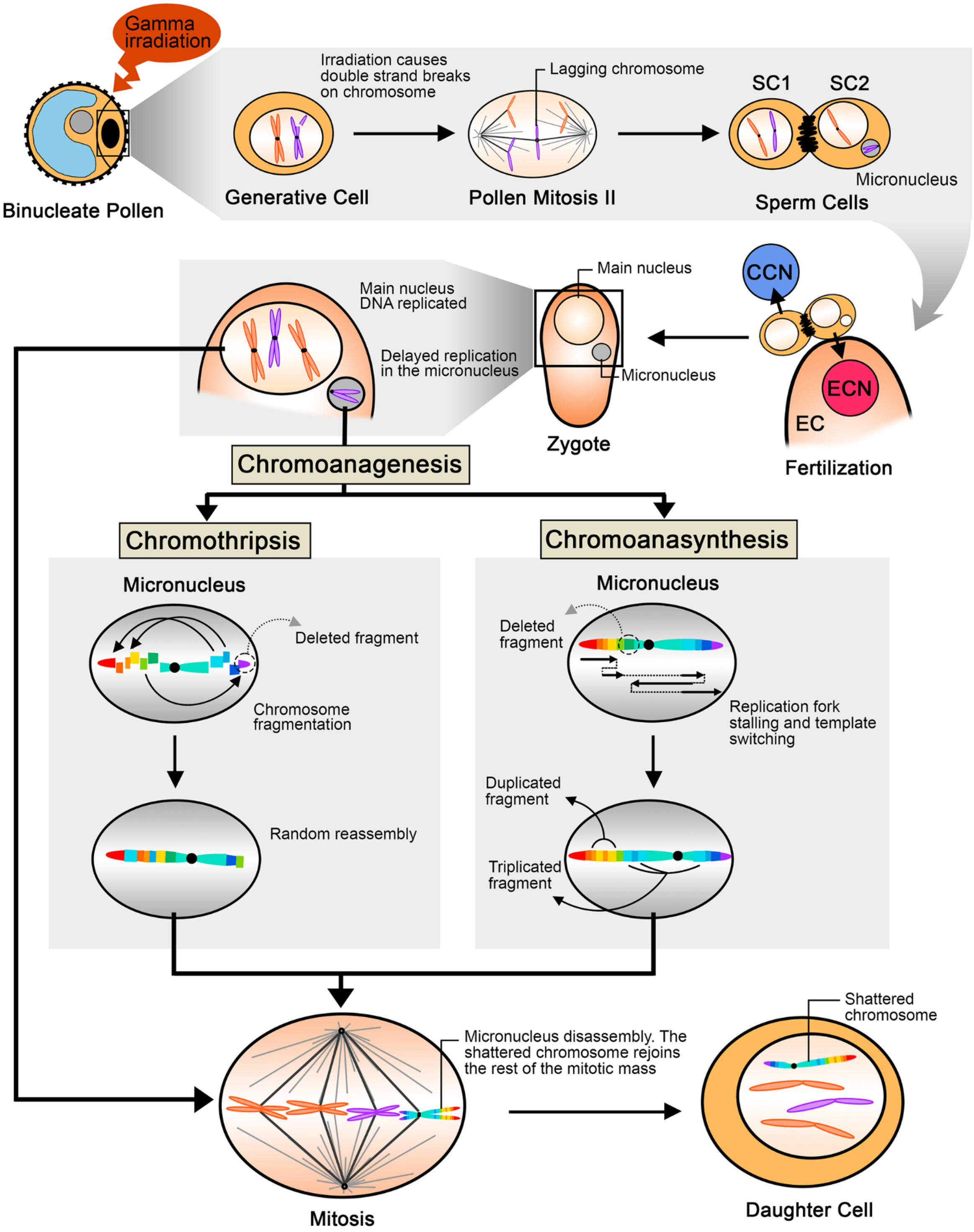


Figure 3.8. See next page for caption.

Figure 3.8. Proposed model illustrating the steps leading to chromoanagenesis following pollen irradiation. Gamma irradiation of binucleate pollen induces double stranded DNA breaks in the generative cell, and results in chromosome lagging or in bridge formation [36] during the second pollen mitosis. The lagging chromosome is excluded from the main nucleus and forms a micronucleus. The sperm cell carrying the micronucleus undergoes karyogamy with the egg cell, and produces a zygote with a (2n-1) nucleus and a micronucleus containing a single paternal chromosome. DNA replication in micronuclei is delayed and leads to chromoanagenesis via two possible mechanisms, chromothripsis and chromoanasynthesis, which were both observed in our poplar lines. Chromothripsis involves fragmentation and random reassembly, while chromoanasynthesis results from replication fork stalling and template switching. The highly rearranged chromosome is eventually released from the micronucleus and reunites with the main nuclear genome during mitotic division. The shattered chromosome is thereafter retained in the main nucleus. SC: sperm cell; EC: egg cell; ECN: egg cell nucleus; CCN: central cell nucleus.

Supplementary Materials

Supplementary Figures

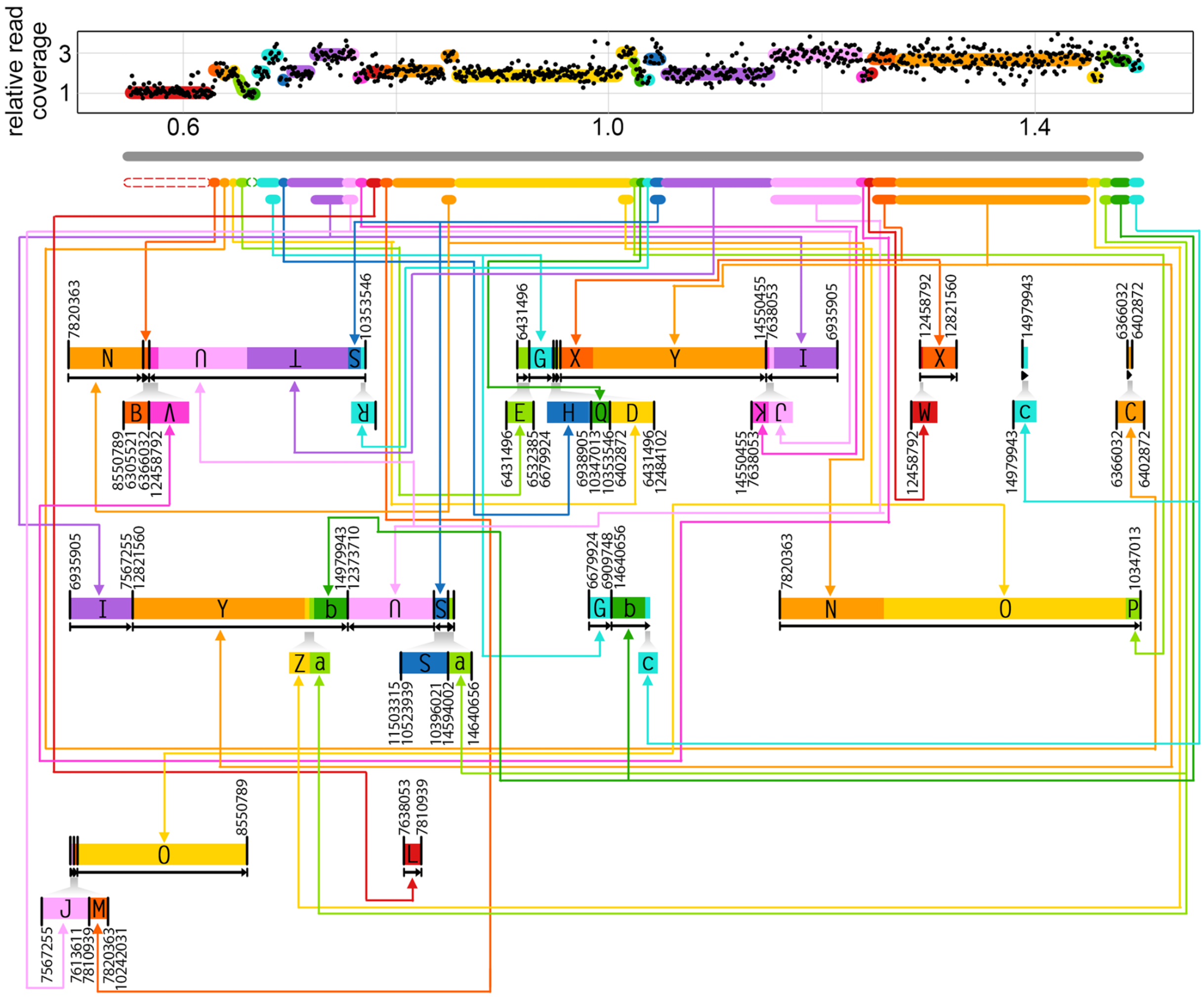


Figure S3.1. Summary of the fragments reconstructed on line POP30_88 Chromosome 2. The diagram follows the same criteria as in Figure 3.6B. The rearranged fragments were constructed based on the data of novel junction observation in POP30_88.

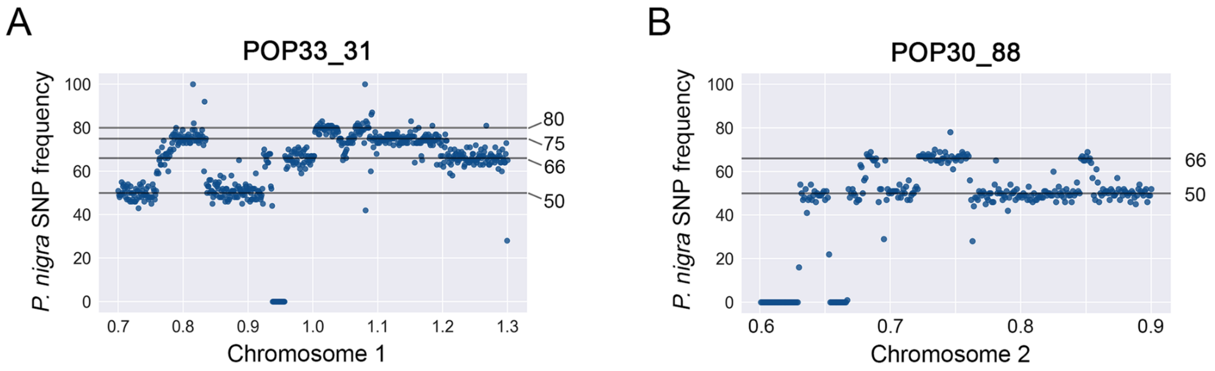


Figure S3.2. Detailed view of the *P. nigra* SNP frequency pattern in the shattered regions. The genome was divided into consecutive non-overlapping 10kb bins. Each blue dot represents the average *P. nigra* SNP frequency for a 10kb bin. Horizontal lines exhibit the expected frequency levels for different copy number states, with their numbers labeled on the right.

Supplementary Tables

Table S3.1. PCR primers used in novel junction validation. List of primers used for PCR amplification.

Table S3.2. Summary of indels in 2 shattered lines. Large-scale indels in two shattering lines (POP33_31=21, POP30_88=11) were identified based on dosage variation patterns and SNP frequency. The locations and copy number states of indels are indicated, as well as the parental genotype they originated from. D: *P. deltoides*; N: *P. nigra*.

Table S3.3. Summary of all validated novel DNA junctions. List of validated novel DNA junctions in the two shattering lines (POP33_31=12, POP30_88=14). Each junction is indicated with its junction type, two breakpoints positions, orientation, and its correlation with CNV edges.

Table S3.4. DNA context at the breakpoints. List of all breakpoints identified, as well as information about the affected genes when the breakpoints occurred within a gene.

Table S3.5. Summary of the possible gene fusion events at novel DNA junctions. List of junctions containing gene to gene fusion within the two shattering lines. Instances where two genes are fused in the same direction are labelled in green, indicating that these fusions might form novel gene products.

Table S3.6. DNA context surrounding the novel junctions. The novel DNA junctions identified in the two lines (POP33_31=12, POP30_88=14) exhibiting clustered patterns are preferentially located in regions that are rich in gene sequences and poor in repeated sequences, compared to the rest of the genome. Enrichment ratio represents the comparison between the means of genomic feature density at real breakpoints and the means of a similar set of randomly selected features. Ratio >1 indicates validated breakpoints that have a higher density of features than the genome average, while ratio <1 indicates the lower density of validated breakpoints compared to the genome. Genes are enriched near breakpoints of both lines with 100kb-window size (POP33_31 $p<0.001$; POP30_88 $p<0.05$). The lack of repeated elements surrounding breakpoints are observed in both lines as well (POP33_31 $p<0.001$; POP30_88 $p<0.001$).

All Supplementary Tables are available at https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1009735#sec015

Chapter 4

Chromoanagenesis in the *asy1* meiotic mutant of *Arabidopsis*

[Published in: G3 Genes| Genomes| Genetics]

Weier Guo[1], Luca Comai[1] and Isabelle M. Henry[1]*

[1]Genome Center and Dept. Plant Biology, University of California Davis, 95616

*Corresponding author: imhenry@ucdavis.edu

Abstract

Chromoanagenesis is a catastrophic event that involves localized chromosomal shattering and reorganization. In this study, we report a case of chromoanagenesis resulting from defective meiosis in the MEIOTIC ASYNAPTIC MUTANT 1 (asy1) background in *Arabidopsis thaliana*. We provide a detailed characterization of the genomic structure of this individual with a severely shattered segment of chromosome 1. We identified 260 novel DNA junctions in the affected region, most of which affect gene sequence on one or both sides of the junction. Our results confirm that asy1-related defective meiosis is a potential trigger for chromoanagenesis. This is the first example of chromoanagenesis associated with female meiosis and indicates the potential for genome evolution during oogenesis.

Short summary

Chromoanagenesis is a complex and catastrophic event that results in severely restructured chromosomes. It has been identified in cancer cells and in some plant samples, after specific triggering events. Here, we identified this kind of genome restructuring in a mutant that exhibits defective meiosis in the model plant system *Arabidopsis thaliana*.

Background

Complex chromosomal rearrangements (CCRs) refer to genomic structure variation that involve at least three double strand DNA breaks among two or more chromosomes [77]. These changes can cause the truncation, relocation, or copy number variation of multiple genes or gene regulatory elements, which can subsequently lead to dramatic phenotypic changes [160]. Chromoanagenesis, caused by a single catastrophic genome restructuring event, and diagnosed by the presence of tens to hundreds of copy number variations (CNVs) on a single chromosome, has been identified in many systems in the last decade [30,32–35,38,161,162]. It can be associated with multiple types of human cancer [35,163], or with transgenic modifications used for plant genetic engineering [30,38,41,44]. The origin, mechanism and potential effects of chromoanagenesis are just starting to be deciphered. Chromothripsis is one type of chromoanagenesis, characterized by the pulverization of a single chromosome and its random reassembly with limited copy number changes [32,80,82]. Chromothripsis has been used to describe many extreme chromosome rearrangements in various systems [37,44,124]. Besides chromothripsis, the two other types of processes included in chromoanagenesis - chromoanasynthesis and chromoplexy - can produce rearranged chromosomes as well, but exhibit different features and result from different mechanisms [33,34]. Here, in the absence of mechanistic information, we use the broader term chromoanagenesis to describe the chromosome restructuring patterns observed in our study.

MEIOTIC ASYNAPTIC MUTANT 1 (ASY1) is the *Arabidopsis* homolog of the yeast chromosome axis component HOP1. ASY1 plays an important role in meiotic recombination by regulating crossover assurance and interference [109,110,164]. First observed in transgenic *Arabidopsis* mutants exhibiting reduced synapsis [109,165], the presence of aneuploidy in the progeny of ASY1 mutants suggests that the ASY1 mutation can also result in chromosome mis-segregation [166,167].

Here, we report a case of chromoanagenesis resulting from defective meiosis in the *asy1* mutant background in *Arabidopsis thaliana*. Specifically, a homozygous *asy1* mutant was crossed as a female to a wild-type male, and aneuploids were observed in the progeny [111]. Detailed characterization of the genome of one of these aneuploid individuals detected a severely shattered segment of chromosome 1, which was reminiscent of the consequences of chromoanagenesis. Our analyses identified 260 potential novel DNA junctions in this region, suggesting that defective *asy1* can trigger chromoanagenesis.

Results and Discussion

A recent study [111] demonstrated that the genome of one offspring from a cross between a Col-0/Ler-1 hybrid *asy1* mutants (asy1$^{Col\text{-}0}$ x asy1$^{Ler\text{-}1}$, female) and a wild-type Col-0 (male), carries drastic genomic rearrangement. These rearrangements resemble the consequence of chromoanagenesis. Among the population of 176 individuals, a single line exhibited multiple CNVs, on chromosome 1, all clustered within the first half of the chromosome (from 1 to 16.1Mb) (Fig. 4.1A-C).

To confirm the occurrence of extreme chromosomal rearrangements in this individual, we searched for novel DNA junctions expected at the sites of chromosomal fragments reassembly. Specifically, we searched for Illumina sequencing reads that mapped to two distant genomic locations (>2,000bp), indicating that two regions expected to be distant from each other in the reference genome are next to each other in the rearranged chromosome. We also expect that these novel DNA junctions are unique to the genome of this particular individual, and not present in its siblings. Based on these criteria, we identified 260 novel DNA junctions (Fig. 4.1D, E, File S4.2). For 95.7% (249 out of 260) of these junctions, both breakpoints fall within the shattered region on chromosome 1 (Fig. 4.1E). The breakpoints of the remaining 11 junctions are both located elsewhere in the genome. This frequency of one breakpoint every 32 kb across the shattered region is much higher than previously observed following chromoanagenesis in other plant systems. Specifically, the frequency of breakpoints was 1 / 400 kb in chromoanagenetic individuals that originated from haploid induction crosses in *A. thaliana* [30], and 1 breakpoint per 250 kb for the chromoanagenesis events observed in the progeny of gamma irradiated poplar pollen grains [38].

Next, we attempted to reconstruct the structure of the novel chromosome based on the position and orientation of these breakpoints. In total, we were able to reconstruct 91 fragments from these 260 novel DNA junctions identified. The longest segment involved 13 novel DNA junctions (Fig. 4.2 and File S4.3). The other reconstructed fragments are shorter, presumably because we did not identify all junctions in this sample, resulting in broken pieces in our reconstruction. Junctions that occurred within repeated regions for example, are more likely to have been missed due to poor mapping specificity. Nevertheless, these results are consistent with the hypothesis that the shattered pieces reassembled randomly, in terms of orientation and order, resulting in a completely reorganized chromosome. This is consistent with the characteristics of chromothripsis [82], and what we observed previously in poplar [38].

To characterize the properties of these novel DNA junctions, we investigated the DNA sequence context among breakpoint loci. Two window sizes, 1kb and 10kb, were used to calculate the enrichment ratio of gene space around the breakpoint loci. Statistical analysis suggested that breakpoint loci were significantly associated with gene-rich regions for both 1kb and 10kb bins (p-value < 0.001) (Table 4.1). These results are consistent with previously documented chromoanagenesis events in plants, which also exhibited higher than expected frequency of breakpoints occurrence in genic regions [30,38]. In addition to gene density, we also characterized the potential enrichment of other genomic features, including chromatin states, transposable elements, or replication origins. The results suggest that breakpoint loci occurred more often in accessible chromatin regions, such as near transcription start sites, while they were significantly depleted in heterochromatin regions (Table 4.1) [168].

Using in silico assembly of the junctions (see Methods), we were able to identify the exact location of the junctions and their exact sequences. Based on the specific sequence at these junctions, we determined that 50.4% (131 out of 260) of these junctions involved the joining of fragments in inverted configuration, while the other half involved two fragments coming together in the same orientation (File S4.2). Junctions could be divided into three junction types : i) microhomology is defined by the presence of an identical sequence (1-29bp) on both sides of the junction, and resulting in a single repeat of the micro-homologous fragment at the resulting novel junction, instead of the expected two copies if the two fragments had come together directly; ii) perfect junctions involved the joining of the two ends with no modification at all and, iii) insertions involved the addition of a few base pairs (1-80bp) between the two original DNA sequences. All three types were observed in this shattered line at the following rates: microhomology (63.8%), perfect joining (11.2%), and insertion (25%). Analysis of these precise locations demonstrated that 69.4% (361 out of 520) breakpoints occurred within a gene sequence. For 49.2% (128 out of 260) of the novel junctions, both breakpoints are located within coding regions (File S4.4). This is expected to result in the loss of function of several genes and possibly in a few novel gene functions, in cases where the junction joined two different coding regions in phase.

Notably, the vast majority of the previously identified chromoanagenesis events in animals [36,98,99,169,170], and all of the characterized events in plants [30,38,44], have been associated with mitosis, usually during early embryo development [30,171], or male gametogenesis [38]. Only a few studies have reported that chromoanagenesis can be correlated with meiotic divisions. Specifically, in human germ cells, chromoanagenesis has been demonstrated to occur during the meiotic divisions of spermatogenic cells and spermiogenesis [169,172]. Extreme chromosomal rearrangements are also expected to occur following defects in female meiosis [173], but no case has been observed so far.

In this study, chromoanagenesis was detected in the offspring of an *asy1* homozygous mutant in a Ler-1/Col-0 background. Specifically, the *asy1* allele in the Ler-1 background was caused by a G nucleotide insertion caused by CRISPR-Cas9. The *asy1* allele in Col-0 corresponds to SALK_046272 T-DNA insertion line, near the ASY1 gene, which is located at 25,240,000 Mb on the lower arm of Chr1. Since both events occurred on chromosome 1, we cannot fully exclude the possibility that the T-DNA insertion played a role in the observed genomic instability. Nevertheless, it seems unlikely based on the fact that the shattering and the T-DNA insertion are located on two different arms of chromosome one, and the observation that none of the other progeny of this T-DNA insertion line exhibited chromoanagenesis.

The more likely explanation is that the ASY-1 mutation, which is known to affect crossover assurance and interference, resulted in altered recombination patterns and unbalanced chromosome segregation during meiosis [110]. Cytological evidence has shown the presence of unequal chromosome segregation during microsporogenesis in *asy1* mutants [165,166,174,175]. Cytological analysis of female sporogenesis in these plants also documented abnormal chromosome pairing and uneven chromosome segregation [176,177].

To investigate at which stage of meiosis this missegregation occurred, we performed a haplotype analysis to examine the origin of the shattered chromosome. This analysis indicates that the percentage of Ler-1 alleles oscillates between 33% and 50% within the shattered region. In contrast, it remains stably around 50% for the rest of chromosome 1 (Fig. 4.3). Furthermore, the presence of many regions with 33% Ler-1 alleles indicates trisomy of the upper arm of chromosome 1. This data is consistent with the following scenario: two copies are intact, one Col-0 haplotype from the Col-0 parent, the other is a Ler-1 haplotype from the hybrid mutant parent. The frequency of Ler-1 alleles over the shattered region goes back and forth between 33% and 50%. This indicates that the shattered chromosome carries the Col-0 haplotype, adding a copy of the Col-0 haplotype when it is present (33% Ler-1). The regions that are missing from the shattered chromosome remain at 50% Ler-1 from the two intact chromosomes. The fact that allelic frequencies oscillate between those two percentages (50% and 33%) throughout the shattered region suggests that the shattered chromosome was not the product of a recombination even in the mutant hybrid prior to mis-segregation. Finally, based on the expected percentage of parental alleles in various cases (Fig. 4.4), the observed percentages suggest that the shattered chromosome 1 originated from mis-segregation during meiosis I of megasporogenesis.

Micronuclei have been observed during male sporogenesis in *asy1* mutants with Col-0/Ler-1 background [111]. Micronuclei have also been observed in haploid induction crosses, which also have generated chromoanagenetic events [30]. It is thus possible that a similar suite of events are at play here. Specifically, chromosome mis-segregation occurred during female sporogenesis, resulting in the formation of a micronucleus carrying a chromosome laggard. Fragmentation and reorganization of the chromosome entrapped within the micronucleus subsequently created the shattered chromosome (Fig. 4.5). Together, our results provide the first example of chromoanagenesis triggered during female meiosis.

The fact that this *Arabidopsis* line produced a viable plant despite carrying an extremely shattered chromosome may be explained by the fact that the rearranged segments are present in three copies in a diploid background. Therefore, any negative functional effect of the shattering and rearrangements, including to protein sequences or reduced gene expression, are buffered by the presence of two intact copies of chromosome 1. The same situation applied to the *Arabidopsis* lines that underwent chromoanagenesis from haploid induction crosses: the rearranged chromosome or chromosomal segments were present as an extra copy of a trisomy [30].

Our result further suggests that *A. thaliana* is able to tolerate this extreme restructuring process during meiosis and survive through fertilization and embryogenesis. Unfortauntely, seed were not collected from this particular line so it is unclear if the shattered chromosome was transmissible sexually. Sexual transmission of a similar shattered chromosome was reported in previous studies in *Arabidopsis* though [30].

Chromoanagenesis-like rearrangements have been previously reported as potentially associated with a role in shaping the genome of camelina and the genus Cucumis [124,125]. Plant species such as *A. thaliana,* with powerful genetic resources could become a valuable system for investigating the mechanisms underlying extreme chromosomal rearrangement, and eventually unraveling the pathways leading to chromoanagenesis, and their potential role in plant genome evolution.

Conclusions

We describe a case of chromoanagenesis that is remarkable by the high frequency of new DNA junctions produced, and because it results from asynapsis during female meiosis. The event demonstrates the potential for karyotypic innovation in connection to oogenesis.

Material and Methods

DNA from the *Arabidopsis* line exhibiting multiple CNVs was prepared for deep sequencing as follows. The genomic DNA was extracted from the leaf tissue, and prepared for Illumina short-read sequencing as previously described [60]. Demultiplexing and quality filtering was performed using a custom Python script (https://comailab.org/data-and-method/barcoded-data-preparation-tools-documentation/). Reads were mapped to the TAIR10 reference genome using BWA [155]. The output files (.sam files) were used for the subsequent analyses. Two controls were generated by pooling the low-sequencing read data from multiple wild-type *Arabidopsis* lines generated from a similar cross (Ler-1/Col-0 x Col-0) (File S4.1) to obtain two control files of similar coverage as the target sample.

Dosage variation along non-overlapping consecutive bins spanning the entire genome was documented as previously described [30,60]. Bin coverage was normalized to the corresponding bin in a diploid control individual for normalization, by using a customized Python script (https://github.com/Comai-Lab/bin-by-sam). The expected relative read coverage of a diploid individual is expected to be close to 2, while values close to 1 and 3 represent deletion and duplication, respectively.

Novel DNA junctions were identified as described previously [38]. Specifically, we searched for sequencing reads that span two genomic locations originally located at distant positions (>2000 bp apart, or on different chromosomes), and that appear uniquely in the target *Arabidopsis* line but not either of the two control samples. A custom Python script (https://github.com/guoweier/Poplar_Chromoanagenesis) was used to identify the potential genomic locations of the two breakpoints for each novel junction. Potential false positives were discarded based on a coverage threshold calculated as previously described [38]. Next, PRICE assembly was applied to construct contigs spanning the novel DNA junctions [157]. These contigs were compared to the sequence of the *Arabidopsis* genome by BLAST [158], with the expectation that the two sides of these *in silico* confirmed novel DNA junctions should map to the expected regions specifically (no multiple mapping allowed).

To identify the origin of the shattered chromosome, single nucleotide polymorphism (SNP) between Col-0 and Ler-1 were collected as previously reported [111]. Next, we calculated the Ler-1 allelic frequency along the genome of the shattered *Arabidopsis* line, following a method developed previously [60]. Specifically, an mpileup file was created to record the allele and read coverage in each genomic position, followed with a simplification step to create a parsed-mpileup file, using a custom Python pipeline (http://comailab.genomecenter.ucdavis.edu/index.php/Mpileup) based on Samtools [159]. The parsed-mpileup file was used for calling SNPs between Col-0 and Ler-1, and data from the selected SNPs were pooled into consecutive non-overlapping bins. Ler-1 allele frequency for each bin was calculated and visualized. Since the shattered *Arabidopsis* line was produced from a cross between a Col-0/Ler-1 hybrid (female) and a WT-Col-0 (male), at least 50% Col-0 alleles were expected. The transmission of a Ler-1 chromosome from the female parent results in 50% Ler-1 alleles, while the transmission of a Col-0 chromosome results in 0% Ler-1 alleles. Detailed discussion of allele frequency in copy number variations are in Figure 4.4.

To analyze the genomic features surrounding the breakpoints, the frequency of gene space surrounding them was compared to the frequency of pseudo breakpoints randomly selected along the *Arabidopsis* genome. The annotation files of various genomic features of *Arabidopsis thaliana* (TAIR10) were acquired from the GitHub repository (https://github.com/KorfLab/FRAG_project) associated with the breakpoint analysis previously performed on aneuploid *Arabidopsis* [30]. Statistical analysis was performed as previously described [38].

Data availability

The sequences reported in this paper have been deposited in the National Center for Biotechnology Information BioProject database (BioProject ID: PRJNA723952).

Conflict of interest

The authors declare no competing interest.

Authors’ contributions

WG analyzed and interpreted the data, and drafted the manuscript. LC and IMH conceptualized the project, helped interpret the data, and reviewed and edited the manuscript. IMH managed the project. All authors read and approved the final manuscript.

Funding

This work was supported by the joint USDA DOE Feedstock Genomics Programs (DOE Grant DE-SC0005581).

Acknowledgments

We thank Dr Arp Schnittger for providing DNA from the chromoanagenetic individual. We thank the UC Davis DNA Technologies core for sequencing assistance.

Tables and Figures

Table 4.1. Enrichment ratio of novel breakpoints on other genomic features. ***: p-value < 0.001; **: p-value < 0.01; *: p-value < 0.0.5.

Genomic features		1kb window		10kb window		Description
		Enrichment ratio	P-value	Enrichment ratio	P-value	
Chromatin states	Chromatin state 1	1.58	1.59E-6***	1.31	7.27E-11***	TSS, promoter, 5' UTR
	Chromatin state 2	1.28	0.026*	1.14	0.005**	intergenic regions with proximal promoter elements
	Chromatin state 3	1.75	3.47E-8***	1.33	1.86E-9***	transcription elongation signature
	Chromatin state 4	0.91	0.28	1.05	0.34	intergenic regions with distal promoter elements
	Chromatin state 5	0.91	0.34	1.01	0.83	polycomb-regulated chromatin, intergenic region
	Chromatin state 6	0.93	0.54	1.24	1.55E-4***	gene bodies, intragenic region
	Chromatin state 7	1.23	0.09	1.3	2.68E-4***	intragenic region, 55.6% coding sequence, 34.3% intros
	Chromatin state 8	0.38	4.00E-13***	0.49	3.43E-19***	AT-rich heterochromatin
	Chromatin state 9	0.04	4.91E-145***	0.11	8.47E-104***	GC-rich heterochromatin
Dnase I hypersensitive sites		1.36	3.24E-5***	1.21	3.15E-12***	
Gene		1.33	5 82E-21***	1.23	8.00E-30***	
mRNA		1.17	6.18E-9***	1.12	4.94E-12***	
Replication origin		1.5	0.047*	1.29	0.13	
Transposable element		0.31	3.59E-45***	0.42	2.44E-53***	

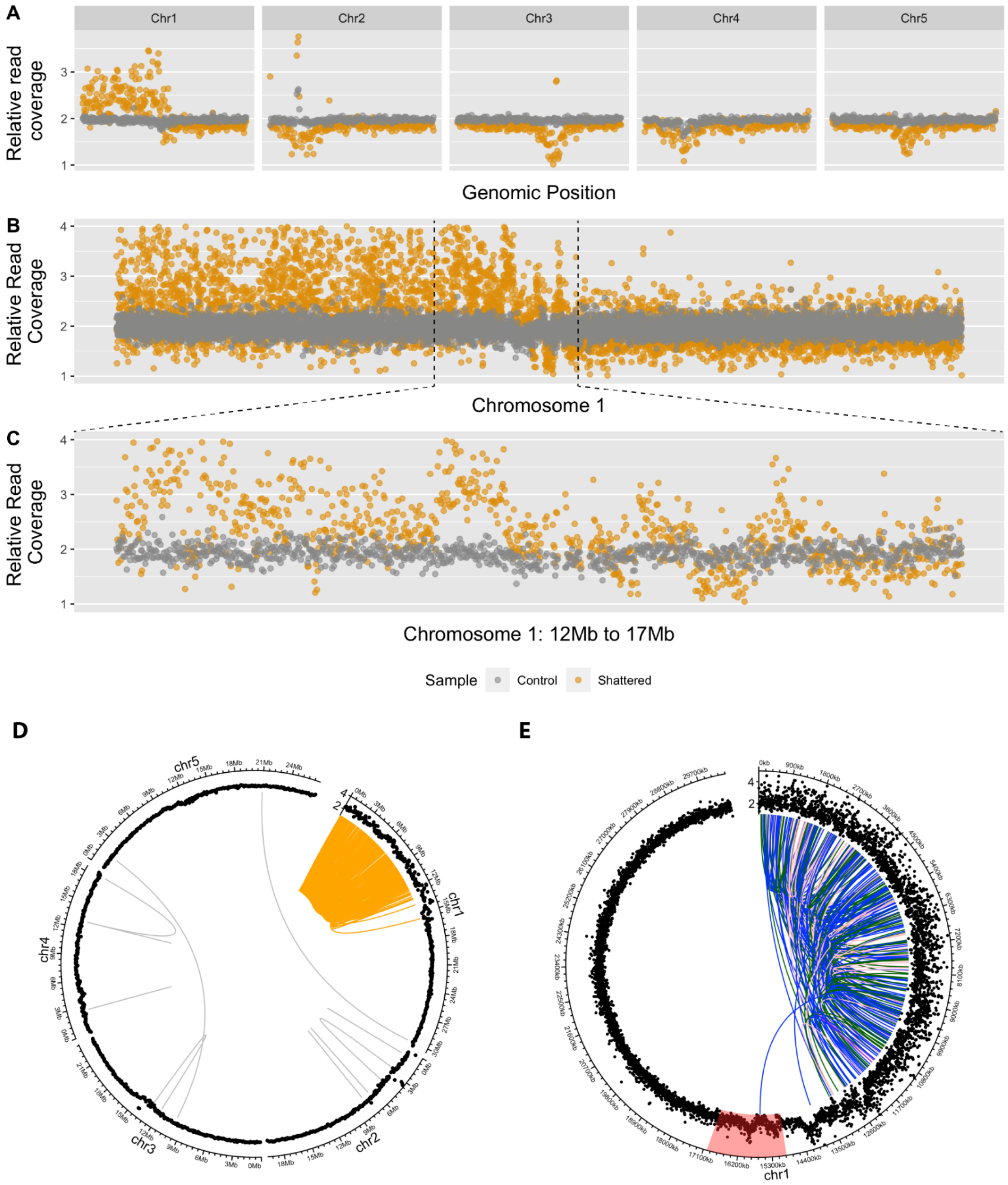


Figure 4.1. See next page for caption.

Figure 4.1. Characteristics of the genomic region with extreme dosage variations in the progeny of *asy1* mutant *A.thaliana*. (A-C) Extremely dense copy number variations on chromosome 1. Each dot represents the normalized read coverage in a bin set along the genome. (A) Relative read coverage across the whole genome (100kb bins). (B) Close-up of chromosome 1 (5kb bins). (C) Further close-up on the region of chromosome 1 that displays dense CNVs (5kb bins). (D, E) Breakpoints were highly enriched over the regions of chromosome 1 exhibiting clustered CNVs. The distribution of DNA junctions on all chromosomes of the *Arabidopsis* genome (D) and just Chromosome 1 (E) are shown with circos plots. The outermost layer indicates chromosomes. The next layer indicates relative read coverage, with 100kb bins (D) and 5kb bins (E). The center arcs represent the locations of breakpoints pairs of the DNA junctions identified. (D) The center arcs are colored in orange if both breakpoints are located within the CNV cluster, and in gray if both breakpoints fall elsewhere in the genome. (E) The center arcs are colored in blue if the breakpoints are connected tail to head, in dark green if they are connected head to head, and in pink if they are connected tail to tail. The red highlighted region represents the chromosome 1 centromere. The CNV cluster represents the first 16.1Mb of Chromosome 1.

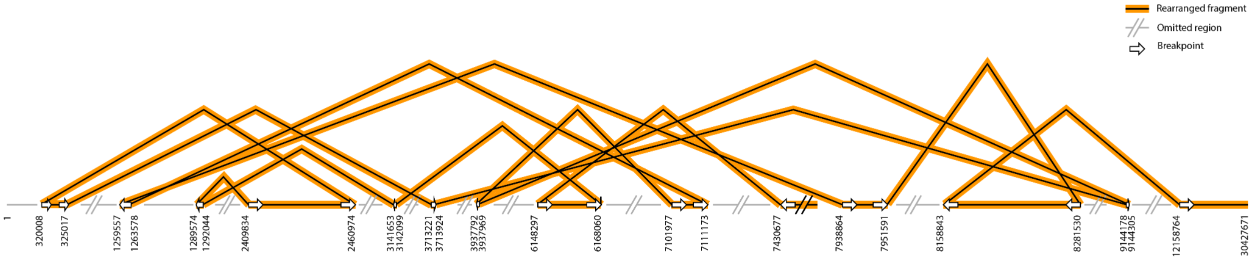


Figure 4.2. Potential structure of the rearranged chromosome. One of the restructured fragments resulting from the chromoanagenesis event, and reconstructed based on the structure of the DNA junctions identified. The horizontal line represents chromosome 1. Segments involved in the rearrangement are shown in black, while segments not involved in this particular rearrangement are shown in gray. Each breakpoint is labeled with an arrow representing the joining orientation, and also its original genomic positions.

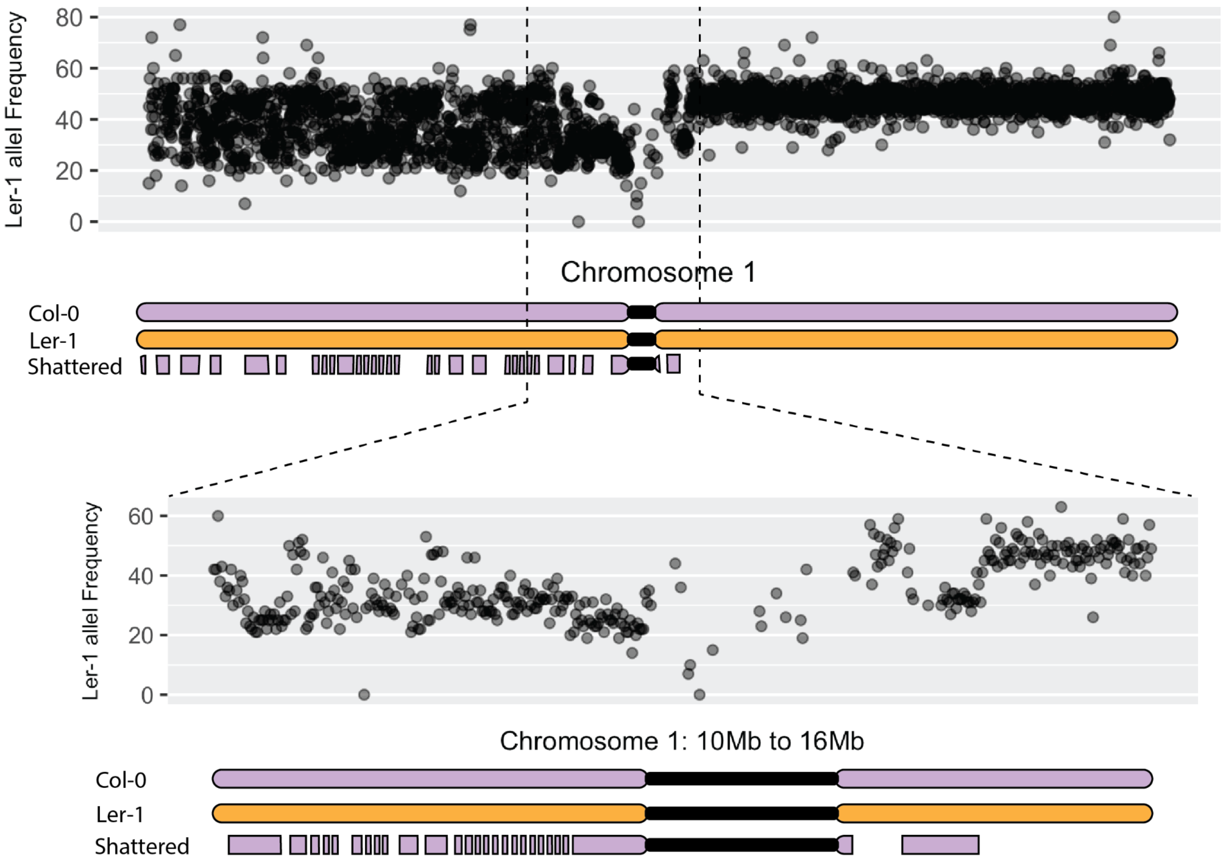
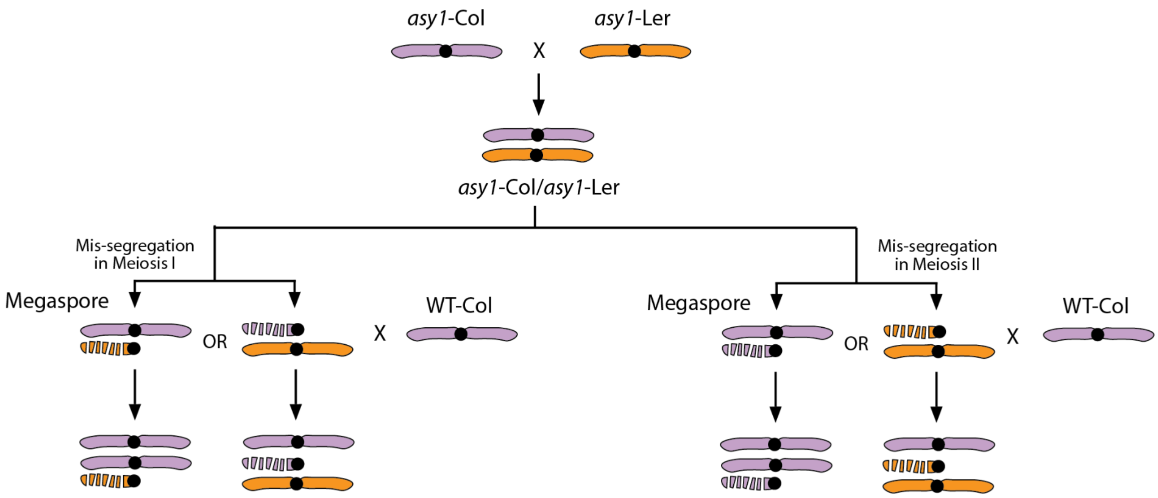


Figure 4.3. Parental allele variation in the shattered region. Variation in Ler-1 allele frequency along chromosome 1 in the shattered *Arabidopsis* line. Each dot represents the percentage of Ler-1 alleles within a 10kb bin. A Ler-1 allele frequency of 50% represents one Ler-1 chromosome, and one Col-0 chromosome. A frequency of 33% represents a 1:2 ratio of Ler-1 and Col-0 alleles, with one chromosome copy from the Col-0 parent and one Col-0 and one Ler-1 copy from the asy1 mutant hybrid parent. The top graph represents Ler-1 allele frequencies along the entire chromosome 1. The bottom graph represents a close-up of the pericentromeric region, from 10Mb to 16Mb. The schematic drawing at the bottom of each graph represents the inferred karyotype of that region, with Col-0 in purple and Ler-1 in orange. The fact that allelic percentages around the pericentromeric region switches between 50% and 33% indicates that two different haplotypes are inherited from the *asy1* mutant hybrid parent. This demonstrates that the shattered chromosome originated from mis-segregation during meiosis I (See Figure 4.4).



Pattern	Type 1		Type 2	
	Shattered	No shattered	Shattered	No shattered
Ler Freq	33%	0%	33%	50%

Pattern	Type 1		Type 2	
	Shattered	No shattered	Shattered	No shattered
Ler Freq	0%	0%	66%	50%

Figure 4.4. Schematic representation of the expected haplotype frequency for the *Arabidopsis* line with a shattered chromosome. The shattered *Arabidopsis* line is the product of two subsequent crosses. The first cross was performed between an *asy1*-Col-0 (female) and an *asy1*-Ler-1 (male), producing *asy1*-Col-0/*asy1*-Ler-1 hybrids. Next, the *asy1*-Col-0/*asy1*-Ler hybrids were used as females, and crossed to Col-0 (WT). During the second cross, the shattered chromosome resulted from mis-segregation during meiosis, either during meiosis I or meiosis II. Meiosis I mis-segregation causes the two homologous chromosomes to be inherited into the megaspore. In this case, the shattered chromosome is expected to carry a different haplotype from the intact chromosome. Meiosis II mis-segregation leads to the incorporation of two sister chromatids into the megaspore. In this case, the shattered chromosome is expected to carry the same haplotype as the intact chromosome. Considering the Col-0 genotype from the male parent, the final Ler-1 genotype frequency in the shattered individual can inform us about the origin of the shattered chromosome. Meiosis I mis-segregation leads to 33% of Ler-1 alleles along the shattered region, while meiosis II mis-segregation results in either 0% or 66% of Ler-1 alleles. Recombination between Ler-1 and Col-0 is not shown here but could lead to coinheritance of the recombinant chromatids, one of which would then undergo chromoanagenesis. The allele frequencies around the pericentromeric regions would remain the same as depicted here though.

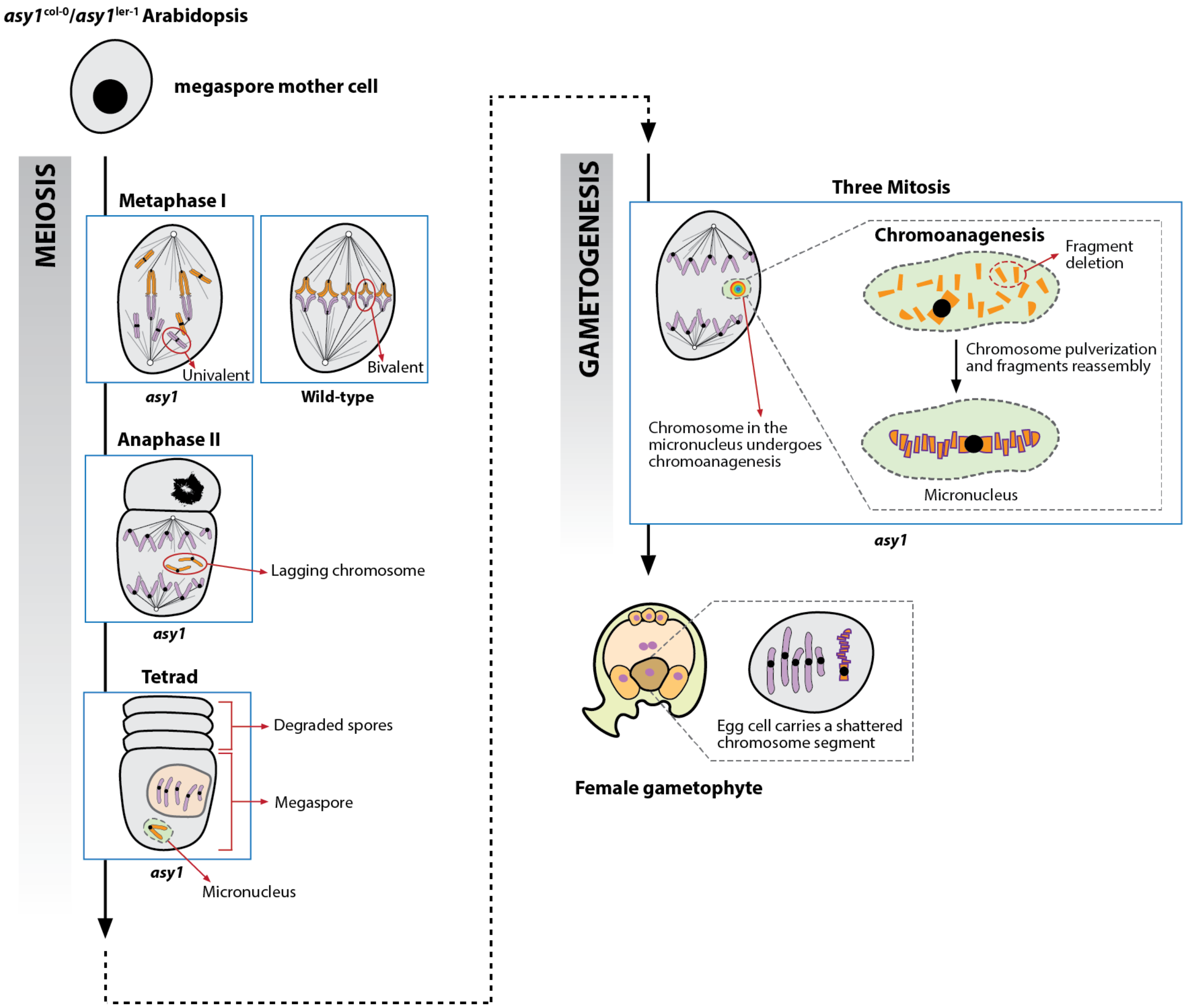


Figure 4.5. Proposed mechanism for chromoanagenesis in the *asy1* homozygous mutant. The megaspore mother cell from *asy1* homozygous mutant exhibits chromosome mis-segregation during female meiosis. Specifically, the *asy1* mutation results in the formation of univalents at metaphase I, which leads to unbalanced chromosome segregation. During meiosis II, the mis-segregated chromosome lags, and is incorporated into a micronucleus. In the following three mitosis during gametogenesis, the chromosome within micronucleus is unable to synchronize with the mitotic division of the main nucleus, and undergoes pulverization and restructuring, resulting in a chromosome with clustered structural variation. This shattered chromosome can be transmitted to the progeny if it is partitioned into the egg cell after micronucleus disassembly.

Supplementary Materials

File S4.1. List of the wild-type *Arabidopsis* lines used for generating two controls.
File S4.2. Summary of all novel DNA junctions.
File S4.3. List of rearranged fragments from novel DNA junctions.
File S4.4. Breakpoints inside gene sequence.

All Supplementary Files are available at https://academic.oup.com/g3journal/advance-article/doi/10.1093/g3journal/jkac185/6654592#supplementary-data

Chapter 5

Phenotypic effect of natural allelic variation and induced dosage variation in *Populus*

[Unpublished]

Abstract

Both allelic variation and dosage variation have important regulatory effects on plant traits. Though many studies have investigated phenotypic variation or natural/dosage variation, very few have documented both and their relative contribution to phenotypic effects remains unclear. The *Populus* genome is highly polymorphic, and poplars are fairly tolerant of gene dosage variation. Here, using a previously established *Populus* hybrid population, we conducted QTL analyses, assessing the effect of natural allelic variation and induced dosage variation, respectively, on biomass, phenology and leaf morphology traits. Our results indicate limited overlap between QTLs from allelic and dosage variation. Overall, integration of QTLs from allelic and dosage variation explains a larger percentage of the phenotypic variance. Our study helps clarify the relationship between allelic and dosage variation and their effects on quantitative traits in *Populus*.

Introduction

Natural allelic variation plays an important role in phenotypic variation in plants [178–187]. The statistical framework raised by R. A. Fisher provides an approach to systematically identify the quantitative trait loci (QTL) responsible for heritable variation [188]. In the last decade, the development of new DNA high-throughput sequencing and genotyping technologies have dramatically improved our ability to identify polymorphic genetic markers between individuals or species [189–191]. This, in turn, enables more accurate QTL identification in plants and animals [192–195].

Despite these technological advances, a wide percentage of the observed variance still remains unexplained by the detected QTLs. This is particularly problematic for complex traits with expected polygenic contributions. For example, the QTL detected following analysis of biomass-related traits in *Populus* explains, on average, 26% of the observed phenotypic variation [67]. To increase biomass yield through tree breeding, we need to consider other types of heritable variations in order to derive a deeper understanding of the regulatory mechanisms at hand.

Dosage variation can affect the phenotypic outcomes of many important traits in plants. Copy number variations (CNVs), especially the ones present on protein-coding regions, can affect phenotypic outcomes in multiple plant species [6,9,11,37,57]. Pan-genomic analyses have identified structural variants across different accessions of multiple plant species, many of which were shown to affect agronomic traits [3,20,56,196,197]. A hypothetical mechanism of how gene dosage affects phenotype was proposed [198]. Gene deletion and duplication can directly affect expression level (*cis*-effect), which in turn affects phenotypes. Dosage variation can also modulate the expression of genes located outside of indels regions (*trans*-effect), since many traits are regulated by a complex network consisting of multiple genetic components [199,200].

We propose that investigating the phenotypic effects of dosage variation and allelic variation, as well as the interplay between these two sources of variation can increase our understanding of the sources of phenotypic variation. For example, when a locus encodes a functional protein whose function is dosage sensitive, the CNV-induced expression changes affect the phenotype. However, if allelic variation is also present, such as a hypomorphic or null allele, two effects are possible: i) CNV affecting the deficient allele would result in no or little phenotypic variation and ii) CNV affecting the normal allele would result in magnified phenotypic variation. In other words, focusing on either the allelic variation or the dosage variation alone only addresses part of the mechanisms at play. A more comprehensive approach, which integrates both types of variations may be better suited to understand the genetic regulatory factors on complex traits such as biomass yield.

Populus is an attractive system to study the interplay between allelic and dosage variation. It is dioecious and therefore an obligate outcrosser and its genome is highly polymorphic, both in terms of sequence polymorphisms and CNVs [56,72]. Genomic structural variation can be maintained through vegetative propagation. In our previous study, we established a *Populus* F1 hybrid population (592 lines) from an interspecific cross between a wild-type egg cell from *P. deltoides* and a gamma-irradiated pollen from *P. nigra [60]*. Whole-genome sequencing analysis revealed that 58% of the F1 lines carry large-scale insertions or deletions (indels). Using this resource, we have identified dosage QTLs associated with a variety of traits including biomass, phenology, leaf morphology and vessel development [61–63]. Since both parental genomes are highly polymorphic, allelic variation is also expected to play an important role in the observed phenotypic variation, but it was not taken into account in these earlier studies. Here, we aim to describe the contribution of allelic variation on phenotypic variation, and to document the possible interaction between allelic and dosage variation in this population.

Specifically, we selected 343 F1 lines from this *Populus* population to investigate the effects of allelic and dosage variation on *Populus* phenotypes. Our results suggest limited overlaps of QTLs between allelic and dosage variation on most traits. Direct integration of QTLs from two types of variation significantly increases phenotypic prediction power. For further understanding the interplay between allelic and dosage variation, combined models including both types of variation may be needed.

Results

A custom computational pipeline efficiently genotyped F1 lines with low-coverage genome sequencing data

The *Populus* F1 lines (592) were originally sequenced at low read depth (~0.5x per line), which was sufficient to identify large-scale indels but was not sufficient to reliably haplotype and genotype this population [201–204]. Fortunately, RNA-seq data from 122 of these F1 lines was also available, as well as Illumina short-read sequencing data from two of the three parental lines (*P. deltoides* 45x, *P. nigra* 65x) [60,62]. Using these resources, we designed a custom computational process to derive parental haplotypes and genotype the F1 lines for both parental contributions (Fig. 5.1 and Table 5.1) (See Materials and Methods).

The process is divided into three steps: parental SNP detection, parental haplotype phasing, and genotyping. Because our population is an F1 population, polymorphisms between the two parental genomes are not informative. Instead, we characterize the parental haplotypes separately. We started by identifying positions that were heterozygous within each parental genome. We selected 37,556 and 33,035 positions that were heterozygous in *P. deltoides* and *P. nigra*, respectively. Next, we used the RNA-seq from 122 diploid F1 lines to derive phased haplotypes for the two parents. Finally, the phased haplotypes were applied to the low-coverage genomic data (~0.5x per line) for genotyping of the remaining F1 individuals.

In summary, we were able to obtain reliable genotype information for 343 F1 lines (Fig. 5.1C). We generated binned markers (50 SNPs per bin) to increase genotype robustness, and a final common marker set of 507 binned markers was generated for multi-genotype QTL analysis that applied to both the *P. deltoides* and the *P. nigra* genomes.

F1 lines selected for QTL analysis carry abundant dosage variation

Among our 343 selected F1 lines, 54.2% (186 out of 343) carry at least one indel. These indels were more often deletions (66.5%) than insertions (33.5%), as observed in the original population [60]. To identify the effect of dosage variation on gene expression, an approach was established to systematically investigate the association between induced dosage variation and phenotypes (Fig. 5.2) [61–63]. By following the approach described in those studies, we characterized dosage variation in 546 dosage binned markers, with an average of 6 indels in each dosage marker (Fig. 5.3A). Next, these dosage markers were combined with the allelic information to obtain a unified marker list, including *P. deltoides* haplotypes, *P. nigra* haplotypes and dosage information for each of the 343 F1 individuals, at all 507 binned markers.

Contributions of allelic and dosage variation on phenotypes can be assigned to QTLs

We investigated three phenotype categories (42 traits) in the *Populus* F1 population: leaf morphology (22 traits), phenology (10 traits) and biomass (10 traits). In total, QTLs were observed for all 42 traits and they were located on all 19 chromosomes (Fig. 5.3, S5.1 and S5.2). Using the single model (Trait ~ Genotype), 111, 83 and 321 QTLs were identified for 42 traits from *P. deltoides*, *P. nigra* and dosage genotypes, respectively (Table 5.2 and File S5.1). Specifically, approximately 45% of the dosage QTLs were consistent with previous results (Fig. S5.3) [61,62]. Observation of different dosage QTLs probably due to (a) different number of F1 lines used in two analyses (This study: 343 lines; Previous studies: 592 lines); (b) different statistical methods used for QTL selection (This study: permutation test on t-values; Previous studies: Benjamini-Hochberg method on p-values). Dosage QTLs for leaf morphology and biomass-related traits co-localized on chromosomes 1, 9. 14 and 19 (Fig. 5.3E and S5.1E), while chromosomes 2, 4, 8 showed colocalization of dosage QTLs for phenology (Fig. S5.2E). Interestingly, QTLs from allelic variation and dosage QTL for the same traits seldomly co-localized. For example, *P. deltoides* QTLs on chromosomes 2 and 10 were found to influence multiple leaf morphology and biomass traits, but dosage variation of these genomic regions did not show significant effect on phenotypes (Fig. 5.3C, E, S5.1C, E). This was particularly true for

P. deltoides QTLs, which were almost exclusively non-overlapping with dosage QTLs. *P. nigra* allelic-variation QTLs, on the other hand, exhibited partial overlaps with the dosage QTLs (Fig. 5.3D, E chromosome 1, S5.2D, E chromosome 17). Comparing QTLs from *P. deltoides* and *P. nigra*, there was some overlap for phenology-related traits, but very little consistency on leaf morphology and biomass (Fig. 5.3C, D, S5.1C, D and S5.2C, D). This result suggested that the two parental species, *P. deltoides* and *P. nigra*, may have similar allelic variation effects on phenology, but different genetic influences on leaf morphology and biomass.

Some of the observed allelic QTLs were consistent with previously identified allelic QTLs. For example, *P. deltoides* QTLs on chromosomes 6 and 10 for phenology-related traits (bud burst), on chromosome 10 for biomass-related traits (height, base diameter, volume), and on chromosome 1 for leaf morphology traits (leaf size, leaf shape, leaf serration) were consistent with previous reported allelic QTLs [64,65,67,68]. On the other hand, other QTLs, such as *P. nigra* QTLs on chromosomes 1 and 11 for leaf morphology, or the common *P. deltoides* and *P. nigra* QTL on chromosome 17 for phenology, have not been reported before. Interestingly, chromosome 17 was previously shown to be a hotspot for leaf shape-related traits [68]. Our observation of QTLs on chromosome 17 suggested that leaf shape and phenology may share genetic regulators or have their regulatory genes located nearby in *Populus*.

To investigate whether allelic and dosage variation together improve the explanation on observed phenotypic variance, we next used a multivariate model to detect allelic and dosage QTLs simultaneously. We examined traits which exhibited significant signals on both allelic and dosage variation. 16 traits were selected (File S5.3). For these 16 traits, at least one QTL associated with the trait was identified for each of the three types of variation. For these 16 traits, integration of QTLs from allelic and dosage variation explained 27.45% of the observed phenotypic variance, which was significantly higher than the percentage of variance explanation by the allelic QTLs only (13.6%, Tukey's test, P < 0.001) or the dosage QTLs only (19.5%, Tukey's test, P = 0.033) (Fig. 5.4 and File S5.2).

Discussion

Identifying candidate genes underlying a target trait is a crucial step for understanding the responsible molecular regulatory mechanisms, and for applying this knowledge to plant breeding. Quantitative trait loci (QTL) analysis, which typically correlates SNP to traits or phenotype-associated features such as gene expression and RNA alternative splicing [205,206], is an efficient approach for this endeavor. Besides SNPs, other genetic features such as dosage variation [61–63] can affect traits of interest. A unique *Populus* population, which carries natural allelic and induced dosage variation was previously established [60]. Our study investigated the effects of allelic and dosage variation on quantitative traits. In general, our results indicate overall limited overlaps between QTLs from allelic and dosage variation.

A single model approach was used to describe the correlation between each variation source and target traits. *P. deltoides* and *P. nigra* genotypic information allowed for the identification of QTLs between different haplotype within each parental species. The results showed few shared QTLs between *P. deltoides* and *P. nigra* genotypes on leaf morphology and biomass related traits, while a couple of phenology QTL overlapped. This suggested that these two *Populus* species share a common regulatory network for phenology. With the current data, it is difficult to determine if the pathways that control leaf morphology and biomass are similar are not. *P. deltoides* and *P. nigra*'s pathways may have diverged after speciation and now be distinct, or they could still be similar but the QTL analysis failed to identify common loci because they are not polymorphic in both parents.

Dosage variation was induced by γ irradiation of *P. nigra* pollen and all indels are located on the *P. nigra* chromosomes [60]. Therefore, I expected to observe some overlap between *P. nigra* QTLs and dosage QTLs. For overlapping QTL, one possible interpretation is that the *P. nigra* QTL carries alleles affecting gene expression levels, which in turn affect protein production. In that case, dosage and allelic variation would have similar effects. with decreased protein level to 0 in the case of a deletion or increased levels to two-folds in the case of an insertion. According to this model, both *P. nigra* QTL and dosage QTL act through dosage-dependent regulation on the target trait. The dosage-dependent behavior is consistent with additivity and was described as the basis of quantitative variation in some studies [207,208].

Many *P. nigra* QTLs, however, did not show significant dosage-mediated signals (Fig. 5.3D, chromosome11; Fig. S5.1D, chromosome 2; Fig. S5.2D, chromosome 6). This may be because 0X to 2X constitutive dosage variation is insufficient to affect protein function, while allelic variation could affect gene function in other ways, such as by modifying the expression pattern, or directly affecting the function of the protein if the AA sequence changes. Alternatively, it is possible that dosage QTLs were not identified because of insufficient dosage variation at those loci in the population. Indeed, over 50% of the *P. nigra* loci are connected to fewer than 5 indels, limiting the statistical power of my dosage QTL analysis. Gene dosage compensation is another possible explanation, where the structural gene dosage effect is canceled by an inverse regulatory effect, exerted either within the same locus or from an unlinked region [198,209]. The combination of these two opposite effects would result in no significant changes of gene expression.

Conversely, dosage QTLs that did not overlap with allelic QTLs were detected in several instances (Fig. 5.3E, chromosomes 9,14,17; Fig. S5.1E, chromosomes 9,14,19; Fig. S5.2E, chromosomes 2,4,8). In those cases, it is possible that allelic variation at these loci has too subtle an impact on the gene expression level to enable statistical identification through QTL analysis, or that allelic variation is just not present for those loci. Induced dosage variation, instead, by directly affecting the presence or level of the responsible protein affects traits in a distinct way. In these cases, the gene balance hypothesis can explain the success in detecting dosage QTLs and the failure of detecting allelic QTLs [198]. According to this hypothesis, traits regulated by multisubunit complexes are particularly sensitive to dosage. Copy number variations involving the genes encoding these subunits can perturb their stoichiometry dramatically altering the complex function and ultimately the connected traits. Genetic variation that alters the dosage of these subunits, on the other hand, may be limited because changes in gene product concentration may be detrimental and subject to purifying selection [198,210]. Genetic variation with subtle effects would be difficult to identify [211].

Integration of QTLs from dosage and allelic variation, compared to either allelic QTLs or dosage QTLs, provide significant improvement on the variance explanation (Fig. 5.4). Specifically, dosage QTLs have larger effects on phenotypes compared to allelic variation. These results suggest that a large percentage of the phenotypic variation was caused by the induced large-scale indels, but not all of it. Some of the phenotypic variation is caused by natural allelic variation, and taking both the allelic and dosage variation into account improves phenotypic prediction. However, integration of all identified QTLs from the single models only explained on average 27.45% of the observed phenotypic variance, indicating that the majority of the variation remains unexplained. One possibility is that it is due to the interaction between allelic and dosage variation. For example, dosage effects are expected to be allele-sensitive if the responsible gene is heterozygous for a null allele (Fig. 5.5). As a result, single models including only the allelic variation or the dosage variation are not able to identify the interactive effects. An integrative model is needed. One option is to fit allelic and dosage variation together into a multivariate linear regression model. For example, Trait ~ *P. deltoides* + *P. nigra* + Dosage can identify significant signals for each variation type while considering the effect of other covariates simultaneously. A limitation is that our three variables may be correlated. For example, if dosage is 0 (deletion) at the *P. nigra* locus*,* the *P. nigra* haplotype becomes "NA". This problem is likely to cause the loss of significant signals. The other option is to assume that there are interactions between each pair of variation types. In that case, we derive genotypic states for every individual that encompass all three types of information. For example, D1.N1.1 on Marker 1 represents the individuals with *P. deltoides 1*, *P. nigra 1*, and 1 *P. nigra* copy in Marker 1. In this scenario, all individuals would fit in one of 10 possible states and we can fit integrated genotypic states into a univariate model such as Trait ~ States. To further understand the differences between each genotype states, pairwise comparisons can be performed after linear regression. Contrasts showing significant differences between two genotypic states indicate the source of phenotypic variation. The limitation of this method is that some genotypic states may not be represented by enough data points. For example, indels are originally less prevalent compared to regular states in the population. When adding allelic information, sometimes the indel genotypic state only contains 1 or 2 individuals, which cannot be used to perform statistical tests. Future work can focus on the investigation of interactive effects between allelic and dosage variation, probably by establishing integrative approaches.

Taken together, I investigated the contribution of natural allelic variation and induced dosage variation in F1 *Populus* hybrids on quantitative traits. I found limited overlap between allelic and dosage variation QTLs, suggesting that the naturally occurring sequence polymorphisms and the induced structural variation may act on the examined traits through different regulatory factors. The phenotypic variance explained by only the allelic variation or the dosage variation is still connected to a large missing heritability. Integrating the QTLs from allelic and dosage variation significantly increases the phenotypic variance explanation compared to only allelic or dosage QTLs. New methods may be needed for future study to investigate the interaction between allelic and dosage variation in detail, which in turn should improve the manipulation of beneficial traits in *Populus*.

Materials and Methods

Data acquisition and preprocessing

Genomic sequencing data, RNA-seq data, and phenotypic information were obtained from previous studies [60–62,212]. Briefly, an interspecific cross between wild-type *P. deltoides* and pollen-irradiated *P. nigra* produced 592 F1 hybrid lines. High-coverage Illumina short-read sequences were obtained from the two parental lines with read depth around 45x and 65x for *P. deltoides* and *P. nigra*, respectively. Additionally, low-coverage Illumina genome sequences were obtained from each of the F1 hybrid clones (read depth around 0.5x per line). Leaf RNA sequencing was performed on 166 F1 lines, each in triplicates. The raw RNA-seq reads were pooled per clone and used to assist in haplotype phasing. The collection and statistical analysis of phenotypic information were described in previous studies [61,62]. Three categories of phenotypes - leaf morphology, phenology, and biomass - were used in our study.

The preprocessing of sequencing data followed a custom pipeline developed previously. It starts with a demultiplexing step by using the custom pipeline (https://github.com/Comai-Lab/allprep) for separating raw reads into individual libraries. Reads were aligned to the *Populus* reference *P. trichocarpa* v3.0 [72], using a custom Python script based on BWA [155] (https://comailab.org/data-and-method/bwa-doall-a-package-for-batch-library-processing-and-alignment/). Bam files were generated in this step, which were used to obtain an mpileup file using a custom Python package (https://github.com/Comai-Lab/mpileup-tools) based on Samtools [159], followed by a simplification step to convert the mpileup file into a parsed-mpileup file.

Haplotype phasing

To describe the allelic variation within the two parental clones, we identified heterozygous positions in each parent and determined the phasing between these positions, using a custom computational pipeline (https://github.com/guoweier/QTL_manuscript). Specifically, we started by identifying single nucleotide polymorphisms (SNPs) that can distinguish between two haplotypes within a parent (Fig. 5.1A). The parsed-mpileup file of two parents was generated through the preprocessing approach described above. The parsed-mpileup file was used to obtain desired SNPs. In short, we selected two lists of SNPs, one for *P. deltoides* and the other for *P. nigra*. The example of *P. deltoides* SNPs selection is shown in Fig. 5.1A. For *P. deltoides*, we selected positions that show heterozygous in *P. deltoides* homozygous in *P. nigra*; or positions that have heterozygous in *P. deltoides* and different heterozygous allele combinations in *P. nigra*.

Next, we used RNA-seq data obtained from a subset of the F1 individuals to derive phased parental haplotypes (Fig. 5.1B). Briefly, we first used the RNA-seq raw data from the diploid F1 lines for haplotype phasing, after retaining the positions that are well highly covered in the RNA-Seq data. Second, we treated RNA-seq raw data as genomic sequencing data, with the preprocessing approaches that have been described above. Parsed-mpileup file with 122 RNA-seq lines was obtained after running the pipeline. Then, the RNA-seq parsed-mpileup file was used to identify inherited alleles from *P. deltoides* and *P. nigra*, respectively. Then, we collected the adjacent SNPs combination orders, and recorded the order as parental haplotypes when more than 90% (109 out of 122) of RNA-seq lines carried it.

Genotyping

The adjusted phased haplotypes were applied to low-coverage sequencing data for genotyping. Specifically, for each SNP marker, genotype in F1 hybrids were only recorded when it inherited the alternative allele. Then recorded genotypes were binned by SNP numbers (50 SNPs per bin) to increase the robustness of genotype information. The same genotyping process using adjusted phased haplotypes was also applied with RNA-seq data. The transcriptomic genotypes and genomic genotypes were compared manually (File S5.4). Specifically, we sorted F1 lines based on their read-depth of low-coverage genome sequencing data. Then we selected lines for subsequent QTL analysis based on a) Genomic genotypes can clearly show the pattern along the whole genome and, b) for RNA-sequenced F1 lines with similar read-depth, its genomic genotypes and transcriptomic genotypes are consistent. Finally, 343 lines were selected to proceed for QTL analysis. Transcriptomic and genomic genotypes comparison of chromosome 1 on the selected F1 line with the lowest read-depth was shown in Fig. S5.4.

Dosage variation quantification

Methods of quantifying dosage variation have been described in previous studies [61]. Shortly, we defined bins based on indels breakpoints and tiled bins on the chromosomes. For each bin, the dosage genotype was determined by counting the number of *P. nigra* chromosome copies, since all the dosage variation comes from *P. nigra*. For example, one *Populus* line carries a deletion which occupies 4 bins on the chromosome 10. The dosage genotype for these 4 bins were set to 0, while the rest of bins on chromosome 10 were set to 1. Dosage genotypes were acquired for 343 lines carrying SNPs genotypes.

QTL analysis

To perform a QTL analysis that includes both allelic and dosage variation at the same time we created a common marker list for three types of variation *P. deltoides* haplotype, *P. nigra* haplotype, dosage, using a custom Python pipeline (https://github.com/guoweier/QTL_manuscript). First, we identified physical positions of binned markers in *P. deltoides* and *P. nigra* genotypes, respectively. We then imputed genotypes in the unknown regions using information from their flanking binned markers. For example, on the *P. nigra* genotype, marker 1 is Chr01_1_10000 with genotype N1 and marker 2 is Chr01_20000_30000 with genotype N1. So the genotype in Chr01_10001_19999 is N1. If two flanking markers contained different genotypes or there is NA in flanking markers, the genomic region in between was assigned NA. Second, we built a common marker list for the two parents, using *P. deltoides* markers as the reference and imputed *P. nigra* genotypes based on markers' physical positions. Last, we applied the common marker onto dosage genotype and obtained dosage state for each new marker.

Single models were established for analyzing the correlation between phenotypes and each variation type. The model is specified as:

$$Y_i = \beta_0 + \beta_1 gt_i + \varepsilon_i$$

where Y_i is the phenotype; β_0 is the intercept; β_1 is the unknown coefficient; gt_i is one of the examining genotypes (*P. deltoides* or *P. nigra* or dosage); ε_i is the residual variance. *P. deltoides* haplotype is given the levels D1 and D2. *P. nigra* haplotype is given the levels N1 and N2, while deletion regions are assigned as NA. Dosage is given the levels 0 (deletion), 1 (regular) and 2 (insertion). QTLs were selected with a permutation test on t-values [213]. In short, for each trait and each genotype, the phenotype data were randomized for 343 F1 lines and then performed a linear regression with all the markers along the genome. The maximum t-value of all markers was selected. This randomization process was repeated 1000 times. Then we selected the top 5%

and 1% of maximum t-values. In the observed dataset, the markers with t-values larger than 5% threshold were considered as significant, and larger than 1% threshold were considered as confirmed. Significant markers next to each other were considered belonging to the same QTL. To investigate how much phenotypic variance that can be explained with a single QTL, we did the QTL mapping with multivariate models including all markers underneath that QTL and extracted adjusted R-squares. For phenotypic variance explained by all QTLs of one trait, we took the most significant marker (marker with the largest t-value) underlying each QTL, and then ran multivariate models including these selected markers. Integration of QTLs from allelic and dosage variation follows the similar approach. For each trait, we collected the most significant marker of each QTL, and fit these selected markers into a multivariate model. Adjusted R-square values were recorded.

Tables and Figures

Table 5.1. Summary of marker numbers in each step of the genotyping pipeline.

	P. deltoides	*P. nigra*
SNPs between two parents (initial list)	1,850,175	
Selection of SNPs present at high coverage in the RNA-Seq data	37,556	33,035
Haplotype phasing using the RNA-Seq data	30,475	25,495
Binned markers (50 SNPs/bin)	618	520
Binned markers after genetic map construction	530	473
Common binned markers	507	507

Table 5.2. Summary of observed QTLs in single models.

Phenotype (Total # of traits)	Genotype	Total # of QTL	# of traits with QTL	Variance explained by single QTL (μ ± σ) (%)	Variance explained by all QTLs of a trait (μ ± σ) (%)
Leaf (22)	*P. deltoides*	56	17	2.4 ± 1	6.8 ± 3
	P. nigra	33	12	3.5 ± 1.8	9.9 ± 7.9
	Dosage	160	19	3 ± 1.6	17.4 ± 8.4
Phenology (10)	*P. deltoides*	33	10	2.6 ± 1.2	6.8 ± 1.8
	P. nigra	32	9	2.9 ± 1.3	10.8 ± 6.9
	Dosage	105	10	3 ± 1.9	23.1 ± 9.9
Biomass (10)	*P. deltoides*	22	8	2.2 ± 0.5	4.2 ± 1.9
	P. nigra	18	4	2 ± 0.5	3.5 ± 2.8
	Dosage	56	9	2.3 ± 0.9	10.3 ± 4.1

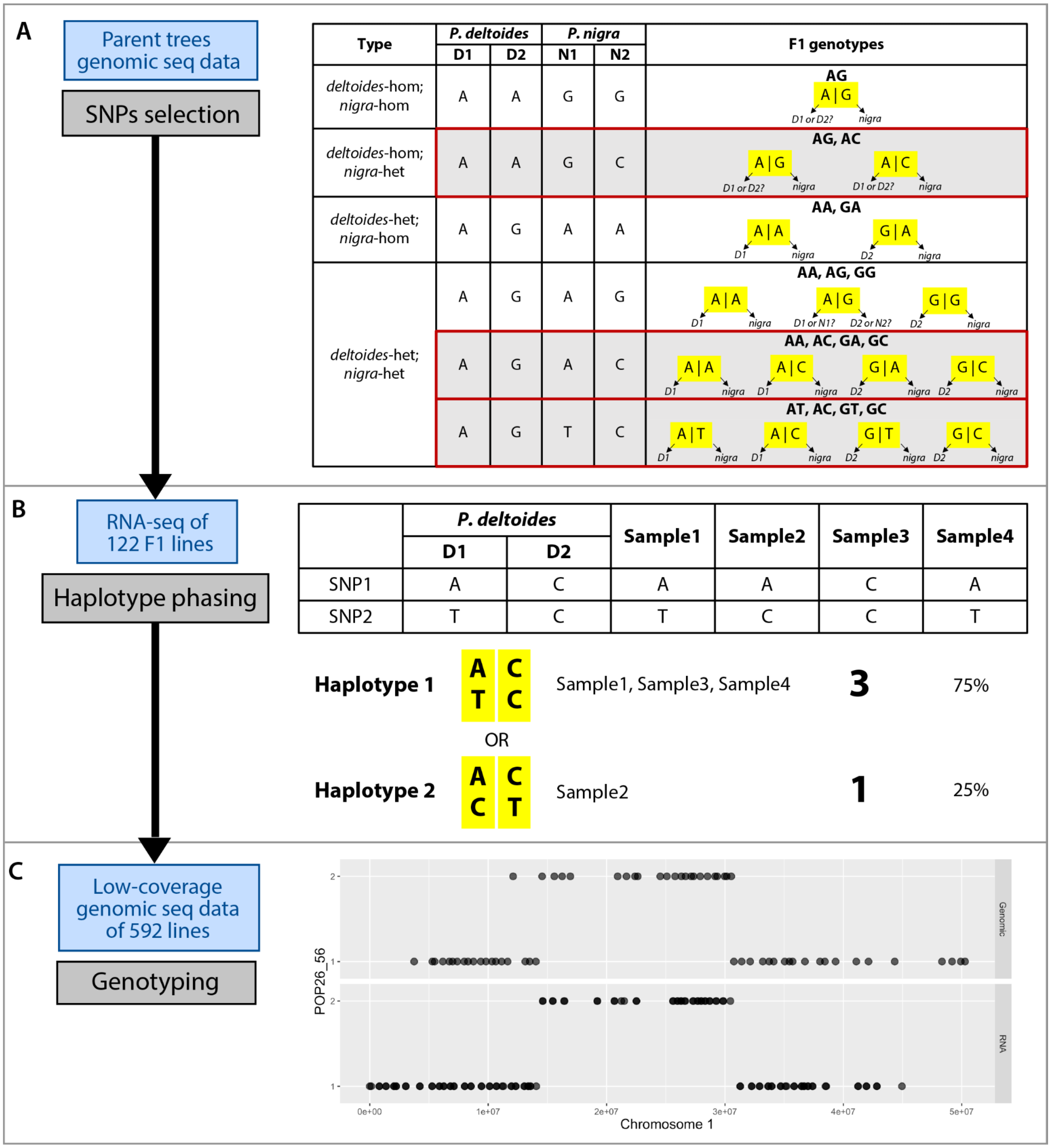
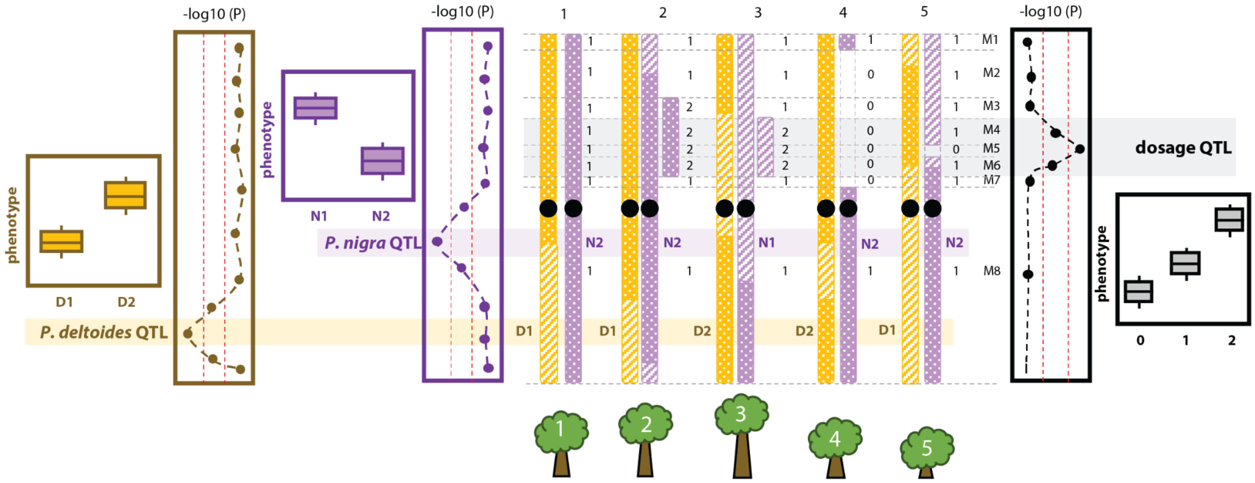


Figure 5.1. See next page for caption.

Figure 5.1. F1 genotyping approach. For each panel on the left, the blue box represents data used, and the black box represents the pipeline step. The diagrams on the right side provide more details for each step. (A) Selection of SNPs that can distinguish between the two parental haplotypes. The table on the right lists all possible F1 genotypes.The rows highlighted in gray indicate the positions selected for *P. deltoides* haplotyping. (B) Haplotype phasing using high-coverage RNA-seq data. The panel on the right exemplifies how we determined parental phasing between two adjacent SNP using only 4 RNA-seq samples. In practice, genotype information from 122 RNA-seq lines were applied at this step and the threshold for an acceptable haplotype combination was set to 90% (110 out of 122). (C) Extrapolate F1 genotypes by applying the phased haplotypes information to the low-coverage sequencing data. The panel on the right shows a comparison of the *P. deltoides* haplotypes obtained from low-coverage sequencing data (top plot) and from RNA-seq data (bottom top) for chromosome 1.



	1	2	3	4	5
Tree height	2 m	3 m	4 m	1 m	0.5 m
P. deltoides	D1	D1	D2	D2	D1
P. nigra	N2	N2	N1	N2	N2
Dosage	1	2	2	0	0

▨ *deltoides 1* ▨ *nigra 1*

▩ *deltoides 2* ▩ *nigra 2*

Figure 5.2. Representative illustration of QTL analysis using both allelic and dosage variation information. QTLs detected from sequence variation within *P. deltoides* (orange), sequence variation within *P. nigra* (purple) and dosage variation (gray) can all contribute to the same trait (here tree height). *P. deltoides* and *P. nigra* haplotypes were acquired through analysis of allelic variation within each parent (D1/D2 or N1/N2). Dosage information was obtained through the calculation of relative copy number states in each bin chromosome bin (See details in Material and Methods).

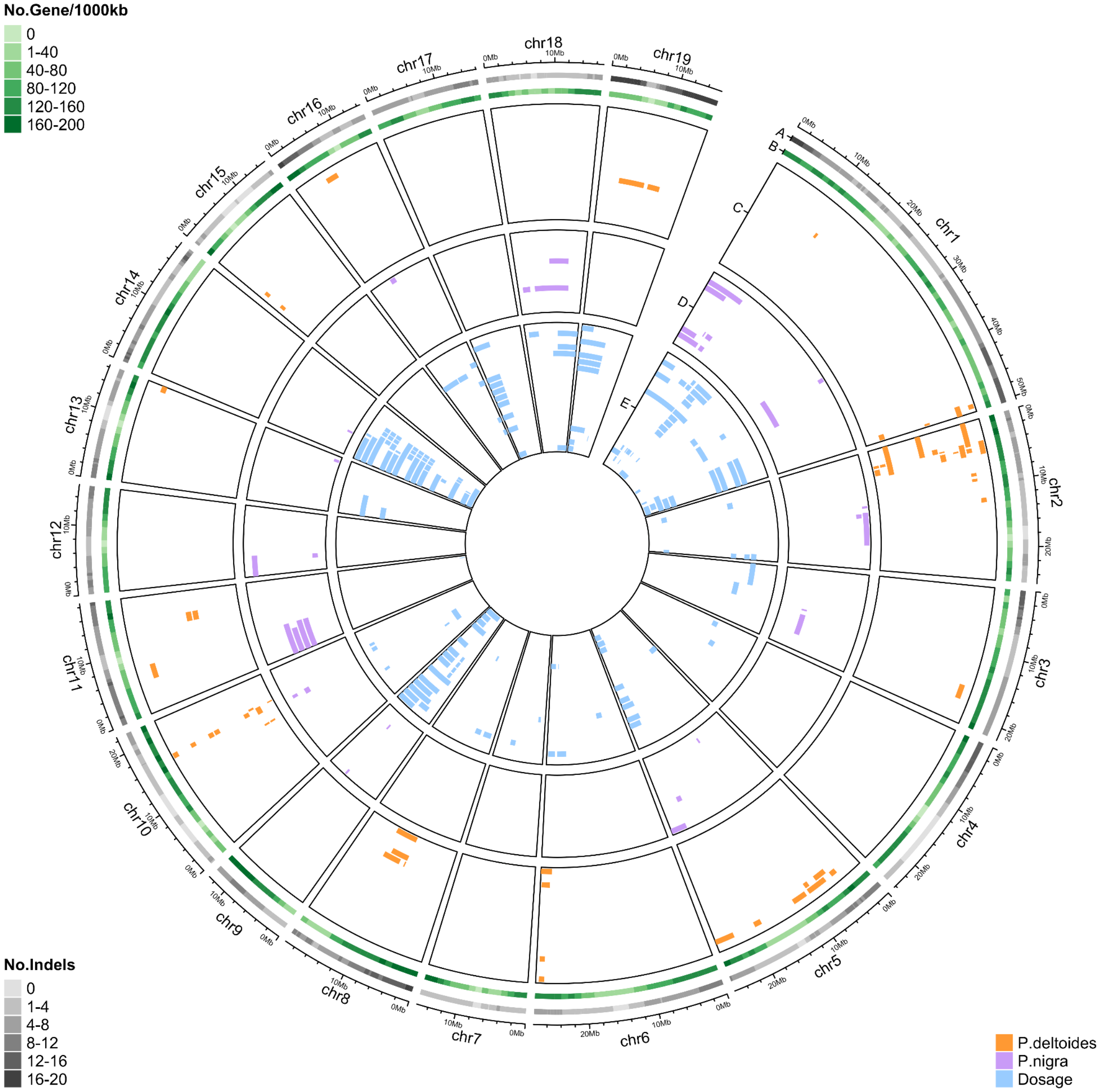


Figure 5.3. See next page for caption.

Figure 5.3. Observed QTLs for leaf morphology traits, using single models. (A) Number of lines carrying indels under each bin. The bins were defined by the boundaries of indels which are tiled on the genome. (B) Gene density across the genome. (C-E) QTLs detected based on variation in *P. deltoides* haplotypes (C)*, P. nigra* haplotypes (D) and dosage (E). The traits from outermost to innermost in each track are: (C) Area_y1_y2, Circularity_y1_y2, Horizontal_symmetry_y1_y2, Width_y1_y2, Indent_depth_y1_y2, Indent_width_y1_y2, Indent_number_y1_y2, PC1:PC2_y1_y2, PC1:PC3_y1_y2, PC1:PC4_y1_y2, PC2:PC3_y1_y2, PC3:PC4_y1_y2, PC3_y1_y2, Perimeter_y1_y2, Perimeter2:Area2_y1_y2, Length_y1_y2, Length:width_y1_y2; (D) Horizontal_symmetry_y1_y2, Width_y1_y2, Indent_depth_y1_y2, Indent_width_y1_y2, PC1:PC2_y1_y2, PC1:PC3_y1_y2, PC1:PC4_y1_y2, PC1_y1_y2, PC2:PC3_y1_y2, Permieter2:Area2_y1_y2, Length_y1_y2, Vertical_symmetry_y1_y2; (E) Area_y1_y2, Circularity_y1_y2, Width_y1_y2, Horizontal_symmetry_y1_y2, Indent_depth_y1_y2, Indent_width_y1_y2, Indent_number_y1_y2, PC1:PC2_y1_y2, PC1:PC3_y1_y2, PC1:PC4_y1_y2, PC1_y1_y2, PC2_y1_y2, PC3:PC4_y1_y2, PC3_y1_y2, PC4_y1_y2, Perimeter_y1_y2, Perimeter2:Area2_y1_y2, Length_y1_y2, Length:width_y1_y2.

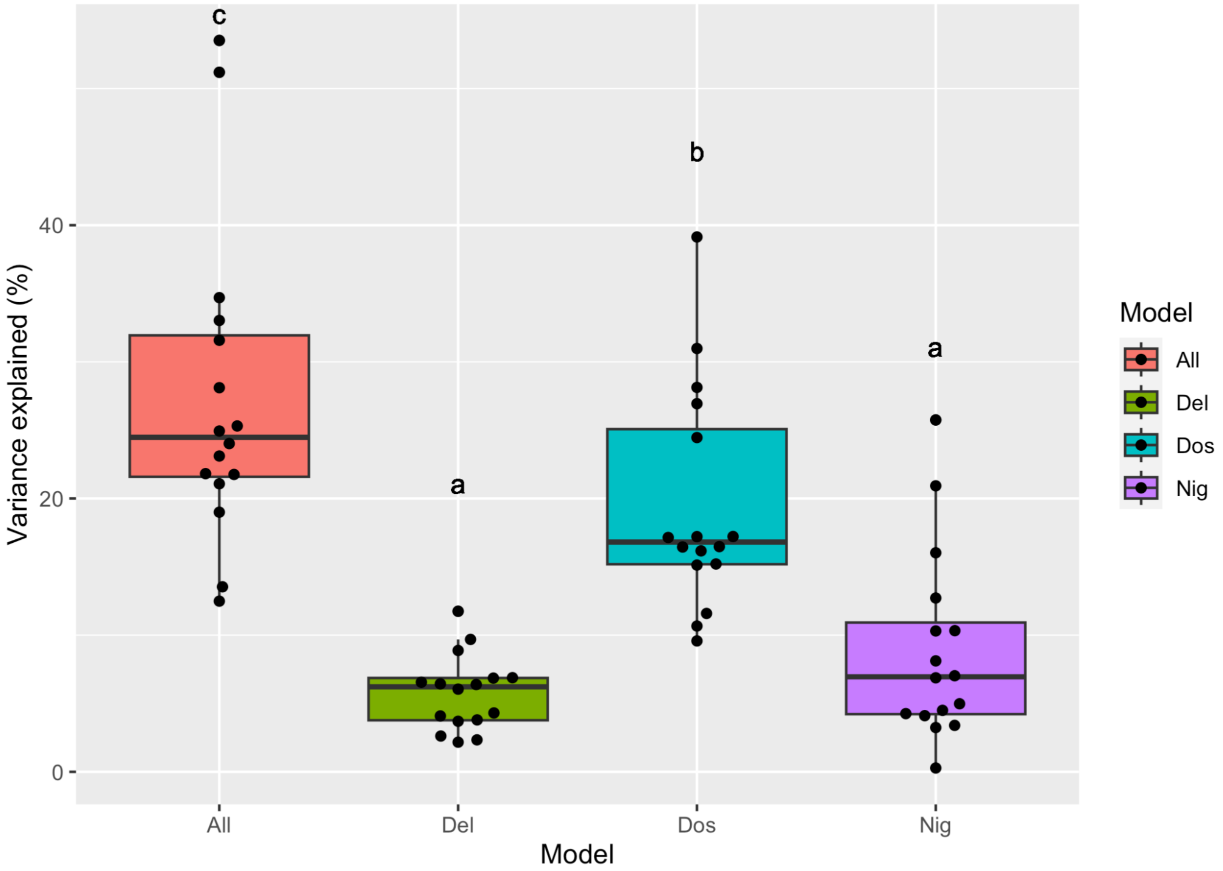
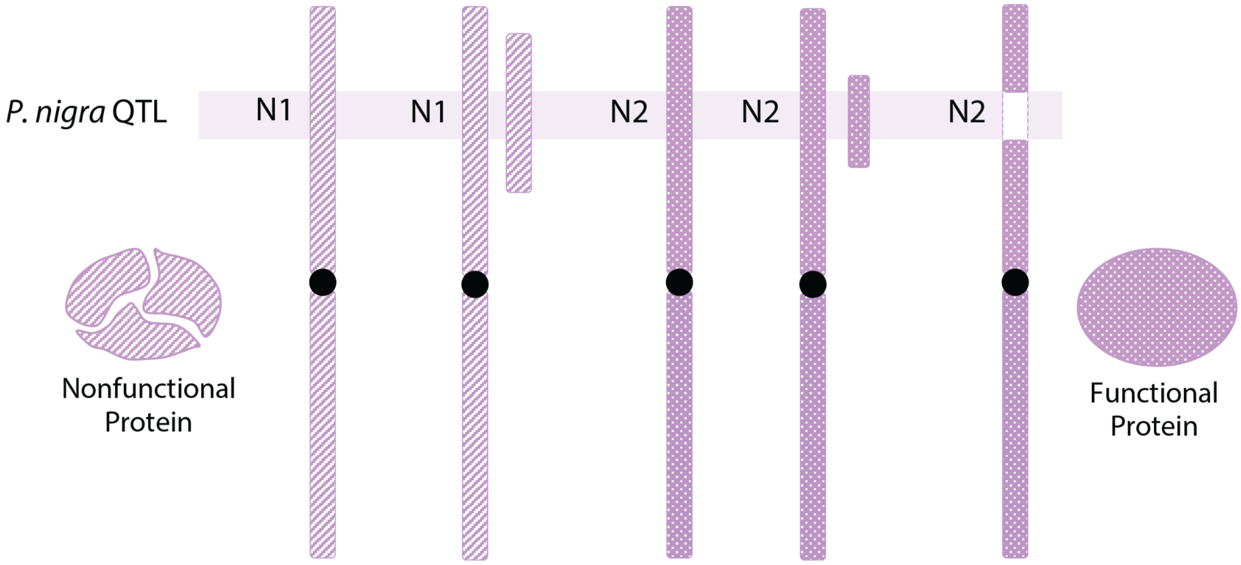


Figure 5.4. Percentage of phenotypic variance explained by QTLs in 16 traits. Allelic and dosage QTLs were both observed in these 16 traits. Variance explained by each model and the collection of QTLs from all three single models. Del, Nig, Dos represent R-square values of models using *P. deltoides* haplotypes, *P. nigra* haplotypes and Dosage, respectively. All represent R-square values of the collection of QTL markers from three single models.



Functional protein	**0**	**0**	**1**	**2**	**0**
Dosage	1	2	1	2	0
P. nigra genotype	N1	N1	N2	N2	*N/A*
Dosage of N2	0	0	1	2	0

▨ *nigra 1*

▦ *nigra 2*

Figure 5.5. Representative diagram of interplay between *P. nigra* haplotypes and dosage variation. N1 (*nigra 1*) encodes a functional protein, while N2 (*nigra 2*) encodes a nonfunctional protein. Copy number changes on N2 have no effect on phenotypes, while copy number changes on N1 result in dramatic differences on phenotypic outcomes.

Supplementary Materials

Supplementary Figures

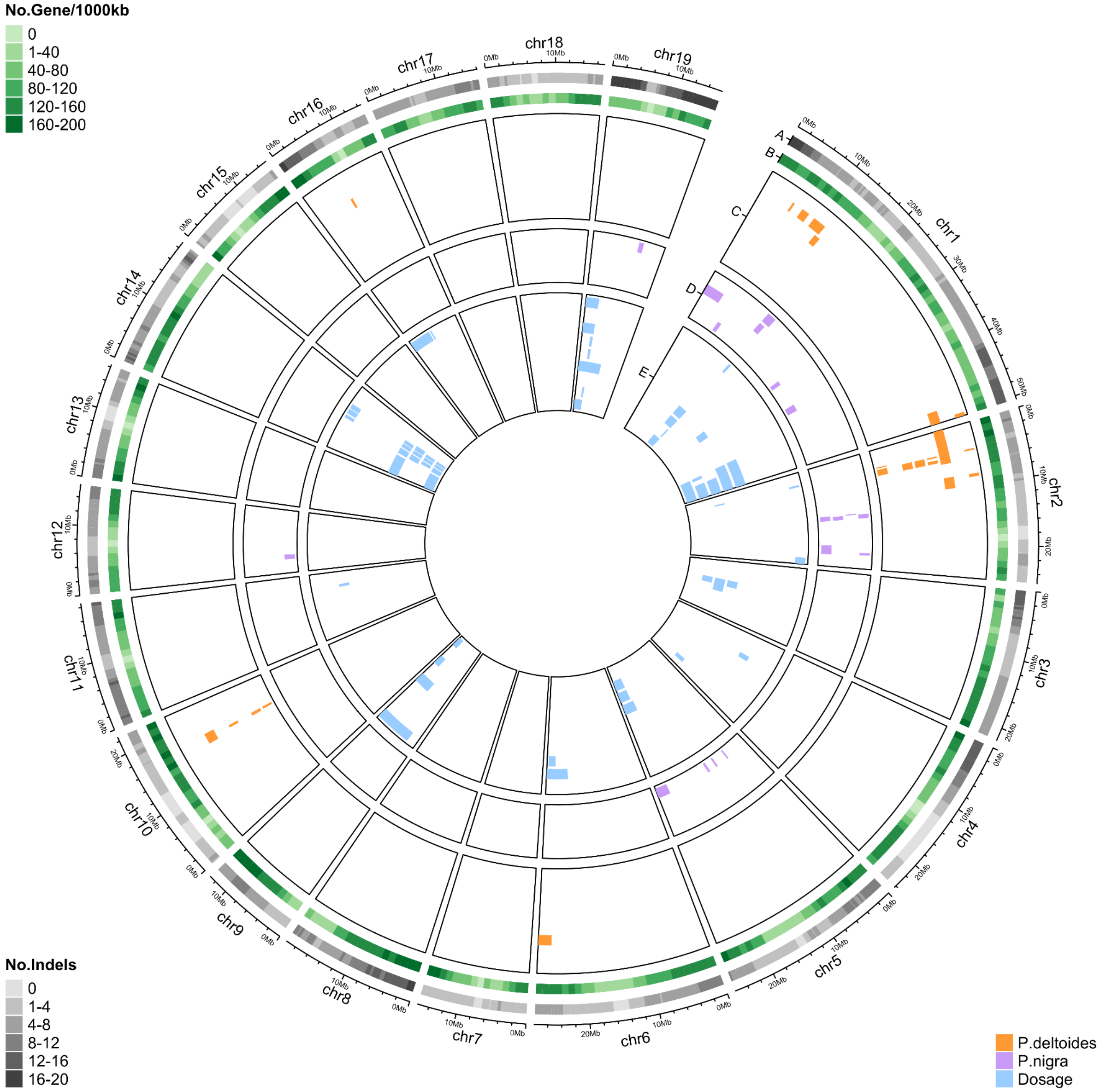


Figure S5.1. See next page for caption.

Figure S5.1. Observed QTLs for biomass-related traits with single models. Similar diagram format with Figure 5.3. (A) Number of lines carrying indels under each bin. (B) Gene density across the genome. (C-E) QTLs detected from *P. deltoides* (C)*, P. nigra* (D) and dosage (E) genotypes. The traits from outermost to innermost in each track are: (C) AUC_height, Coppicing_y1, Diameter_base, Height, Time_serie_diameter_base, Time_serie_diameter_base_height, Time_serie_height, Volume; (D) AUC_height, Coppicing_y2, Time_serie_diameter_base_height, Time_serie_height; (E) AUC_height, Coppicing_y1, Coppicing_y2, Diameter_base, Height, Time_serie_diameter_base, Time_serie_height, Time_serie_volume, Volume.

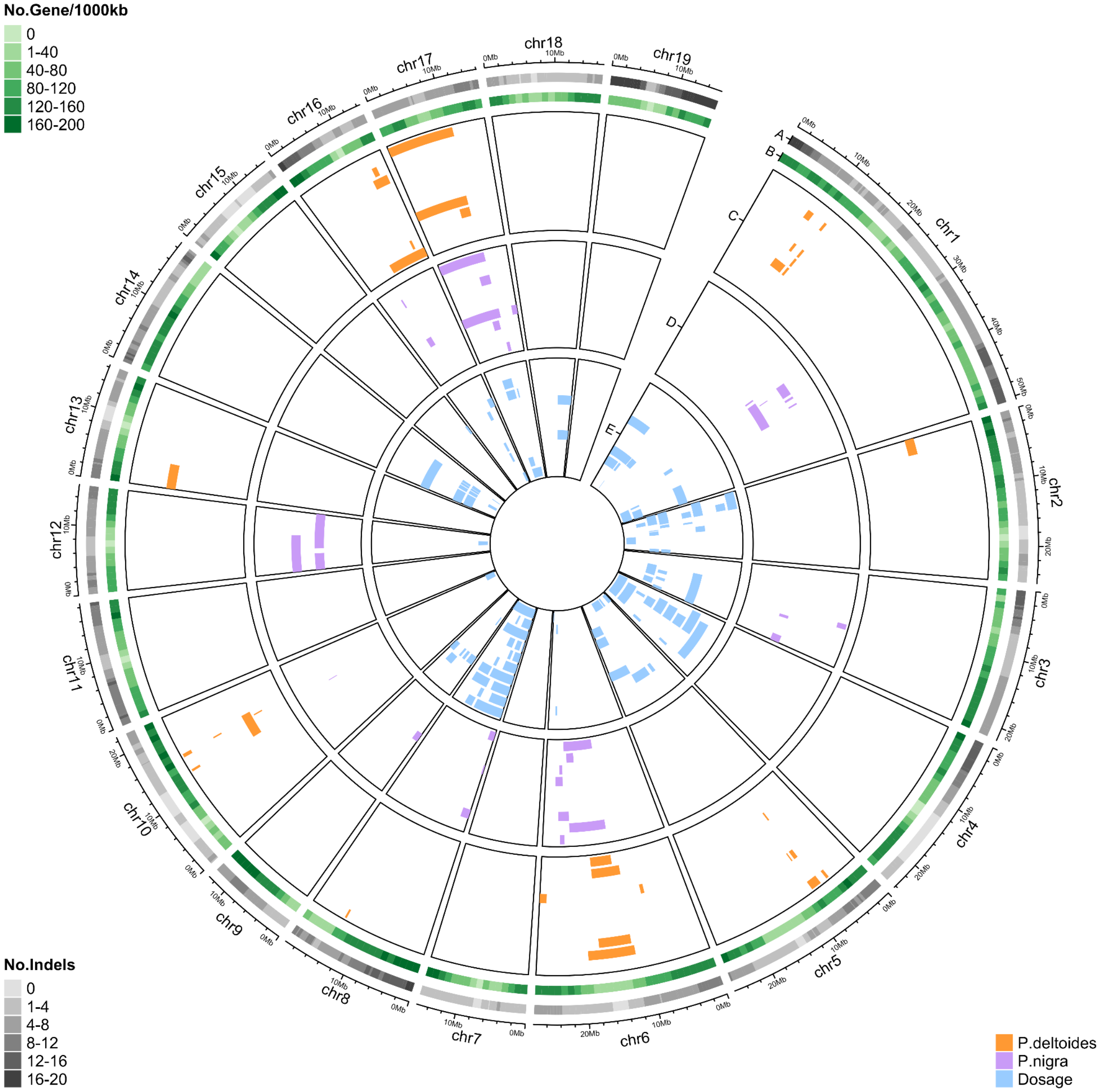


Figure S5.2. Observed QTLs for phenology traits with single models. Similar diagram format with Figure 5.3. (A) Number of lines carrying indels under each bin. (B) Gene density across the genome. (C-E) QTLs detected from *P. deltoides* (C)*, P. nigra* (D) and dosage (E) genotypes. The traits from outermost to innermost in each track are: (C) AUC_bud_burst_y1_y2, AUC_color_y1_y2_y3, AUC_drop_y1_y2_y3, Bud_burst_y1_y2, Color_y1_y2_y3, Drop_y1_y2_y3, Time_serie_bud_burst_y1_y2, Time_serie_color_y1_y2_y3, Time_serie_drop_y1_y2_y3, Green_canopy_duration_y1_y2; (D) AUC_bud_burst_y1_y2, AUC_color_y1_y2_y3, Bud_burst_y1_y2, Color_y1_y2_y3, Drop_y1_y2_y3, Time_serie_bud_burst_y1_y2, Time_serie_color_y1_y2_y3, Time_serie_drop_y1_y2_y3, Green_canopy_duration_y1_y2; (E) AUC_bud_burst_y1_y2, AUC_color_y1_y2_y3, AUC_drop_y1_y2_y3, Bud_burst_y1_y2, Color_y1_y2_y3, Drop_y1_y2_y3, Time_serie_bud_burst_y1_y2, Time_serie_color_y1_y2_y3, Time_serie_drop_y1_y2_y3, Green_canopy_duration_y1_y2.

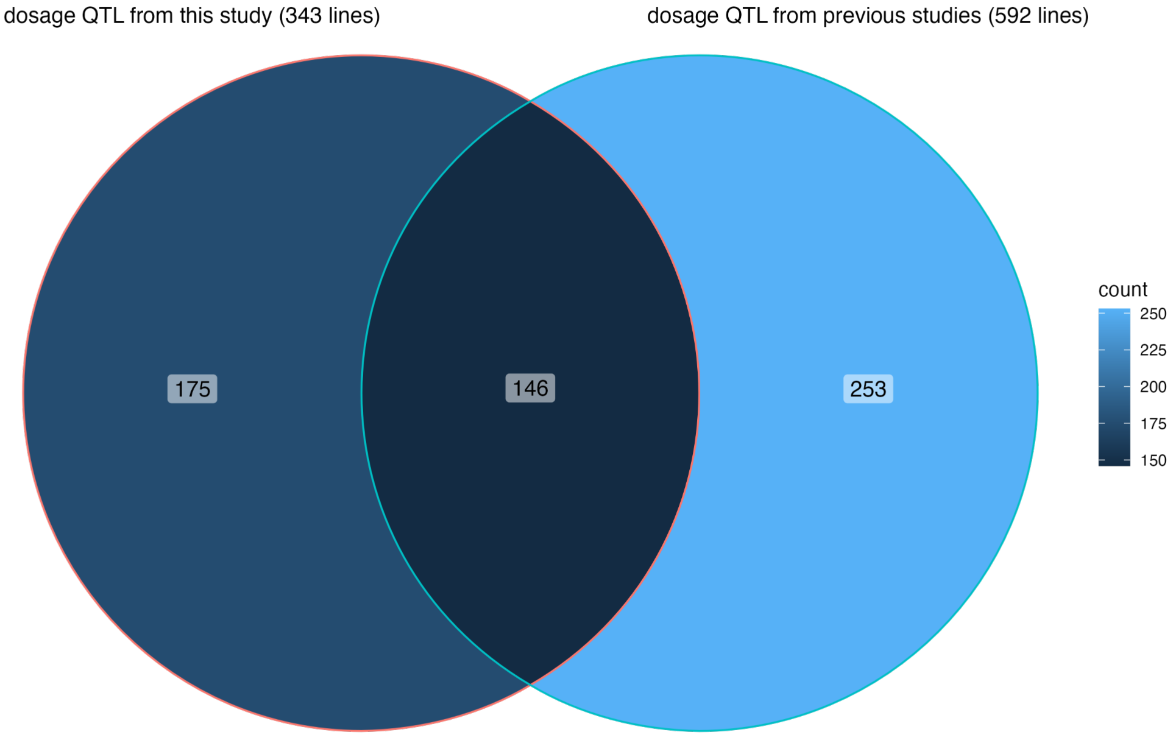


Figure S5.3. Comparison of dosage QTLs observed in this study and previous studies. Left circle represents dosage QTLs observed through QTL model Trait ~ Dosage in this study. Right circle represents dosage QTLs observed in previous studies [61,62]. The overlap region represents the common dosage QTLs generated from two analyses. Numbers in the circles indicate QTL counts.

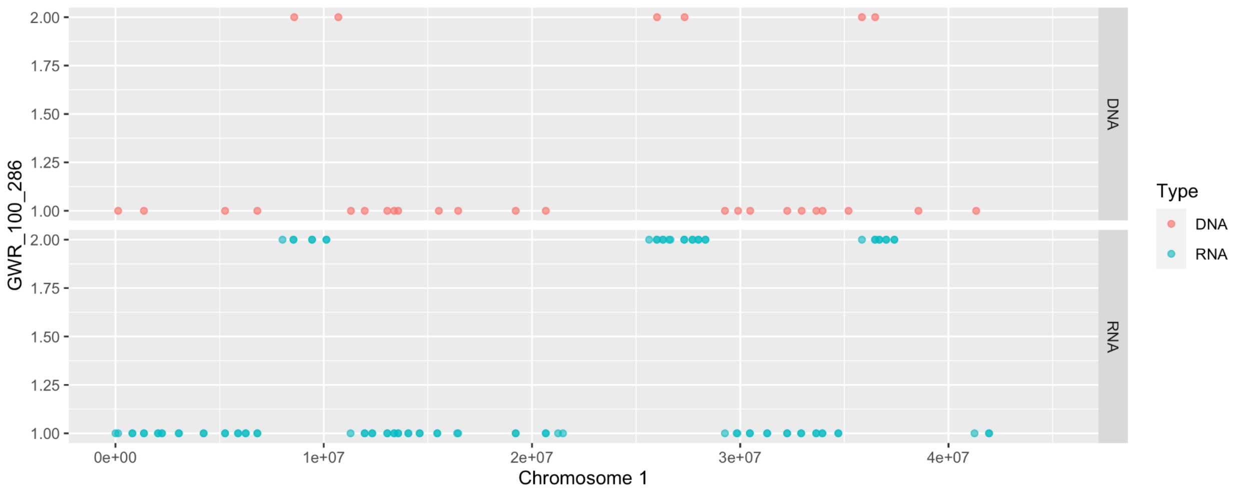


Figure S5.4. Comparison of genomic and transcriptomic genotypes of *P. nigra* haplotypes on chromosome 1 of the F1 lines GWR_100_286. x-axis represents genome positions of chromosome 1. y-axis represents two haplotypes, labeling 1 and 2 here. Top panel showed genotypes generated from low-coverage genome sequencing. Bottom panel showed genotypes generated from RNA-seq.

Supplementary Files

File S5.1. List of identified quantitative trait loci (QTLs).
File S5.2. Summary of variance explained by all QTLs within a trait.
File S5.3. List of 16 traits used for calculating phenotypic variance explained by integrated QTLs.
File S5.4. F1 genotyping results from transcriptomic and genomic sequencing data.

All Supplementary Files are available at https://github.com/guoweier/QTL_manuscript

Chapter 6

Conclusions

Overview of dissertation research

The primary goal of this dissertation was to investigate the origin and effects of genomic structural variation in plants. Using two model systems, *Populus* and *Arabidopsis*, this dissertation addressed the following questions: 1) What processes can potentially trigger chromoanagenesis in plants? How can we identify and characterize chromoanagenesis in plants? 2) What are the phenotypic effects of allelic and dosage variation in *Populus*? Can we increase the power of phenotypic prediction by integrating allelic and dosage variation?

Chromoanagenesis, one of the novel types of structural variations (SV), has been observed to be associated with human cancer [32,35]. A limited number of studies identifying and characterizing chromoanagenesis in plants have been published, probably because of: (a) The lack of significant phenotypic evidence associated with chromoanagenesis in plants; (b) The deficiency of standard methods for characterizing clustered breakpoints on localized regions. Previous studies indicate that centromere deficiency [30] and biolistic transformation [44] can result in chromoanagenesis in plants. This suggests that mutagenesis and genetic engineering may be the potential triggering processes for chromoanagenesis. A *Populus* hybrid population from an interspecific cross between wild-type *P. deltoides* (female) and gamma-irradiated *P. nigra* (male) was investigated and revealed 2 hybrid lines with shattered chromosomes. Illumina short-read sequencing confirmed the presence of chromosomal rearrangement in both lines, and the rearranged chromosomes were inherited from the gamma-affected parent (*P. nigra*). 12 and 14 novel DNA junctions were respectively identified and validated from two *Populus* lines, and these junctions were demonstrated to be highly clustered and associated with the rearrangements. The predicted architecture of rearranged chromosomes based on novel DNA junctions indicates broken genomic segments are randomly restructured into a novel chromosome. Genomic feature enrichment characterization suggests breakpoints are more likely to occur in genic regions. This study demonstrates that gamma irradiation can trigger chromoanagenesis in plants, and the induced extreme rearrangement can be tolerated in *Populus* hybrids.

An *Arabidopsis thaliana* hybrid population with MEIOTIC ASYNAPTIC MUTANT 1 (*asy1*) mutation background was screened and we observed one line with extreme copy number oscillations on chromosome 1. ASY1 gene plays an important role in determining recombination patterns during meiosis, and the mutated protein can lead to mis-segregation in Meiosis 1. We characterized the rearranged chromosome in the *Arabidopsis* line, and identified 260 novel DNA junctions in the shattered region. Specifically, 249 junctions were found to have both ends associated with the CNV clustering region. Breakpoints were proved to significantly enrich in genic regions, which is consistent with the findings in other chromoanagenesis cases in plants [30,38]. Additionally, we demonstrated that the rearranged chromosome resulted from the mis-segregation during Meiosis 1, suggesting that *asy1* mutation is the major cause of chromoanagenesis. This study demonstrated that *asy1*-mediated defective meiosis can be a potential triggering process for chromoanagenesis.

With the two characterizations of chromoanagenesis events in plants, we concluded that chromoanagenesis may occur more frequently in plants than previously expected. Mutagenesis plays an important role in triggering this genome catastrophic event. Plants have a great tolerance to this type of chaos, which turns out to be an attractive system for in-depth investigation of complex genome instability events.

SV has been demonstrated to affect phenotypes in many crop species [6,16,17,51,52,214]. Studies about the effects of SV in forest trees are relatively new. *Populus*, as the model tree plant, has a highly heterozygous genome, and also possesses naturally occurring SV. Both allelic variation and SV in *Populus* plants can affect phenotypic outcomes through the modulation of responsive genes corresponding to traits. Previous studies have identified regulatory genomic loci for quantitative traits in *Populus*, independently using SNPs [64,65,67,68] or induced large-scale indels [61–63] as DNA markers. However, a large portion of phenotypic variance remains unexplained, making it difficult for genomic selection in *Populus* breeding. This is probably because of the interaction among SNPs and SV, since *Populus* carries both allele polymorphism and structural variants. In this study, we independently analyzed the effects of allelic and dosage variation on quantitative traits in *Populus*, and compared the QTLs from two variation types. The analysis was conducted on a previously established *Populus* hybrid population, which have induced large-scale indels across the whole genome and possess abundant phenotypic variation. Our results suggested that allelic and dosage variation have distinct responsive loci to most examined traits such as biomass and leaf morphology. Integration of QTLs from allelic and dosage variation significantly improved the explanation of observed phenotypic variance compared to only allelic or dosage QTLs. To further investigate the interaction between allelic and dosage variation, new methods may be needed. Our findings provide a snapshot of the relationship between allelic and dosage variation as well as their effects on quantitative traits in *Populus*.

The findings of this study enrich our understanding of the origin and effects of structural variations in plants. Characterization of chromoanagenesis in plants demonstrates various plant species can provide a good system for in-depth investigation of this genome chaotic event. The computational approach conducted for identifying clustered breakpoints can be applied for fast prediction of chromoanagenesis. Characterization of the effects of SV and SNPs on quantitative traits in *Populus* preliminarily unravel the relationship between allelic and dosage variation on plant phenotypes. The discoveries and accumulated knowledge of SV can improve our ability to better predict the phenotypes and select elite cultivars in plant breeding.

References

1. Escaramís, G. *et al.* (2015) A decade of structural variants: description, history and methods to detect structural variation. *Brief. Funct. Genomics* 14, 305–314
2. Chiang, C. *et al.* (2017) The impact of structural variation on human gene expression. *Nat. Genet.* 49, 692–699
3. Alonge, M. *et al.* (2020) Major impacts of widespread structural variation on gene expression and crop improvement in tomato. *Cell* in press, 1–17
4. Zhou, Y. *et al.* (2019) The population genetics of structural variants in grapevine domestication. *Nat Plants* 5, 965–979
5. Wang, Y. *et al.* (2015) Copy number variation at the GL7 locus contributes to grain size diversity in rice. *Nat. Genet.* 47, 944–948
6. Díaz, A. *et al.* (2012) Copy number variation affecting the Photoperiod-B1 and Vernalization-A1 genes is associated with altered flowering time in wheat (Triticum aestivum). *PLoS One* 7, e33234
7. Beales, J. *et al.* (2007) A pseudo-response regulator is misexpressed in the photoperiod insensitive Ppd-D1a mutant of wheat (Triticum aestivum L.). *Theor. Appl. Genet.* 115, 721–733
8. Zhu, J. *et al.* (2014) Copy number and haplotype variation at the VRN-A1 and central FR-A2 loci are associated with frost tolerance in hexaploid wheat. *Theor. Appl. Genet.* 127, 1183–1197
9. Li, Y. *et al.* (2012) A tandem segmental duplication (TSD) in green revolution gene Rht-D1b region underlies plant height variation. *New Phytol.* 196, 282–291
10. Xu, K. *et al.* (2006) Sub1A is an ethylene-response-factor-like gene that confers submergence tolerance to rice. *Nature* 442, 705–708
11. Cook, D.E. *et al.* (2012) Copy number variation of multiple genes at Rhg1 mediates nematode resistance in soybean. *Science* 338, 1206–1209
12. Hu, Y. *et al.* (2018) Analysis of Extreme Phenotype Bulk Copy Number Variation (XP-CNV) Identified the Association of rp1 with Resistance to Goss's Wilt of Maize. *Front. Plant Sci.* 9, 110
13. Maron, L.G. *et al.* (2013) Aluminum tolerance in maize is associated with higher MATE1 gene copy number. *Proc. Natl. Acad. Sci. U. S. A.* 110, 5241–5246
14. Sutton, T. *et al.* (2007) Boron-toxicity tolerance in barley arising from efflux transporter amplification. *Science* 318, 1446–1449
15. Cao, J. *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.* 43, 956–963
16. Xu, X. *et al.* (2011) Resequencing 50 accessions of cultivated and wild rice yields markers for identifying agronomically important genes. *Nat. Biotechnol.* 30, 105–111
17. McHale, L.K. *et al.* (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.* 159, 1295–1308
18. Lu, P. *et al.* (2012) Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome Res.* 22, 508–518
19. Saintenac, C. *et al.* (2011) Targeted analysis of nucleotide and copy number variation by exon capture in allotetraploid wheat genome. *Genome Biol.* 12, R88
20. Zmienko, A. *et al.* (2020) AthCNV: A Map of DNA Copy Number Variations in the Arabidopsis Genome. *Plant Cell* 32, 1797–1819
21. Bayer, P.E. *et al.* (2020) Plant pan-genomes are the new reference. *Nat Plants* 6, 914–920
22. Gaines, T.A. *et al.* (2010) Gene amplification confers glyphosate resistance in Amaranthus palmeri. *Proc. Natl. Acad. Sci. U. S. A.* 107, 1029–1034
23. Pearce, S. *et al.* (2011) Molecular characterization of Rht-1 dwarfing genes in hexaploid wheat. *Plant Physiol.* 157, 1820–1831
24. Żmieńko, A. *et al.* (2014) Copy number polymorphism in plant genomes. *Theor. Appl. Genet.* 127, 1–18

25. Lorenz, A.J. *et al.* (2011) Genomic selection in plant breeding: knowledge and prospects. *Adv. Agron.* 110, 77–123
26. Hämälä, T. *et al.* (2021) Genomic structural variants constrain and facilitate adaptation in natural populations of Theobroma cacao, the chocolate tree. *Proc. Natl. Acad. Sci. U. S. A.* 118
27. Weisweiler, M. *et al.* (2022) Structural variants in the barley gene pool: precision and sensitivity to detect them using short-read sequencing and their association with gene expression and phenotypic variation. *Theor. Appl. Genet.* 135, 3511–3529
28. Gaeta, R.T. *et al.* (2013) In vivo modification of a maize engineered minichromosome. *Chromosoma* 122, 221–232
29. Tan, E.H. *et al.* (2023) Establishment and inheritance of minichromosomes from Arabidopsis haploid induction
30. Tan, E.H. *et al.* (2015) Catastrophic chromosomal restructuring during genome elimination in plants. *Elife* 4, 1–16
31. Guo, W. *et al.* (2022) Chromoanagenesis in plants: triggers, mechanisms, and potential impact. *Trends Genet.* DOI: 10.1016/j.tig.2022.08.003
32. Stephens, P.J. *et al.* (2011) Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* 144, 27–40
33. Liu, P. *et al.* (2011) Chromosome catastrophes involve replication mechanisms generating complex genomic rearrangements. *Cell* 146, 889–903
34. Baca, S.C. *et al.* (2013) Punctuated evolution of prostate cancer genomes. *Cell* 153, 666–677
35. Cortés-Ciriano, I. *et al.* (2020) Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat. Genet.* 52, 331–341
36. Umbreit, N.T. *et al.* (2020) Mechanisms generating cancer genome complexity from a single cell division error. *Science* 368
37. Carbonell-Bejerano, P. *et al.* (2017) Catastrophic Unbalanced Genome Rearrangements Cause Somatic Loss of Berry Color in Grapevine. *Plant Physiol.* 175, 786–801
38. Guo, W. *et al.* (2021) Chromoanagenesis from radiation-induced genome damage in *Populus*. *PLoS Genet.* 17, e1009735
39. Guo, W. *et al.* (2022) Chromoanagenesis in the asy1 meiotic mutant of Arabidopsis*bioRxiv*, 2022.04.27.489737
40. Park, J.-S. *et al.* (2020) Genome analysis of tissue culture-derived variations in regenerated Brassica rapa ssp. pekinensis plants using next-generation sequencing. *Horticulture, Environment, and Biotechnology* 61, 549–558
41. Fossi, M. *et al.* (2019) Regeneration of Solanum tuberosum Plants from Protoplasts Induces Widespread Genome Instability. *Plant Physiol.* 180, 78–86
42. Pucker, B. *et al.* (2021) Large scale genomic rearrangements in selected Arabidopsis thaliana T-DNA lines are caused by T-DNA insertion mutagenesis. *BMC Genomics* 22, 599
43. Gernand, D. *et al.* (2007) Tissue culture triggers chromosome alterations, amplification, and transposition of repeat sequences in Allium fistulosum. *Genome* 50, 435–442
44. Liu, J. *et al.* (2019) Genome-Scale Sequence Disruption Following Biolistic Transformation in Rice and Maize. *Plant Cell* 31, 368–383
45. Ellis, B. *et al.* (2010) Why and How *Populus* Became a "Model Tree." In *Genetics and Genomics of Populus* (Jansson, S. et al., eds), pp. 3–14, Springer New York
46. Schreiner, E.J. (1959) *Production of Poplar Timber in Europe and Its Significance and Application in the United States*, U.S. Department of Agriculture, Forest Service
47. Roller, K.J. *et al.* (1984) *A guide to the identification of poplar clones in Ontario*, Maple : Ontario Ministry of Natural Resources
48. Arreghini, R.I. *et al.* (2000) Poplar clones: identification in the nursery. *Poplar clones: identification in the nursery.* at <https://www.cabdirect.org/cabdirect/abstract/20013014213>
49. Li, S.-W. *et al.* (2005) Progress and strategies in cross breeding of poplars in China. *For. Stud. China* 7, 54–60

50. Stanton, B.J. *et al.* (2010) *Populus* Breeding: From the Classical to the Genomic Approach. In *Genetics and Genomics of Populus* (Jansson, S. et al., eds), pp. 309–348, Springer New York
51. Beló, A. *et al.* (2010) Allelic genome structural variations in maize detected by array comparative genome hybridization. *Theor. Appl. Genet.* 120, 355–367
52. Schiessl, S. *et al.* (2017) Targeted deep sequencing of flowering regulators in Brassica napus reveals extensive copy number variation. *Sci Data* 4, 170013
53. Prunier, J. *et al.* (2017) CNVs into the wild: screening the genomes of conifer trees (Picea spp.) reveals fewer gene copy number variations in hybrids and links to adaptation. *BMC Genomics* 18, 97
54. Prunier, J. *et al.* (2017) Gene copy number variations in adaptive evolution: The genomic distribution of gene copy number variations revealed by genetic mapping and their adaptive role in an undomesticated species, white spruce (Picea glauca). *Mol. Ecol.* 26, 5989–6001
55. Porth, I. *et al.* (2023) Structural genomic variations and their effects on phenotypes in *PopulusbioRxiv*, 2023.02.14.528455
56. Pinosio, S. *et al.* (2016) Characterization of the Poplar Pan-Genome by Genome-Wide Identification of Structural Variation. *Mol. Biol. Evol.* 33, 2706–2719
57. Prunier, J. *et al.* (2019) Gene copy number variations involved in balsam poplar (*Populus* balsamifera L.) adaptive variations. *Mol. Ecol.* 28, 1476–1490
58. Zhang, B. *et al.* (2019) The poplar pangenome provides insights into the evolutionary history of the genus. *Commun Biol* 2, 215
59. Goessen, R. *et al.* (2022) Coping with environmental constraints: Geographically divergent adaptive evolution and germination plasticity in the transcontinental *Populus* tremuloides. *Plants People Planet* 4, 638–654
60. Henry, I.M. *et al.* (2015) A system for dosage-based functional genomics in poplar. *Plant Cell* 27, 2370–2383
61. Bastiaanse, H. *et al.* (2019) A comprehensive genomic scan reveals gene dosage balance impacts on quantitative traits in *Populus* trees. *Proc. Natl. Acad. Sci. U. S. A.* 116, 13690–13699
62. Bastiaanse, H. *et al.* (2020) A systems genetics approach to deciphering the effect of dosage variation on leaf morphology in *Populus*. *Plant Cell* at <https://academic.oup.com/plcell/advance-article-abstract/doi/10.1093/plcell/koaa016/6007531>
63. Rodriguez-Zaccaro, F.D. *et al.* (2021) Genetic regulation of vessel morphology in *Populus*. *Front. Plant Sci.* 12, 705596
64. Frewen, B.E. *et al.* (2000) Quantitative trait loci and candidate gene mapping of bud set and bud flush in *populus*. *Genetics* 154, 837–845
65. Rohde, A. *et al.* (2011) Bud set in poplar - genetic dissection of a complex trait in natural and hybrid populations. *New Phytol.* 189, 106–121
66. Carletti, G. *et al.* (2016) QTLs for Woolly Poplar Aphid (Phloeomyzus passerinii L.) Resistance Detected in an Inter-Specific *Populus* deltoides x P. nigra Mapping Population. *PLoS One* 11, e0152569
67. Rae, A.M. *et al.* (2009) Five QTL hotspots for yield in short rotation coppice bioenergy poplar: the Poplar Biomass Loci. *BMC Plant Biol.* 9, 23
68. Xia, W. *et al.* (2018) Construction of a high-density genetic map and its application for leaf shape QTL mapping in poplar. *Planta* 248, 1173–1185
69. Wu, R. *et al.* (1997) Molecular genetics of growth and development in *Populus* (Salicaceae). v. mapping quantitative trait loci affecting leaf variation. *Am. J. Bot.* 84, 143
70. Della Coletta, R. *et al.* (2021) How the pan-genome is changing crop genomics and improvement. *Genome Biol.* 22, 3
71. Alkan, C. *et al.* (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.* 12, 363–376
72. Tuskan, G.A. *et al.* (2006) The genome of black cottonwood, *Populus* trichocarpa (Torr. & Gray). *Science* 313, 1596–1604
73. Weckselblatt, B. and Rudd, M.K. (2015) Human Structural Variation: Mechanisms of Chromosome

Rearrangements. *Trends Genet.* 31, 587–599

74. Amundson, K.R. *et al.* (2020) Genomic Outcomes of Haploid Induction Crosses in Potato (Solanum tuberosum L.). *Genetics* 214, 369–380
75. Di Meo, G.P. *et al.* (2006) Cattle rob(1;29) originating from complex chromosome rearrangements as revealed by both banding and FISH-mapping techniques. *Chromosome Res.* 14, 649–655
76. Ducos, A. *et al.* (2007) Chromosomal control of pig populations in France: 2002–2006 survey. *Genet. Sel. Evol.* 39, 583
77. Pellestor, F. *et al.* (2011) Complex chromosomal rearrangements: origin and meiotic behavior. *Hum. Reprod. Update* 17, 476–494
78. Shen, M.M. (2013) Chromoplexy: a new category of complex rearrangements in the cancer genome*Cancer cell*, 23567–569
79. Holland, A.J. and Cleveland, D.W. (2012) Chromoanagenesis and cancer: mechanisms and consequences of localized, complex chromosomal rearrangements. *Nat. Med.* 18, 1630–1638
80. Pellestor, F. *et al.* (2021) Chromoanagenesis, the mechanisms of a genomic chaos. *Semin. Cell Dev. Biol.* DOI: 10.1016/j.semcdb.2021.01.004
81. Zepeda-Mendoza, C.J. and Morton, C.C. (2019) The Iceberg under Water: Unexplored Complexity of Chromoanagenesis in Congenital Disorders. *Am. J. Hum. Genet.* 104, 565–577
82. Korbel, J.O. and Campbell, P.J. (2013) Criteria for inference of chromothripsis in cancer genomes. *Cell* 152, 1226–1236
83. Ly, P. *et al.* (2017) Selective Y centromere inactivation triggers chromosome shattering in micronuclei and repair by non-homologous end joining. *Nat. Cell Biol.* 19, 68–75
84. Lee, J.A. *et al.* (2007) A DNA Replication Mechanism for Generating Nonrecurrent Rearrangements Associated with Genomic Disorders. *Cell* 131, 1235–1247
85. Hastings, P.J. *et al.* (2009) A microhomology-mediated break-induced replication model for the origin of human copy number variation. *PLoS Genet.* 5
86. Berger, M.F. *et al.* (2011) The genomic complexity of primary human prostate cancer. *Nature* 470, 214–220
87. Haffner, M.C. *et al.* (2010) Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nat. Genet.* 42, 668–675
88. Anderson, N.D. *et al.* (2018) Rearrangement bursts generate canonical gene fusions in bone and soft tissue tumors. *Science* 361
89. Fenech, M. *et al.* (2011) Molecular mechanisms of micronucleus, nucleoplasmic bridge and nuclear bud formation in mammalian and human cells. *Mutagenesis* 26, 125–132
90. Klaasen, S.J. *et al.* (2022) Nuclear chromosome locations dictate segregation error frequencies. *Nature* 607, 604–609
91. Crasta, K. *et al.* (2012) DNA breaks and chromosome pulverization from errors in mitosis. *Nature* 482, 53–58
92. Terradas, M. *et al.* (2010) Genetic activities in micronuclei: is the DNA entrapped in micronuclei lost for the cell? *Mutat. Res.* 705, 60–67
93. Hatch, E.M. *et al.* (2013) Catastrophic nuclear envelope collapse in cancer cell micronuclei. *Cell* 154, 47–60
94. Otsuka, S. *et al.* (2016) Nuclear pore assembly proceeds by an inside-out extrusion of the nuclear envelope. *Elife* 5
95. Denais, C.M. *et al.* (2016) Nuclear envelope rupture and repair during cancer cell migration. *Science* 352, 353–358
96. Liu, S. *et al.* (2018) Nuclear envelope assembly defects link mitotic errors to chromothripsis. *Nature* 561, 551–555
97. Maciejowski, J. *et al.* (2015) Chromothripsis and Kataegis Induced by Telomere Crisis. *Cell* 163, 1641–1654
98. Zhang, C.-Z. *et al.* (2015) Chromothripsis from DNA damage in micronuclei. *Nature* 522, 179–184
99. Kneissig, M. *et al.* (2019) Micronuclei-based model system reveals functional consequences of

chromothripsis in human cells. *Elife* 8

100. Tang, S. *et al.* (2022) Breakage of cytoplasmic chromosomes by pathological DNA base excision repair. *Nature* DOI: 10.1038/s41586-022-04767-1
101. Maciejowski, J. *et al.* (2020) APOBEC3-dependent kataegis and TREX1-driven chromothripsis during telomere crisis. *Nat. Genet.* DOI: 10.1038/s41588-020-0667-5
102. Morishita, M. *et al.* (2016) Chromothripsis-like chromosomal rearrangements induced by ionizing radiation using proton microbeam irradiation system. *Oncotarget* 7, 10182–10192
103. Leibowitz, M.L. *et al.* (2021) Chromothripsis as an on-target consequence of CRISPR–Cas9 genome editing. *Nat. Genet.*
104. Yi, K. and Ju, Y.S. (2018) Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.* 50, 1–11
105. Kuppu, S. *et al.* (2015) Point Mutations in Centromeric Histone Induce Post-zygotic Incompatibility and Uniparental Inheritance. *PLoS Genet.* 11, e1005494
106. Ravi, M. and Chan, S.W.L. (2010) Haploid plants produced by centromere-mediated genome elimination. *Nature* 464, 615–618
107. Marimuthu, M.P.A. *et al.* (2021) Epigenetically mismatched parental centromeres trigger genome elimination in hybrids. *Sci Adv* 7, eabk1151
108. Li, X. *et al.* (2017) Single nucleus sequencing reveals spermatid chromosome fragmentation as a possible cause of maize haploid induction. *Nat. Commun.* 8, 991
109. Caryl, A.P. *et al.* (2000) A homologue of the yeast HOP1 gene is inactivated in the Arabidopsis meiotic mutant asy1. *Chromosoma* 109, 62–71
110. Lambing, C. *et al.* (2020) ASY1 acts as a dosage-dependent antagonist of telomere-led recombination and mediates crossover interference in Arabidopsis. *Proc. Natl. Acad. Sci. U. S. A.* 117, 13647–13658
111. Pochon, G. *et al.* (2022) The Arabidopsis Hop1 homolog ASY1 mediates cross-over assurance and interference*bioRxiv*, 2022.03.17.484635
112. Nacry, P. *et al.* (1998) Major chromosomal rearrangements induced by T-DNA transformation in Arabidopsis. *Genetics* 149, 641–650
113. Yokota, E. *et al.* (2011) Stability of monocentric and dicentric ring minichromosomes in Arabidopsis. *Chromosome Res.* 19, 999–1012
114. Majhi, B.B. *et al.* (2014) A novel T-DNA integration in rice involving two interchromosomal translocations. *Plant Cell Rep.* 33, 929–944
115. Valente, A.S. *et al.* (2018) T-DNA associated reciprocal translocation reveals differential survival of male and female gametes. *Plant Gene* 15, 37–43
116. Gheysen, G. *et al.* (1987) Integration of Agrobacterium tumefaciens transfer DNA (T-DNA) involves rearrangements of target plant DNA sequences. *Proc. Natl. Acad. Sci. U. S. A.* 84, 6169–6173
117. Gang, H. *et al.* (2019) Comprehensive characterization of T-DNA integration induced chromosomal rearrangement in a birch T-DNA mutant. *BMC Genomics* 20, 311
118. Salomon, S. and Puchta, H. (1998) Capture of genomic and T-DNA sequences during double-strand break repair in somatic plant cells. *EMBO J.* 17, 6086–6095
119. Di Gaspero, G. and Foria, S. (2015) 2 - Molecular grapevine breeding techniques. In *Grapevine Breeding Programs for the Wine Industry* (Reynolds, A., ed), pp. 23–37, Woodhead Publishing
120. Lukaszewski, A.J. (1995) Chromatid and chromosome type breakage-fusion-bridge cycles in wheat (Triticum aestivum L.). *Genetics* 140, 1069–1085
121. Zheng, Y.Z. *et al.* (1999) Time course study of the chromosome-type breakage-fusion-bridge cycle in maize. *Genetics* 153, 1435–1444
122. Kato, A. *et al.* (2005) Minichromosomes derived from the B chromosome of maize. *Cytogenet. Genome Res.* 109, 156–165
123. Han, F. *et al.* (2007) Minichromosome analysis of chromosome pairing, disjunction, and sister chromatid cohesion in maize. *Plant Cell* 19, 3853–3863

124. Mandáková, T. *et al.* (2019) Origin and Evolution of Diploid and Allopolyploid Camelina Genomes Were Accompanied by Chromosome Shattering. *Plant Cell* 31, 2596–2612
125. Zhao, Q. *et al.* (2021) Reconstruction of ancestral karyotype illuminates chromosome evolution in the genus Cucumis. *Plant J.* 107, 1243–1259
126. Mandáková, T. and Lysak, M.A. (2018) Post-polyploid diploidization and diversification through dysploid changes. *Curr. Opin. Plant Biol.* 42, 55–65
127. Luo, M.C. *et al.* (2009) Genome comparisons reveal a dominant mechanism of chromosome number reduction in grasses and accelerated genome evolution in Triticeae. *Proc. Natl. Acad. Sci. U. S. A.* 106, 15780–15785
128. Wang, H. and Bennetzen, J.L. (2012) Centromere retention and loss during the descent of maize from a tetraploid ancestor. *Proc. Natl. Acad. Sci. U. S. A.* 109, 21004–21009
129. Hoang, P.T.N. and Schubert, I. (2017) Reconstruction of chromosome rearrangements between the two most ancestral duckweed species Spirodela polyrhiza and S. intermedia. *Chromosoma* 126, 729–739
130. Pinton, A. *et al.* (2009) Influence of sex on the meiotic segregation of at (13; 17) Robertsonian translocation: a case study in the pig. *Hum. Reprod.* 24, 2034–2043
131. Villagómez, D.A.F. *et al.* (2008) Extensive nonhomologous meiotic synapsis between normal chromosome axes of an rcp(3;6)(p14;q21) translocation in a hairless Mexican boar. *Cytogenet. Genome Res.* 120, 112–116
132. Bertelsen, B. *et al.* (2016) A germline chromothripsis event stably segregating in 11 individuals through three generations. *Genet. Med.* 18, 494–500
133. Zeng, D. *et al.* (2020) A transcriptomic view of the ability of nascent hexaploid wheat to tolerate aneuploidy. *BMC Plant Biol.* 20, 97
134. Yuan, Y. *et al.* (2021) Current status of structural variation studies in plants. *Plant Biotechnol. J.* DOI: 10.1111/pbi.13646
135. Comai, L. and Tan, E.H. (2019) Haploid Induction and Genome Instability. *Trends Genet.* 35, 791–803
136. Liu, Y. *et al.* (2020) Rapid Birth or Death of Centromeres on Fragmented Chromosomes in Maize. *Plant Cell* 32, 3113–3123
137. Pellestor, F. (2019) Chromoanagenesis: Cataclysms behind complex chromosomal rearrangements. *Mol. Cytogenet.* 12, 1–12
138. Pellestor, F. and Gatinois, V. (2018) Chromoanasynthesis: Another way for the formation of complex chromosomal abnormalities in human reproduction. *Hum. Reprod.* 33, 1381–1387
139. Pellestor, F. and Gatinois, V. (2020) Chromoanagenesis: a piece of the macroevolution scenario. *Mol. Cytogenet.* 13, 3
140. Marimuthu, M.P.A. *et al.* (2021) Biased removal and loading of centromeric histone H3 during reproduction underlies uniparental genome elimination*bioRxiv*, 2021.02.24.432754
141. Nasuda, S. *et al.* (1998) Gametocidal genes induce chromosome breakage in the interphase prior to the first mitotic cell division of the male gametophyte in wheat. *Genetics* 149, 1115–1124
142. van Harten, A.M. (1998) *Mutation breeding: theory and practical applications*, Cambridge University Press
143. Forment, J.V. *et al.* (2012) Chromothripsis and cancer: Causes and consequences of chromosome shattering. *Nat. Rev. Cancer* 12, 663–670
144. Brunner, H. (1995) Radiation induced mutations for plant selection. *Appl. Radiat. Isot.* 46, 589–594
145. Hase, Y. *et al.* (2018) Physiological status of plant tissue affects the frequency and types of mutations induced by carbon-ion irradiation in Arabidopsis. *Sci. Rep.* 8, 1394
146. Belfield, E.J. *et al.* (2012) Genome-wide analysis of mutations in mutant lineages selected following fast-neutron irradiation mutagenesis of Arabidopsis thaliana. *Genome Res.* 22, 1306–1315
147. Henry, I.M. *et al.* (2014) Efficient Genome-Wide Detection and Cataloging of EMS-Induced Mutations Using Exome Capture and Next-Generation Sequencing. *Plant Cell* 26, 1382–1397
148. Sakamoto, A.N. *et al.* (2017) An ion beam-induced Arabidopsis mutant with marked chromosomal

rearrangement. *J. Radiat. Res.* 58, 772–781

149. Lieber, M.R. (2010) The mechanism of double-strand DNA break repair by the nonhomologous DNA end-joining pathway. *Annu. Rev. Biochem.* 79, 181–211
150. Tsai, A.G. and Lieber, M.R. (2010) Mechanisms of chromosomal rearrangement in the human genome. *BMC Genomics* 11 Suppl 1, S1
151. Przybytkowski, E. *et al.* (2014) Chromosome-breakage genomic instability and chromothripsis in breast cancer. *BMC Genomics* 15, 579
152. Huang, K. and Rieseberg, L.H. (2020) Frequency, Origins, and Evolutionary Role of Chromosomal Inversions in Plants. *Front. Plant Sci.* 11, 296
153. Shoshani, O. *et al.* (2020) Chromothripsis drives the evolution of gene amplification in cancer. *Nature* DOI: 10.1038/s41586-020-03064-z
154. Koo, D.-H. *et al.* (2018) Extrachromosomal circular DNA-based amplification and transmission of herbicide resistance in crop weed Amaranthus palmeri. *Proc. Natl. Acad. Sci. U. S. A.* 115, 3332–3337
155. Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* 25, 1754–1760
156. Henry, I.M. *et al.* (2010) Phenotypic consequences of aneuploidy in Arabidopsis thaliana. *Genetics* 186, 1231–1245
157. Ruby, J.G. *et al.* (2013) PRICE: Software for the targeted assembly of components of (Meta) genomic sequence data. *G3: Genes, Genomes, Genetics* 3, 865–880
158. Camacho, C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10, 421
159. Li, H. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25, 2078–2079
160. Poot, M. and Haaf, T. (2015) Mechanisms of Origin, Phenotypic Effects and Diagnostic Implications of Complex Chromosome Rearrangements. *Mol. Syndromol.* 6, 110–134
161. Anand, R.P. *et al.* (2014) Chromosome rearrangements via template switching between diverged repeated sequences. *Genes Dev.* 28, 2394–2406
162. Blanc-Mathieu, R. *et al.* (2017) Population genomics of picophytoplankton unveils novel chromosome hypervariability. *Sci Adv* 3, e1700239
163. Koltsova, A.S. *et al.* (2019) On the complexity of mechanisms and consequences of chromothripsis: An update. *Front. Genet.* 10, 393
164. Sanchez-Moran, E. *et al.* (2008) ASY1 coordinates early events in the plant meiotic recombination pathway. *Cytogenet. Genome Res.* 120, 302–312
165. Ross, K.J. *et al.* (1997) Cytological characterization of four meiotic mutants of Arabidopsis isolated from T-DNA-transformed lines. *Chromosome Res.* 5, 551–559
166. Wei, F. and Zhang, G.-S. (2010) Meiotically asynapsis-induced aneuploidy in autopolyploid Arabidopsis thaliana. *J. Plant Res.* 123, 87–95
167. Ferdous, M. *et al.* (2012) Inter-homolog crossing-over and synapsis in Arabidopsis meiosis are dependent on the chromosome axis protein AtASY3. *PLoS Genet.* 8, e1002507
168. Sequeira-Mendes, J. *et al.* (2014) The Functional Topography of the Arabidopsis Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States. *Plant Cell* 26, 2351–2366
169. Kloosterman, W.P. *et al.* (2011) Chromothripsis as a mechanism driving complex de novo structural rearrangements in the germline. *Hum. Mol. Genet.* 20, 1916–1924
170. Ly, P. *et al.* (2019) Chromosome segregation errors generate a diverse spectrum of simple and complex genomic rearrangements. *Nat. Genet.* 51, 705–715
171. Papathanasiou, S. *et al.* (2021) Whole chromosome loss and genomic instability in mouse embryos after CRISPR-Cas9 genome editing. *Nat. Commun.* 12, 5855
172. Chiang, C. *et al.* (2012) Complex reorganization and predominant non-homologous repair following chromosomal breakage in karyotypically balanced germline rearrangements and transgenic integration. *Nat. Genet.* 44, 390–397
173. Pellestor, F. *et al.* (2014) Chromothripsis: potential origin in gametogenesis and preimplantation cell

divisions. A review. *Fertil. Steril.* 102, 1785–1796

174. Cuacos, M. *et al.* (2021) Meiotic chromosome axis remodelling is critical for meiotic recombination in Brassica rapa. *J. Exp. Bot.* 72, 3012–3027
175. Yuan, W. *et al.* (2009) Mutation of the rice gene PAIR3 results in lack of bivalent formation in meiosis. *Plant J.* 59, 303–315
176. Golubovskaya, I. *et al.* (1992) Effects of several meiotic mutations on female meiosis in maize. *Dev. Genet.* 13, 411–424
177. Cao, L. *et al.* (2021) The Inactivation of Arabidopsis UBC22 Results in Abnormal Chromosome Segregation in Female Meiosis, but Not in Male Meiosis. *Plants* 10
178. Huang, X. and Han, B. (2014) Natural variations and genome-wide association studies in crop plants. *Annu. Rev. Plant Biol.* 65, 531–551
179. Alonso-Blanco, C. *et al.* (2009) What has natural variation taught us about plant development, physiology, and adaptation? *Plant Cell* 21, 1877–1896
180. Huang, X. *et al.* (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proc. Natl. Acad. Sci. U. S. A.* 108, 4488–4493
181. Alonso-Blanco, C. *et al.* (1999) Natural allelic variation at seed size loci in relation to other life history traits of Arabidopsis thaliana. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4710–4717
182. Zhang, S. *et al.* (2021) Natural allelic variation in a modulator of auxin homeostasis improves grain yield and nitrogen use efficiency in rice. *Plant Cell* 33, 566–580
183. Todesco, M. *et al.* (2010) Natural allelic variation underlying a major fitness trade-off in Arabidopsis thaliana. *Nature* 465, 632–636
184. Duan, Z. *et al.* (2022) Natural allelic variation of GmST05 controlling seed size and quality in soybean. *Plant Biotechnol. J.* 20, 1807–1818
185. Satbhai, S.B. *et al.* (2017) Natural allelic variation of FRO2 modulates Arabidopsis root growth under iron deficiency. *Nat. Commun.* 8, 15603
186. Jin, J.-Q. *et al.* (2016) Natural allelic variations of TCS1 play a crucial role in caffeine biosynthesis of tea plant and its related species. *Plant Physiol. Biochem.* 100, 18–26
187. Huang, X. *et al.* (2012) Epistatic natural allelic variation reveals a function of AGAMOUS-LIKE6 in axillary bud formation in Arabidopsis. *Plant Cell* 24, 2364–2379
188. Fisher, R.A. (1919) XV.—The Correlation between Relatives on the Supposition of Mendelian Inheritance. *Earth Environ. Sci. Trans. R. Soc. Edinb.* 52, 399–433
189. Gupta, P.K. *et al.* (2008) Array-based high-throughput DNA markers for crop improvement. *Heredity* 101, 5–18
190. Davey, J.W. *et al.* (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. *Nat. Rev. Genet.* 12, 499–510
191. Elshire, R.J. *et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS One* 6, e19379
192. McMullen, M.D. (2003) Quantitative trait locus analysis as a gene discovery tool. *Methods Mol. Biol.* 236, 141–154
193. Rafalski, J.A. (2010) Association genetics in crop improvement. *Curr. Opin. Plant Biol.* 13, 174–180
194. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661–678
195. Jamann, T.M. *et al.* (2015) QTL mapping using high-throughput sequencing. *Methods Mol. Biol.* 1284, 257–285
196. Yan, H. *et al.* (2023) Pangenomic analysis identifies structural variation associated with heat tolerance in pearl millet. *Nat. Genet.*
197. Golicz, A.A. *et al.* (2016) The pangenome of an agronomically important crop plant Brassica oleracea. *Nat. Commun.* 7, 13390
198. Birchler, J.A. and Veitia, R.A. (2012) Gene balance hypothesis: Connecting issues of dosage sensitivity across biological disciplines. *Proc. Natl. Acad. Sci. U. S. A.* 109, 14746–14753
199. Veitia, R.A. *et al.* (2013) Gene dosage effects: nonlinearities, genetic interactions, and dosage

compensation. *Trends Genet.* 29, 385–393

200. Birchler, J.A. and Veitia, R.A. (2010) The gene balance hypothesis: implications for gene regulation, quantitative traits and evolution. *New Phytol.* 186, 54–62
201. Howie, B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.* 5, e1000529
202. Williams, A.L. *et al.* (2012) Phasing of many thousands of genotyped samples. *Am. J. Hum. Genet.* 91, 238–251
203. Hager, P. *et al.* (2020) SmartPhase: Accurate and fast phasing of heterozygous variant pairs for genetic diagnosis of rare diseases. *PLoS Comput. Biol.* 16, e1007613
204. Martin, M. *et al.* (2016) WhatsHap: fast and accurate read-based phasing*bioRxiv*, 085050
205. Brem, R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science* 296, 752–755
206. Li, Y.I. *et al.* (2016) RNA splicing is a primary link between genetic variation and disease. *Science* 352, 600–604
207. Lukens, L.N. and Doebley, J. (1999) Epistatic and environmental interactions for quantitative trait loci involved in maize evolution. *Genet. Res.* 74, 291–302
208. Frary, A. *et al.* (2000) fw2.2: a quantitative trait locus key to the evolution of tomato fruit size. *Science* 289, 85–88
209. Birchler, J.A. *et al.* (1990) Analysis of autosomal dosage compensation involving the alcohol dehydrogenase locus in Drosophila melanogaster. *Genetics* 124, 679–686
210. Birchler, J.A. and Yang, H. (2022) The multiple fates of gene duplications: Deletion, hypofunctionalization, subfunctionalization, neofunctionalization, dosage balance constraints, and neutral variation. *Plant Cell* 34, 2466–2474
211. Birchler, J.A. and Veitia, R.A. (2021) One Hundred Years of Gene Balance: How Stoichiometric Issues Affect Gene Expression, Genome Evolution, and Quantitative Traits. *Cytogenet. Genome Res.* 161, 529–550
212. Zinkgraf, M. *et al.* (2016) Creation and Genomic Analysis of Irradiation Hybrids in *Populus*. *Curr Protoc Plant Biol* 1, 431–450
213. Doerge, R.W. and Churchill, G.A. (1996) Permutation tests for multiple loci affecting a quantitative character. *Genetics* 142, 285–294
214. Yu, P. *et al.* (2011) Detection of copy number variations in rice using array-based comparative genomic hybridization. *BMC Genomics* 12, 372