The plasticity of plant genomes

Causes and consequences: a survey of data on structural genome variation in plants

CGM 2020-04

ONDERZOEKSRAPPORT

The plasticity of plant genomes

Causes and consequences: a survey of data on structural genome variation in plants

Ruud A. de Maagd, Clemens van de Wiel, and Henk. J. Schouten

Wageningen Plant Research, Wageningen, The Netherlands

SUPERVISORY COMMITTEE

Prof. dr. Johan Memelink (Chair), Universiteit Leiden, member of COGEM

Dr. Jan Kooter, Vrije Universiteit Amsterdam, member of COGEM

Dr. Marcel Proveniers, Universiteit Utrecht, external member

Dr. Cynthia van Rijn, GMO office, National Institute of Public Health and the Environment

Dr. Maria Koster, COGEM secretariat

**Disclaimer:

This report was commissioned by COGEM. The content of this publication is the sole responsibility of the author(s) and does not necessarily reflect the views of COGEM.

Preface

The EU has been debating for years, decennia almost, whether plants manufactured by new breeding techniques should be exempted from GMO regulations. In an effort to advance decisions about this matter, in 2017 the Netherlands introduced a discussion paper in the European debate. This paper proposes that GM plants manufactured with new breeding techniques that are assessed as being at least as safe as plants obtained by traditional breeding should be exempted from GMO regulations. COGEM issued in 2019 an advice on this paper at the request of the Ministry of Infrastructure and Water Management. The advice addresses among others so-called intragenic GM plants, which are modified exclusively with DNA sequences from plant species with which the target plant species can also exchange genetic material using traditional breeding methods. However, other than in cisgenesis, in intragenesis genetic elements, such as for example a protein coding region, can be recombined with regulatory signals such as a promoter from other genes, provided they are from crossable relatives. In the advice, COGEM stated that it is not known whether the risks of intragenic GM plants are comparable with the risks of traditionally bred plants. Although it is known that new combinations of genomic sequences, known as structural variations, may also arise naturally, it is not clear to everyone to what extent or how frequently these changes occur.

To gain insight into the extent and frequency of structural changes occurring naturally in the genome, COGEM has commissioned a literature study on genome plasticity. The goal was to obtain an overview of current knowledge about genome plasticity and on the time scale wherein structural genome variations occur. The idea was that the study can contribute to defining a baseline for risk analysis of plants genetically modified by intragenesis.

The resulting report starts with an overview of natural mechanisms whereby structural variations in genomes can arise, and then continues with examples especially from crop plants where structural variations that naturally occurred gave rise to desirable phenotypes. Finally the report attempts to estimate the relative frequency of different types of structural variation.

The report notes that most of the existing structural variation is still hidden because the sequencing techniques until very recently generate data that allow detection of structural variation only to a limited extent. Looking at the mechanisms it becomes clear that transposons are responsible for a large amount of structural variation. The report also notes that traditional breeding using mutagenesis, which is exempted from European GMO legislation, can lead to highly increased structural variation within one generation, especially when ionizing radiation is used as the mutagen. No further spoilers here, it is up to the reader of the report to decide whether structural variations introduced by intragenesis fall within the range of structural variations found "naturally" in traditional breeding (supported by mutagenesis).

Prof. dr. J. Memelink, Chairman Advisory Committee

Contents

Samenvatting

De afgelopen jaren is de kennis over genoomplasticiteit enorm toegenomen. Om deze reden had de COGEM de behoefte aan een actuele inventarisatie van spontane genoomveranderingen in eukaryoten met de nadruk op planten, en met name de structurele veranderingen die optreden tijdens de conventionele veredeling en teelt van gewassen. Deze informatie is relevant voor risicobeoordeling van "intragene" gewassen, waar nieuwe combinaties van genetische elementen van de gastheer (of kruisbare verwanten) worden geïntroduceerd. De centrale vraag is of dergelijke nieuwe combinaties ook spontaan kunnen ontstaan in conventionele veredeling, inclusief conventionele mutagenese, en met welke frequentie. Dit draagt bij tot het vaststellen van een "basislijn" waarmee intragene planten of de creatie van nieuwe structurele variatie in het algemeen, kan worden vergeleken bij de risicobeoordeling. Het rapport richt zich op "Structurele Variatie" (SV) in genomen: verschillen tussen genomen van individuen van dezelfde (of kruisbare) soort, groter dan 50 basenparen. Dit omvat deleties (verwijderingen), inserties (invoegingen, inclusief transposons), duplicaties, inversies, translocaties en complexe herschikkingen.

Voor dit rapport hebben we twee informatiebronnen gebruikt: literatuuronderzoek en interviews met professionals in plantenveredelingsbedrijven over hun ervaring met SV in hun werk. Dat laatste leidde tot de conclusie dat SV waarschijnlijk een rol speelt in hun veredelingswerk maar dat ook veel structurele veranderingen waarschijnlijk nog onopgemerkt zijn. De voorbeelden die ze bereid waren te delen, waren meestal al in de literatuur beschreven.

Dit rapport beschrijft eerst de detectiemethoden voor SV's. Genoombrede detectie van SV is afhankelijk van drie technieken: cytogenetica (microscopie), hele-genoom (her)sequencing en de analyse ervan, en op hybridisatie gebaseerde technieken. Helaas lijden deze methoden nog steeds aan hoge vals-positieve en vals-negatieve frequenties en lage resolutie. Dit betekent dat zelfs als SV wordt ontdekt, de exacte aard en het effect op genen meestal nog grotendeels onbekend is. Dat zou een precieze sequentie van het SV-breekpunt of de rand vereisen. Gelukkig beginnen de opkomende sequencing-technologieën voor aaneengesloten grote DNA fragmenten deze beperkingen al op te heffen, en zullen dit in de nabije toekomst nog meer doen. Verder heeft een gedetailleerde analyse van bepaalde fenotypische eigenschappen al een beter inzicht gegeven in de structurele veranderingen die deze eigenschappen veroorzaken.

De processen zelf die leiden tot waargenomen SV worden meestal niet gezien. Er zijn echter veel modellen voorgesteld, die waargenomen SV achteraf kunnen verklaren. We beschrijven een reeks van dergelijke modellen. De voorgestelde mechanismen die leiden tot SV kunnen grofweg worden ingedeeld in:

1. "Kopieerfouten" tijdens DNA-replicatie van delende cellen.
2. Dubbelstrengs DNA-breuken, gevolgd door reparatie. Dit kan leiden tot deleties, inversies en het opnemen van "opvullend" DNA in de gerepareerde breuken.
3. Ontstaan bij recombinatie. Tijdens meiose vindt vaak recombinatie plaats tussen allelsequenties van chromosoomparen. Dit kan leiden to duplicatie, deletie, of translocatie.
4. De activiteit van transposons (TE's). Transposons zijn DNA-sequenties die in genomen kunnen bewegen door replicatie en insertie ("kopiëren en plakken") of door excisie en re-integratie ("knippen en plakken").

Deze studie had als doel bij te dragen aan een degelijke basis voor een milieurisicobeoordeling van intragene gewassen. Daarom hebben we ons gericht op gedocumenteerde SV's tussen kruisbare planten. Deze hier gerapporteerde vergelijkingen moeten worden onderscheiden van vergelijkingen tussen verwante soorten en geslachten, die veranderingen op een meer evolutionaire tijdsschaal beschrijven. Verschillende voorbeelden van SV's binnen soorten, leidend tot waarneembare fenotypes of belangrijke eigenschappen, worden beschreven.

Met betrekking tot SV die ontstaat tijdens conventionele veredeling, bespreken we ook variatie veroorzaakt door "conventionele mutagenese" zoals door chemicaliën of straling. Uit de beperkte beschikbare (kwantitatieve) literatuur blijkt dat met name snelle neutronen of zware ionenbestraling SV veroorzaakt,

meestal grote deleties, inversies en translocaties. Gecombineerd met het verschijnsel dat bij herstel van DNA breuken soms opvullend DNA opgenomen wordt, en de neiging tot translocatie, verhoogt dit de frequentie van het creëren van nieuwe combinaties van genetische elementen aanzienlijk. Verder bekijken we mogelijke effecten van weefselkweek op SV, maar in vergelijking met conventionele mutagenese is er van weefselkweek weinig bekend over impact op de creatie van nieuwe SV.

De hoeveelheid beschikbare literatuur neemt sterk af als het gaat om rapporteren van spontane structurele veranderingen gedurende een of enkele generaties, zowel in planten als in andere eukaryoten, inclusief mensen. Dit is gedeeltelijk te wijten aan beperkingen van huidige sequentietechnologieën en bio-informatica, d.w.z. het vermogen om SV's betrouwbaar te identificeren en te karakteriseren, evenals aan hun lage frequentie. Detectiemethoden kunnen binnenkort verbeteren. Mutatie accumulatielijnen, in het bijzonder van planten met een korte levenscyclus en kleine genomen zoals Arabidopsis, zijn bijzonder geschikt voor het verzamelen van gegevens over laagfrequente gebeurtenissen. Een voorbeeld van Arabidopsis nakomelingen over 5 generaties wordt besproken. Dit onderzoek aan afstammelingen van mutatie accumulatielijnen identificeerde specifiek grote deleties en enkele kortere inserties. In geen van deze experimenten werd de vorming van unieke combinaties van sequenties beoordeeld.

Hoewel de hoeveelheid informatie over spontaan optredende SV tijdens de teelt of plantenveredeling beperkt was, laten we verschillende opvallende voorbeelden zien van structurele veranderingen in appel, druif en tomaat, die hebben geleid tot uitgesproken en vaak commercieel aantrekkelijke fenotypische kenmerken.

We hebben verschillende soorten SV gesorteerd in waarschijnlijke volgorde van afnemende frequenties, d.w.z. 1. Uitwisseling van genetische elementen binnen een cluster van tandemherhalingen, zoals resulterend uit ongelijke cross-overs en andere niet-allele homologe recombinaties; 2. Verwijderingen, en het inbrengen van opvullend DNA tijdens reparatie van DNA breuken; 3. Translocaties van transposons inclusief liftend gastheer-DNA; 4. Omgekeerde herhalingen van DNA, leidend tot RNAinterferentie-achtige structuren; 5. Translocaties van niet-TE's; 6. Inversies; 7. Complexe herschikkingen in intragene voorbeeldplanten. Deze lijst is niet volledig en de exacte volgorde kan worden betwist. In het geval van conventionele veredeling zonder mutagenese, beschouwen we de frequenties van type 1 tot 3 voldoende hoog om op te treden tijdens traditionele veredeling door kruising en teelt van gewassen. Wanneer conventionele mutagenese wordt opgenomen, kunnen typen 4 tot 6 ook worden beschouwd als vrij frequent voorkomend tijdens conventioneel veredelen. We hebben twee voorbeelden gegeven van intragene gewassen. De kans op spontaan ontstaan van dergelijke gebeurtenissen binnen een of enkele generaties is bijzonder klein.

Summary

During recent years knowledge on genome plasticity has vastly increased. For this reason, COGEM has expressed the need for an up-to-date inventory of spontaneous genome changes in eukaryotes with an emphasis on plants, and in particular, the structural changes that occur during conventional breeding and cultivation of crops. This information is relevant for risk assessment of "intragenic" crops, where new combinations of genetic elements from the host (or crossable) species may be introduced. The central question is whether such new combinations also can arise spontaneously in conventional breeding, including conventional mutagenesis, and at what frequency. This contributes to establishing a "baseline" against which "intragenesis," or the creation of new structural variation in general, will be compared in risk assessment. The report focusses on "Structural Variation" (SV) in genomes: differences between genomes of individuals of the same (or crossable) species, larger than 50 base pairs. This includes deletions, insertions (including transposable elements), duplications, inversions, translocations, and complex rearrangements.

For this report, we used two sources of information: a literature search, and interviews with professionals in plant breeding companies about their experience with SV in their work. The latter led to the conclusion that SV probably plays a (yet mostly unknown) role in their breeding work, and the examples they were willing to share were mainly already described in the literature. This report first describes the methods of detection of SVs. Genome-wide detection of SV depends on three techniques: cytology (microscopy), whole-genome (re-) sequencing and its analysis, and hybridisation-based techniques. Unfortunately, these methods still suffer from high false-positive and false-negative rates and low resolution, meaning that even if SV is discovered, its exact nature and effect on genes is usually still mostly unknown. This would require precise sequencing of the SV 'breakpoint' or border. Fortunately, the arising long-read sequencing technologies are already starting to remove these limitations and will do more so soon. Further, a detailed analysis of specific phenotypic traits has provided already more precise insight into the structural variation causing these traits.

The processes themselves that lead to observed SV usually are not witnessed. However, many models have been proposed, which may explain observed SV in retrospect. We describe a series of such models. The suggested mechanisms leading to SV can be broadly categorised as:

1. "Copying errors" during DNA replication of multiplying cells.
2. Double-strand DNA breaks, followed by repair. This can lead to deletions, inversions, and in the inclusion of "filler" DNA into the repaired breaks.
3. Recombination-based. During meiosis, recombination frequently occurs between allelic sequences of pairing chromosomes. This can lead to duplication, deletion, and translocation.
4. The activity of transposable elements. Transposable Elements or transposons are DNA sequences that can move through genomes by replication and insertion ("copy and paste") or by excision and reintegration ("cut and paste").

This study aimed at contributing to a foundation for a sound biosafety assessment of intragenic crops. Therefore, we have focussed on documented SVs between crossable plants. These comparisons reported here should be distinguished from comparisons between related species and genera, which describe changes on a more evolutionary timescale. Several examples of intraspecies SVs, leading to discernible phenotypes or important traits, are described.

Concerning SV that arises during conventional breeding, we also discuss variation caused by "conventional mutagenesis" such as by chemicals or radiation. From the limited available literature, it was concluded that in particular, fast-neutron or heavy-ion irradiation causes SV, mostly large deletions, inversions, and translocations. Combined with the capacity of double-strand break repair to incorporate filler DNA, and the propensity for translocation, this increases the frequency of creating new combinations of genomic DNA considerably. Further, we review possible effects of tissue culture on SV, but compared to conventional mutagenesis, little is known about tissue culture's impact on the creation of new SV.

The amount of available literature decreases sharply when it comes to reports of spontaneous structural changes over one or a few generations, both in plants as well as in other eukaryotes, including humans. This is partly due to constraints of current sequence technologies and bioinformatics, i.e. the ability to reliably identify and characterise SVs, as well as to their low frequency. Detection methods may improve soon. Mutation accumulation lines, specifically from fast cycling plants with small genomes such as Arabidopsis, are particularly suitable for collecting data on low-frequency events. One example of Arabidopsis siblings over 5 generations is discussed. In that study, resequencing of descendants from mutation accumulation lines specifically identified large deletions and some shorter insertions. The creation of unique combinations of genome elements was assessed in none of these experiments.

Although the amount of information on spontaneously arising SV during cultivation or plant breeding was limited, we provide several striking examples of structural changes in apple, grape, and tomato, that led to pronounced and often commercially attractive phenotypic traits.

We sorted different types of SV in likely order of decreasing frequencies, i.e. 1. Exchange of genetic elements within a cluster of tandem repeats, such as resulting from unequal crossovers and other non-allelic homologous recombinations; 2. Deletions, including the insertion of filler DNA during repair; 3. Translocations of transposable elements (TEs), including hitch-hiking host DNA; 4. Inverted repeats, leading to RNAi-like structures; 5. Translocations of non-TEs; 6. Inversions; 7. Complex rearrangements in intragenic example plants. This list is not exhaustive, and the exact order can be disputed. In the case of conventional breeding without mutagenesis, we regard the frequencies of types 1 to 3 sufficiently high for occurring during traditional breeding by crossing and cultivation of crops. When including conventional mutagenesis, types 4 to 6 can also be regarded as occurring rather frequently during conventional breeding.

We provided two examples of intragenic crops. The likelihood of spontaneous creation of such events within one or a few breeding generations is very low.

1 Introduction

During recent years our knowledge on genome plasticity has vastly increased, also regarding rearrangements in genomes, such as large deletions, translocations, and inversions. Structural variations (SVs) are the largest source of genetic variation, in the sense that more base pairs are involved than there are with single nucleotide polymorphisms (SNPs) (Wellenreuther *et al.*, 2019). Depending on whether the amount of DNA remains the same, they can be divided into unbalanced SVs, i.e. copy number variations (CNVs) and insertions or deletions, and balanced SVs, i.e. inversions and translocations (MacDonald *et al.*, 2014).

COGEM has expressed the need for an up-to-date inventory of spontaneous genome changes in eukaryotes, including such changes that occur during conventional breeding of crops. This information will be relevant for environmental risk assessment of "intragenics" in crops, where new combinations of sequences from the same or a crossable species may be introduced.

1.1 Genetic modification and intragenesis

Three types of genetic modifications of crops involving the insertion of recombinant DNA in the final product can be broadly distinguished:

1. **Transgenesis**, where recombinant DNA including DNA from other species than the host species is inserted and remains in the host genome, usually at a random position.
2. **Cisgenesis**, as in transgenesis, but inserting (also at a random position) only recombinant DNA originating from the host species, or from a species that can be naturally crossed with the host. The insert is an identical copy of the host's native gene, including its regulatory sequences, such as promoter, introns, and terminator.
3. **Intragenesis,** as in cisgenesis inserting (also at a random position) only DNA sequences from the host species or crossable species. However, in intragenic plants, new combinations of genetic elements may be made, such as a unique combination of a promoter and a coding sequence.

The conceptual difference between cisgenesis and intragenesis is further clarified in Fig. 1.

Recently the term "subgenic" has been coined for events originating from mutagenesis using site-directed nucleases such as CRISPR/Cas9, where no DNA is inserted, but host DNA is deleted (Sticklen, 2015).

We use the wording "new combinations" of genetic elements of the host as follows:

- combinations of coding sequences from different genes into one recombinant coding sequence
- new combinations of regulatory sequences such as promoters with coding sequences,
- both of the above
- combinations of parts of genes or non-coding sequences not resulting in an expressed coding sequence but designed to affect the expression of endogenous genes (such as artificial microRNAs, RNAi constructs, etc.)

1.2 Intragenic organisms and the 'baseline'

Intragenic plants contain only genetic elements from the breeders' gene pool. In intragenesis, also named intragenics, desired traits can be obtained by newly combining elements such as promoters, coding sequences and terminators of different genes within the gene pool of the conventional breeder. In this report, the gene produced by combining different elements is referred to as an intragene. Intragenesis, therefore, offers considerably more options for modifying gene expression and for trait development compared to cisgenesis. In cisgenesis, only entire genes from the breeder's gene pool are used, including their native regulatory sequences and native terminators. Intragenesis can also include hairpin gene

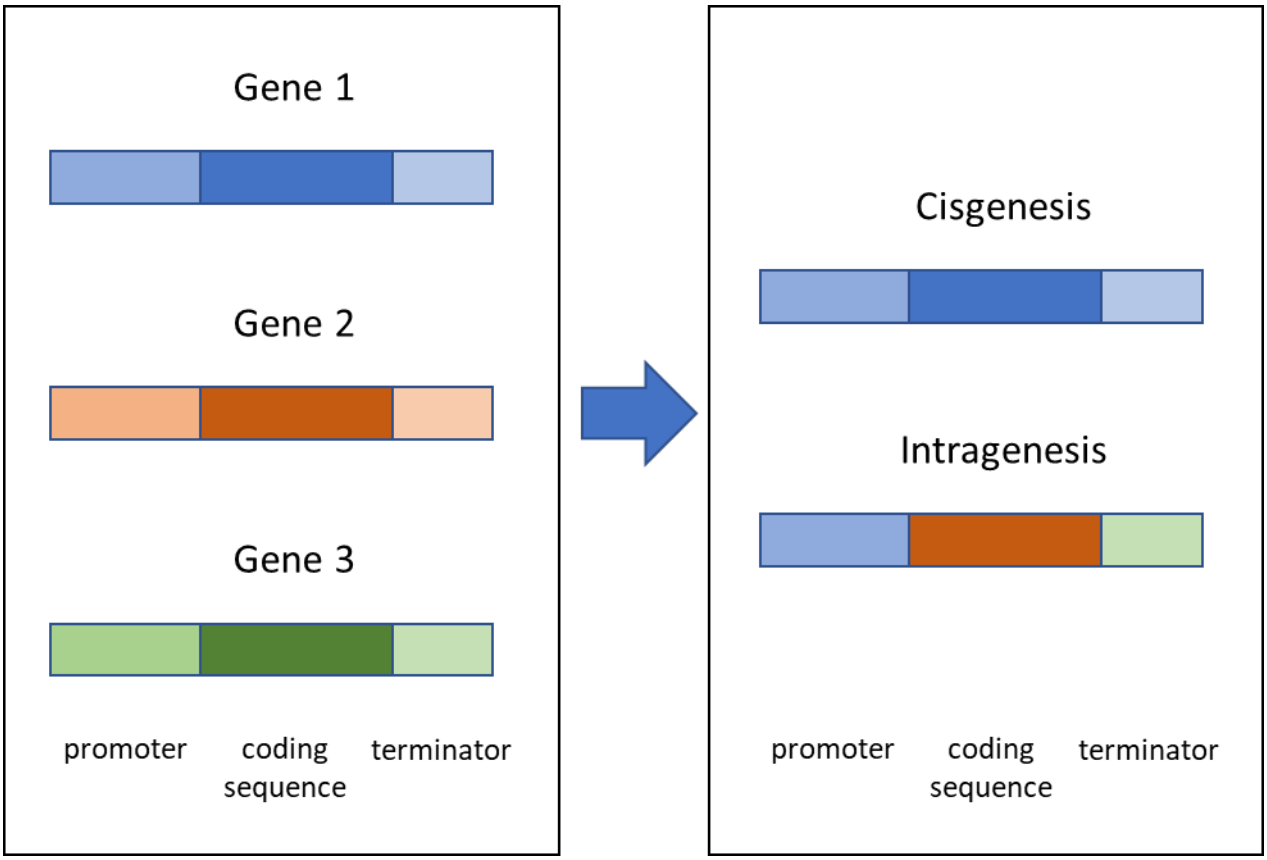

Figure 1. Cisgenesis and intragenesis. *Both methods take existing sequences from the host (left), but where cisgenesis only uses contiguous sequences without rearrangement, intragenesis can combine 'elements' from different host genes or sequences. Promoter, coding sequence, and terminator are used here as examples for 'elements' but this may be any DNA sequence from the host.*

silencing approaches, similar to RNAi (EFSA Panel on Genetically Modified Organisms (GMO), 2012). Intragenic plants do not contain foreign DNA.

A central question in this study is: Can such new combinations also arise spontaneously in conventional breeding, including in conventional mutagenesis, and with which likelihood? This study aims at establishing a "baseline" for genome plasticity, for answering this question. The baseline for structural variation is here defined as the top of the natural range of rearrangements in genomes when comparing sexually compatible organisms, and the genomic rearrangements that can occur spontaneously during conventional breeding and in nature. Whether this is actually a 'line' or rather a 'zone' or a 'bandwidth' is up to the reader to determine after reading this report.

If a genetic modification or gene editing leads to plants that could also be found in nature or could be obtained by conventional breeding, the resulting plants will remain below the baseline of natural variation and traditional breeding. If a genetic modification or gene editing does not lead to novel phenotypes compared to possible phenotypes in nature or breeding, the biosafety risks of that modification will be most probable below the baseline of conventional breeding. However, if new genetic elements are added, or if new combinations of genetic elements are made that cannot occur in nature, cultivation or breeding, or occur with an extremely low likelihood, this may trigger additional risk assessment.

This report focusses on and provides observations and data for discussing the question: To which extent could intragenic-like organisms also arise from breeding or in nature, and which types are improbable to result from these?

To which extent could intragenic-like organisms also arise from conventional breeding or in nature?

1.3 Structural variation

In this report, we will explicitly distinguish SV:

- between related and crossable species (crops and their wild relatives), which has developed at an evolutionary timescale
- between accessions of one crop species, which may have arisen anywhere from tens to thousands of years since domestication, and
- that arose spontaneously within one or a few generations, during breeding history

By far, most of our current crop species belong to the clade of angiosperms or flowering plants. This is the most diverse group of land plants (approximately 300,000 species) which split from other seed-bearing plants (gymnosperms) about 200 million years ago. They are characterised by, among other features, the enclosure of the seeds in a fruit. Since their origin, as well as in the seed plants before that, the evolution of angiosperms is characterised by Whole Genome Duplications (WGDs), rendering plants' progeny polyploid (having more than the usual diploid -2 copies of each chromosome- equivalent of a genome). This had become apparent in the past 20 years when more and more plant genomes were fully sequenced, and the evolutionary history of their genomes could be traced back from their sequence differences (Soltis *et al.*, 2015; Soltis and Soltis, 2016; Salse, 2016; Murat *et al.*, 2017). Whole-genome duplications and segmental duplications (duplications of just a part of a chromosome) generate extra copies of genes, which can be deleterious and therefore disappear rapidly (fractionation). Alternatively, they can persist, be redundant and allow for the evolution of new functions by the divergence of their DNA sequence through accumulating mutations (neofunctionalisation) (see also section 5.8). Thus it is estimated that 65% of plant genes have a duplicate copy (Panchy *et al.*, 2016). Together with chromosome fissions and fusions, whole-genome duplications and fractionation have shaped extant plant genomes since their last common ancestor (Murat *et al.*, 2017). Along with some of the mechanisms described in this report, these processes have created new combinations of coding and regulatory sequences, some of which are species- or lineage-specific and together create diversity. Therefore, structural variation (SV) of genomes has been key to the evolution of plant species.

SV that originated during evolution has been detected frequently in wild germplasm and have been introduced by plant breeders into modern varieties during introgression of new traits. This structural variation has been studied, aided by advanced long-read sequencing technologies. SV identification by whole-genome resequencing of many accessions from one crop species is becoming more commonplace. SV that occurs spontaneously at a more 'daily basis', e.g. during commercial seed propagation or cultivation or in nature, is more difficult to detect. This is probably due to its low frequency, and because the vast majority of structural changes does not lead to a discernible phenotype with a commercial benefit. If offspring with an undesired phenotypic trait is found at a low frequency, that offspring is discarded during breeding, rather than studied in detail. The cause of the phenotypic deviation is then usually unknown. The rate of induction of SV using conventional mutagenesis will be higher, as will be discussed later in this report.

However, with the increasing availability, decreased costs, and the resulting increase of high-throughput genome-wide (re)sequencing of eukaryotic genomes, including long-range sequencing, it has become more and more apparent that these genomes do display plasticity. This plasticity occurs not only on evolutionary timescales, as was already inferred from **interspecies** comparisons of genomes. It also occurs on smaller timescales within a species, as can be inferred from **intraspecies** comparisons. Several new structural variants that spontaneously arose recently in plant varieties have been discovered and studied because these phenotypic traits have been appealing from a commercial or scientific point of view.

1.4 Types of structural variation

Genome plasticity has been described as the alterations in the structure of the genome that allows organisms to adapt to changes in the environment, occupy new niches, for pathogens to coevolve with

their hosts, or to cope with polyploidisation. Genome plasticity can be described using the underlying phenomenon of SV, consisting of many kinds of variations within the genome of one species (Fig. 2).

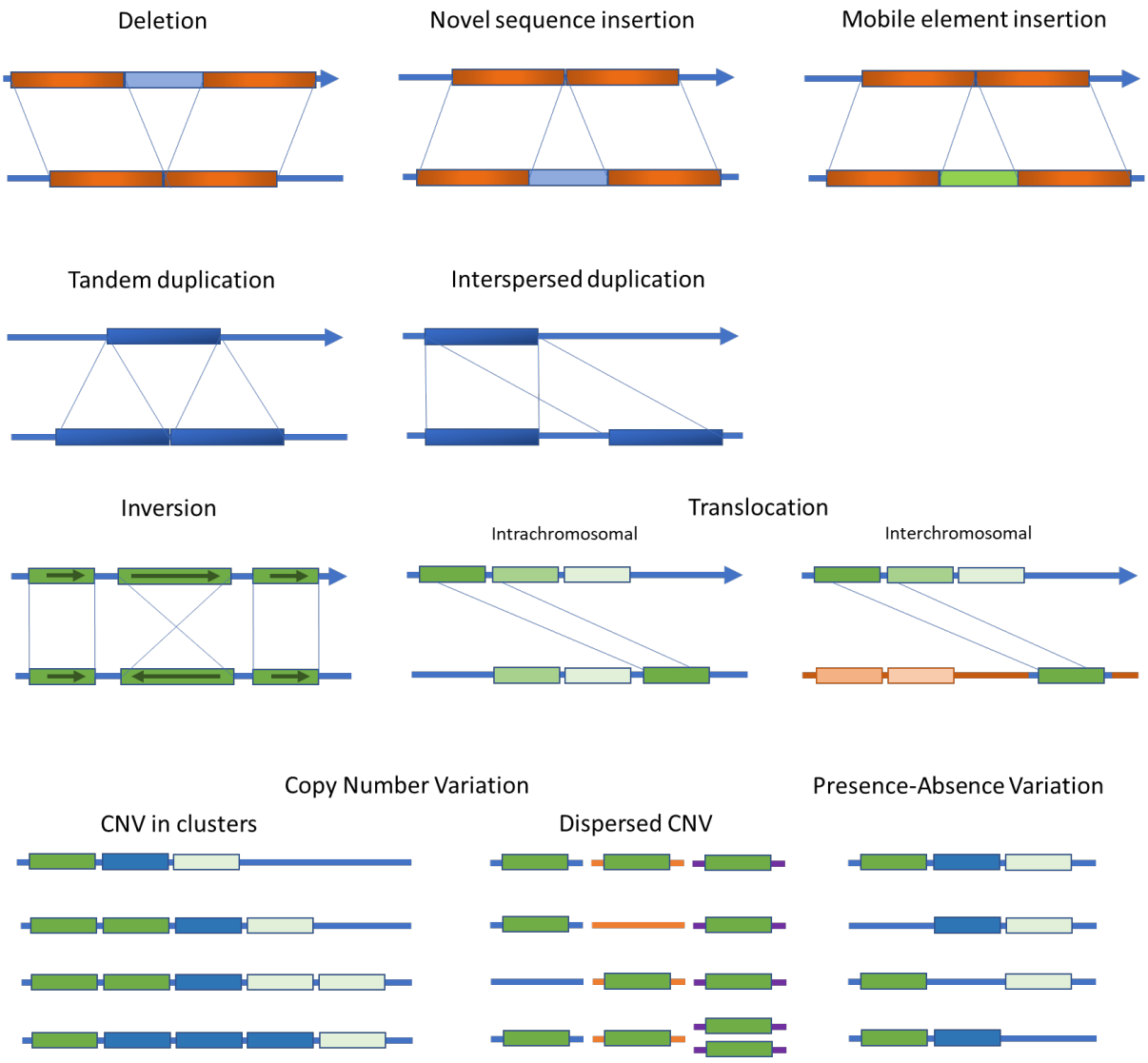

Figure 2. The major types of Structural Variation and Copy Number Variation. *The picture depicts deletions, novel sequence insertions, mobile-element insertions, tandem and interspersed segmental duplications, inversions, and translocations in a test genome (lower line) compared to a reference genome (top arrow). Parts are adapted from* (Alkan *et al.*, 2011)*. Copy number variation (CNV) refers to differences in the frequency of occurrence of a particular sequence in the genome. This may often refer to differences in the numbers of members of a gene cluster (lower left), or of repeated sequences that are dispersed over the genome, like transposons (lower centre). Presence-Absence Variation is a distinct type of SV that only distinguishes whether a sequence is present or absent when comparing genomes.*

SVs refer to insertions/deletions, inversions, translocations and copy number variations (CNVs). SVs were initially defined as genomic variations that involve segments of DNA larger than 1 kb in length (Feuk, Marshall, *et al.*, 2006). More recently, due to advances in discovery methods, its spectrum has been widened to include much smaller events. Therefore, >50 bp is currently used to distinguish SVs from the smaller "indels" (insertions and deletions of less than 50 bp) (Alkan *et al.*, 2011). In contrast, CNVs have been defined as limited to events >1kb. Variation in lengths of short sequence repeats (SSRs such as dinucleotide, trinucleotide repeats, e.g. ATATATAT, ATG ATG ATG ATG) are abundant, and are not covered by 'copy number variation'. We considered this variation in the number of short sequence-repeats out of the scope of this report. Presence/Absence Variation (PAV) is defined as a subset of CNVs. These are

sequences that are present in one genome, but missing in another (and thus can be considered a specific insertion or deletion in one genome relative to another, Fig. 2). This is used to set it apart from a broader definition of CNVs that consists of sequences that are present in all genomes, but with different copy number (Springer *et al.*, 2009).

In this report, we have collected current knowledge on genome plasticity and Structural Variation, particularly of the intraspecies kind, focussed on but not limited to plants and several crops. Related organisms that belong to the same family but are not sexually compatible, also display a lot of structural variation, as has become apparent when comparing *de novo* assemblies of such related organisms. However, we regard that SV less relevant for the baseline for intragenics. Therefore we focussed on SV between sexually compatible plants, looking especially for structural changes that have arisen recently during breeding and cultivation. For this purpose, we have performed an extensive literature search. The results of this search are presented after in this introductory chapter and further detailed in the rest of this report. This leads to an up-to-date description of the types of structural variations that occur in crop plant species, their effects, and their frequency (as far as is determined). We conclude this report with a reflection to the question which structural changes do belong to the baseline, and which structural changes are more likely beyond the baseline.

1.5 Two examples of intragenic plant varieties

As this report was written with an eye on intragenic crops, we here describe two examples. Both instances have already reached or are close to commercialisation. Therefore, these are not just academic examples or 'proofs of principle', but examples that were evaluated for biosafety because of introduction into the market and the food chain.

1.5.1 Example 1: Intragenic tomatoes for resistance to viruses

The University of Queensland in Australia developed a series of six intragenic virus-resistant tomato lines. The company Nexgen Plants Pty Ltd, also from Queensland, is currently trying to commercialise these lines in the USA. In an inquiry letter, dated February 6th 2019, Nexgen Plants asked APHIS in the USA to confirm that the six intragenic lines are not covered by the APHIS mandate for GMO regulation (Schenk and Hervé, 2019). In case of such a confirmation, these intragenic tomato lines can be imported into the USA without the need of a deregulation procedure, as far as APHIS is concerned. It should be noted that other agencies (FDA, EPA) in the United States are also involved in the approval of GMOs.

In this inquiry letter, Nexgen Plants mentioned that the six lines were developed "using an intragenic method that produces events that mimic natural recombination processes." According to Nexgen Plants, the virus-resistant tomatoes do not harbour any foreign DNA sequences. As the promoter, a native tomato sequence was used, i.e. the promoter from the *Actin* gene located on chromosome 3 of tomato. The terminator was taken from the *Rubisco* gene on chromosome 2 of tomato. The heart of the construct was composed of tomato-derived siRNA DNA sequences, that should provide resistance to two viruses (CMV and TSWV). A series of ≥20nt native DNA fragments from the tomato genome sequence were selected, based on homology to the RNA sequences of the two viruses. For CMV, three fragments were selected, and for TSWV four fragments. The overall homology of the tomato fragments to the CMV RNA genome was 86%, and to the other 75%. Subsequently, per anti-viral construct, an inverted repeat was constructed by putting the same fragment in reverse orientation behind the first fragment (Fig. 3). In between these fragments, a tomato derived DNA sequence was inserted. Upon transcription this produces a so-called hairpin structure in the RNA, that is characteristic for RNA interference (RNAi) constructs. Plants recognise such hairpin structures and process these into short interfering RNAs (siRNAs). siRNAs silence complementary gene transcripts sequences, in this case, viral sequences. This should lead to (partial) resistance to the targeted viruses.

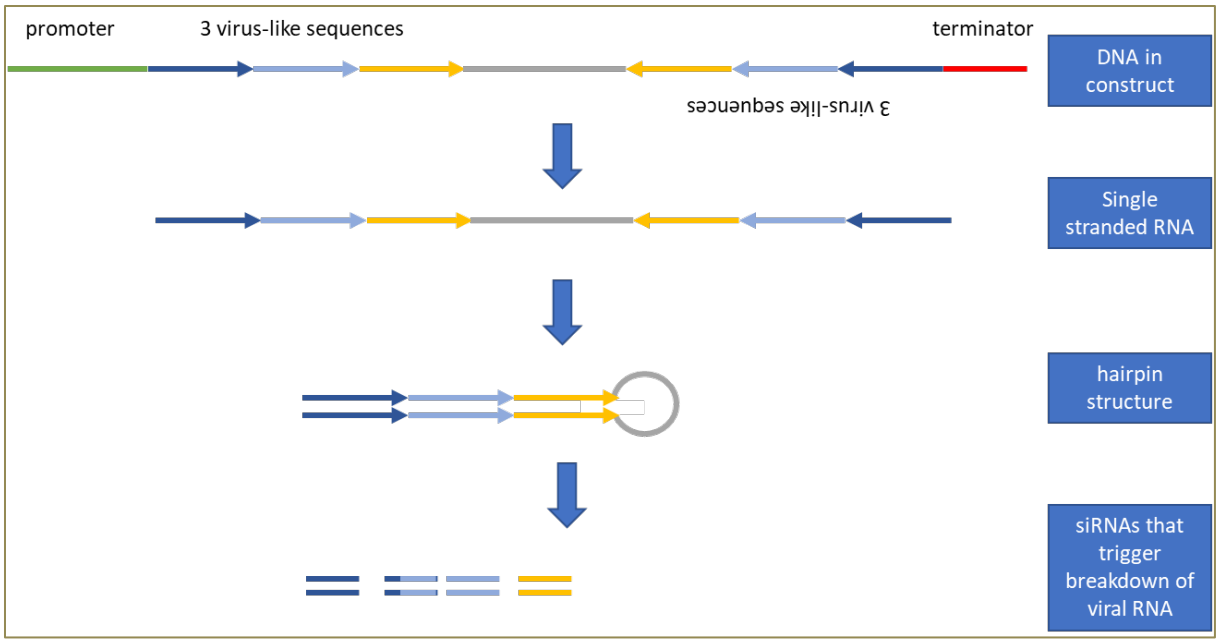

***Figure 3**. **A schematic representation of the production of an intragenic RNAi fragment** by Queensland university and Nexgen Plants Pty, aiming at resistance to viruses. Fragments of only tomato DNA of 20 bp or longer were used. Several virus-like sequences from the tomato genome were concatenated. A copy of this was ligated in reverse orientation, leading to a predicted hairpin (stem-loop) structure in the transcript. Processing of such a hairpin structure by the plant leads to the generation of small interfering RNA (siRNA), that induces the degradation of similar expressed sequences in the plant, in this case, viral sequences* (Schenk and Nowak, 2017).

The constructs were introduced into plant cells by gold particle bombardment. Using particle bombardment rather than the more efficient *Agrobacterium tumefaciens*-mediated transformation, prevented insertion of any *A. tumefaciens* (so foreign) DNA.

The USDA responded on July 2nd 2019 to the above-mentioned question regarding the legal status of these plants, that the intragenic tomato lines are not regarded as plant pests, nor that the construct would lead to a noxious weed. Therefore, the tomato lines are not under APHIS regulation (Firko, 2019). However, USDA also mentioned that these lines may still be subject to other regulatory authorities such as FDA or EPA.

1.5.2 Example 2: Innate™ potatoes

Five years earlier (March 2014), J.R. Simplot Company in the USA had a similar question regarding intragenic potatoes. Simplot did not name their potatoes intragenic, but Innate™. "Simplot's Innate™ technologies allow researchers to isolate genetic elements from any plant genome, rearrange them, or link them together in desired permutations, and introduce them back into the genome (JR Simplot Company, 2015)." Rather than breeding for new potato varieties through crossing, they decided to use the popular 'Russet Burbank' potato variety as the starting point, and improve this variety through Innate™ technologies. They added resistance to the devastating late blight disease caused by the pathogen *Phytopthora infestans*, using the resistance gene *Rpi-vnt1* from a sexually compatible wild potato to this variety. Simultaneously, they added a construct for lowering reducing sugar content, aiming at the reduction of enzymatic browning and discolouration. Further, they added genetic elements for low acrylamide and reduced black spot. For silencing causal genes of these unwanted chemical compounds and discolouration, they applied RNAi, broadly similar to as described for the intragenic virus resistant tomatoes from Australia, although the latter used 'scrambled' tomato genome sequences as opposed to existing target sequences here. The gene construct that was introduced into 'Russet Burbank' was composed of 22 genetic elements, all originating from potato (*Solanum tuberosum*) or wild potato (*S. venturii*) (JR Simplot Company, 2015).

In contrast to the virus-resistant tomatoes from Australia, the potatoes from Simplot had to be evaluated by the USDA for biosafety. In September 2015, one and a half year after the request from Simplot, the USDA allowed commercialisation of the intragenic 'Russet Burbank' in the USA (ISAAA, 2015).

1.6 Overview of this report

In this report, we first describe the applied methods and used sources of information. In the next chapter, we describe mechanisms that can lead to structural variation (SV). This is followed by a series of examples of SV between genotypes belonging to the same species (intraspecies SV). These examples refer to SV that has arisen before modern cultivation and breeding. Not only examples for plants are provided, but also for other eukaryotes. Because the activity of transposable elements (TEs) is a major source of SV, a separate chapter is devoted to TEs. Chapter 7 describes cases of structural changes in crops that occurred during the last century, during cultivation or breeding, and lead to novel phenotypes. As conventional mutagenesis and tissue culture belong to the baseline of conventional breeding, we describe in Chapter 8 effects of these processes on SV. In Chapter 9 we try to provide indications for how frequent spontaneous structural changes occur. These chapters culminate into the last chapter (10): a description of types of structural changes in crops in order of decreasing frequencies, and whether these structural changes are below or beyond the baseline of conventional plant breeding.

2 Methods

2.1 Literature search strategy

We searched literature databases (CAB Abstracts, Google Scholar, Scopus) for reports on structural variation, not only in plants but also in other eukaryotes, as agreed upon with COGEM.

For exploring the literature, we tried several keywords in different combinations: Genome structural variation (SV), genome rearrangement, genome reorganization, genome plasticity, copy number variation (CNV), nonallelic homologous recombination, gene conversion, gene fractionation, duplications, insertions, inversions, inverted repeat, pack-mule, transgenerational, translocations, transposable elements, transposons, retrotransposition.

General terms were not very productive. "Genome structural variation" turned out not really useful, while "genome plasticity" appeared to be biased towards the specialised structural variation evident in fungi and unicellular parasites (discussed in section 6.7) and "genome reorganisation" towards polyploids. More specific terms produced large numbers of publications. Even with more targeted combinations, these numbers could still be substantial. For instance, "Genom* AND (duplication OR insertion OR inversion OR translocat*)" delivered 17508 publications in the CAB Abstracts database, when limited by "Review", still 591, and when by "plant" 11093 or by "plant breeding" 9330. The use of "insertion" led to a lot of papers on genome editing (CRISPR), likely due to "insertion of the CRISP-Cas construct". Keywords could lead to quite many irrelevant documents: translocation may also refer to movements of sugars, RNA, viruses etc.; insertion also applies to molecules (proteins) into membranes. Along the way, other keywords turned up as interesting, such as copy number variation, non-allelic recombination or gene conversion, but these also produced large numbers of papers. Limitation to specific model organisms coming to mind (yeast, Drosophila, human, Arabidopsis) could help but could be limiting the acquired information too stringently. Therefore, we started with recent reviews that were relevant for structural variation and its quantification. As far as possible from these, we further elaborated on relevant examples, such as SVs underlying phenotypes important to breeding. In particular, cases showing similarities to intragenesis, and specific subjects pertinent to breeding and SV, such as mutagenesis and tissue culture effects were further studied. We also used our expertise in topics as resistance gene variability and transposable elements effects. Reassuringly, we came across the same topics and examples repeatedly.

2.2 Interviews with breeding companies

As the baseline of structural variation deals primarily with genomic rearrangements that occur 'on a daily basis', plant breeding companies may be useful sources of information.

Therefore, apart from searching the scientific literature, we also interviewed people working at five different plant breeding companies (BASF Vegetable Seeds, Rijk Zwaan, ENZA Seeds, Syngenta, Managerial Genetics Consulting). We asked them for examples of structural variation in plant genomes in their breeding experience, focussing on the structural variation that arose during the breeding process itself. We mainly talked with pre-breeders and employees of breeding companies who use crop genomes, and select for desired traits and thus have vast knowledge about (deviating) phenotypes, but appeared to be less aware of structural variation at the genome level.

Not all companies were willing to share information. Some of them mentioned that they have encountered more structural variation in plant genomes than they could reveal in the interviews, and only examples that do not harm the commercial interests of the respective companies were shared. Actually, they only mentioned examples that had been described in the literature or had been presented by them during conferences. These examples will be discussed in the following chapters.

However, it was also mentioned that the far majority of the structural variation that arises during the breeding probably has escaped their attention. Reasons for being unnoticed are:

1. If SV does not lead to an apparent phenotype, it will usually not be noticed

2. Only new mutations, including SVs that have a clear commercial interest, will be studied by companies. SVs that lead to an undesired phenotype will be not considered, but discarded, even without knowing whether the phenotype was a result of new structural variation, or because of any other reason

The general consensus is that more SV will be noticed in the coming years, thanks to emerging long-range sequencing technology.

3 Detection of structural variation (SV)

3.1 Introduction

In this chapter, we will discuss methods to detect and quantify structural variations in eukaryotic genomes. This includes limitations of detection methods, despite fast developments in DNA sequencing technology and analysis.

3.2 Methods for detecting SV

3.2.1 Cytogenetics and karyotyping

Before the availability of genome sequencing technologies, microscopy for the visualisation of chromosomes was the only available method for studying structural variation in genomes. Cytology and karyotyping of plant cells originate in the 1920s and '30s. Barbara McClintock was one of the first to use it in the characterisation of the duplications and fusions of the short arm of maize chromosome 9, which was caused by chromosome breaks at what later turned out to be the *Ds* locus ((McClintock, 1942), see further on in this report). Chromosome staining (banding) further improved visualisation and identification of individual chromosomes and the characterisation of structural variation larger than 3 Mb (Feuk, Carson, *et al.*, 2006). This allowed the detection of large translocations and inversions but not, for example, copy number variation. In the 1970s this was complemented by Fluorescence *In Situ* Hybridisation (FISH). This technique uses one or several fluorescently labelled DNA probes to localise, in more detail, and determine the order of genomic sequences on chromosomes with fluorescence microscopy (Ferguson-Smith, 2015).

3.2.2 Whole-genome sequencing

Whole-genome (re)sequencing (WGS) has created the possibility of quantifying the occurrence of structural variation in genomes. Resequencing of individuals, plant lines, varieties or accessions enables obtaining an overview of (structural) variation within a (crop) species when a genome has been assembled as a reference of high quality. This has led to the development of the so-called pan-genome approach (Khan *et al.*, 2019; Gao *et al.*, 2019), which comprises the complete and non-redundant set of genes of a particular species. This set consists of core and variable genes, with the variable genome being mostly related to genes involved in responses to physical and chemical (abiotic) stresses and to biotic stresses, i.e. to pathogens and pest organisms.

However, even with whole-genome sequencing, structural variants are often hard to detect. Until fairly recently, *de novo* sequencing depended on the generation of short DNA reads that were assembled into scaffolds and finally entire chromosomes. This is particularly difficult for repetitive regions because the individual reads often cannot be unambiguously mapped to one of many similar regions or genes. The scaffolds are subsequently anchored to genetic maps, using DNA markers. For anchoring scaffolds, one DNA-marker per scaffold would suffice. To additionally place the scaffold in the correct orientation, at least two DNA markers are required. However, in genetic mapping, closely linked DNA markers can easily be swapped. Thus, the orientation of a scaffold is not always correct. When two *de novo* assemblies are compared, this can easily lead to false-positive identification of inversions, due to assembling errors rather than to the inversions being real. Similarly, false-positive identification of translocations may occur.

When a *de novo* genome assembly of high quality is available, this may be used as the reference genome for re-sequencing of closely related genotypes. Usually, short paired-end Illumina sequence reads are generated and aligned individually to the reference genome. However, aligning to a reference genome, per definition, ignores structural variation between the re-sequenced genotype and the reference. Only close observations of places where the paired-end reads do not align to the reference genome together and nearby ('broken read pairs' or 'discordant reads') may provide some indications of an underlying structural variant. This requires specialised algorithms for inferring SVs from these broken read pairs. This task is

especially challenging for heterozygous genomes (with two alignment options for each read) or polyploid organisms (cotton, potato, strawberry, wheat, oat, leek, apple, chrysanthemum), where 2 or more related genomes are merged in a single genome while still retaining chromosome identity. At least 50-70% of land plants are estimated to be polyploid.

Balanced SVs (those that do not change the total amount of sequence present in a genome), like inversions, may be hard to detect because they will only be clearly recognisable when their breakpoints (ends of the inversion) are found in the alignments. With unbalanced SVs (involving the deletion or addition of sequence), the number of (short) reads for a particular genomic region relative to the average read coverage for the whole genome will provide information on its level of duplication or deletion. Different algorithms for detection of SV sometimes give poorly overlapping results from the same sequencing data (Zmieńko *et al.*, 2014). The false discovery rate can be quite high (more than 10%), even with improved algorithms. For instance, (Z., Zhang *et al.*, 2015) found a false discovery rate of 5.2% for deletions and 9.4% for insertions using PCR analysis. Therefore, there appears to be a need to carefully check data with dedicated PCR tests or long-read sequencing (Z., Zhang *et al.*, 2015).

More recent long-read technologies, including optical mapping, and optimised assembling algorithms have improved matters. For example, in this way, 85% of indels and CNVs averaging 500 bp were shown to be missed by short-read methods (Acuna-Hidalgo *et al.*, 2016). Even then, inversions with highly similar inverted repeats (IRs) at the breakpoints and more significant repetitive regions remained challenging to assess. Thus, an estimated ~308 Mbp (10% of the genome) was inaccessible in a survey of humans by (Audano *et al.*, 2019). By combining several new technologies (Chaisson *et al.*, 2019) determined a 3-7 fold increase in SV identification in human DNA compared to standard high-throughput (HTP) sequencing, including that from the 1000 Genomes Project. Still, dedicated tests may be necessary. An HTP genotyping method for the difficult inversions with highly similar inverted repeats at their breakpoints, using breakpoint sequence information for humans has been reported (Giner-Delgado *et al.*, 2019). Also, in plants, detection problems have been addressed, including breakpoint identification methods (Dolatabadian *et al.*, 2017).

For assessing the complex variation in highly tandem-repeated resistance gene clusters (section 5.2.2 below), a particular enrichment method, RenSeq, was used in combination with long-read technology by (Van de Weyer *et al.*, 2019). Even then, resolving the complete orders of genes in these complex clusters may still be hard to achieve.

3.2.3 Comparative genome hybridisation

Alternatively, array-based comparative genome hybridisation (CGH) can be used for SV detection (Zmieńko *et al.*, 2014). CGH uses DNA probes which together represent an entire species' genome, distributed on an array onto which a sample genome can be hybridised. Hybridised sample DNA can be detected quantitatively for each probe, and so, copy numbers of a particular sequence (probe) can be inferred from the signal strength. Finding any signal depends on the presence of individual sequences but also on the degree of homology to the probes. Therefore, there are biases in this method: lower signals may have more causes than simply deletions compared to the reference genome on the array and sequences in the tested sample that are missing on the array will not be detected at all. Moreover, inversions and translocations may be overlooked.

Currently, most literature is still based on short-read sequencing, hiding the majority of SVs. However, the upcoming long-range sequencing and optical mapping methods will reveal far more SV in the coming years

3.3 Conclusions

In conclusion, most literature is still based on short-read sequencing or comparative genome hybridisation, which may not reveal the majority of SVs. Although long-read (>10 kb) sequencing still has limitations, such as a high rate of small errors, it is promising for obtaining more insight into the nature and frequencies of occurrence of structural variations among genomes. Therefore, the insight in structural variation is likely to increase rapidly in the coming years.

4 Structural variation and gene regulatory effects caused by transposable elements

4.1 Introduction

A variety of mechanisms is underlying structural variation. One of the most important involves transposable elements (TEs). Because TEs represent a significant path to the creation of new combinations of regulatory elements and genes, they warrant a separate discussion, which is the subject of this chapter. Other models for SV creation based on replication errors, and based on repair of double-strand breaks (DSBs) are discussed in the next chapter.

Transposable elements or transposons are relatively autonomously acting DNA sequences that can move through genomes from their location and insert into a new genome location. TEs constitute a large part of all sequenced eukaryote genomes, and this is true for plant genomes as well. Although many of these TEs are inactive, evidence for a role in causing structural variation and phenotypic changes comes from two lines of evidence:

1. From a comparison of genome-wide (re-) sequencing data between individuals (as evidence for activity after divergence from a common ancestor)
2. From the molecular characterisation of relatively recently occurred mutations underlying visible or relevant phenotypes (as evidence for movement in the recent past).

As the number of *de novo* assembled and re-sequenced genomes is increasing, our understanding of the role of TEs in creating structural variation in plant genomes is growing. As will be shown in this chapter, TE activity is detectable in most organisms, although with variable frequency, and the activity is responsible for a vast amount of structural variation in crops as well as in other eukaryotes. Moreover, TE insertions account for a large number of visible and important crop (including mutant) phenotypes, and TE activity is an important mechanism in plant genome evolution and in the creation of new combinations of elements and of genes. There is, however, a large gap between our good knowledge of TE involvement in individual mutant phenotypes and in differences between species at the one hand, and the rather poor knowledge of TE involvement in a generation-to-generation variation on the other hand. Similar to the situation for the detection of other types of SV between generations, this is due to the lack of detailed information from sequencing experiments.

4.2 Classification of TEs and their insertion effects

4.2.1 Types of Transposable Elements

Plant transposable elements can be divided into Class I elements or retrotransposons, which are the most common type, Class II elements (DNA transposons), and Helitrons. Retrotransposons transpose via a "copy-and-paste" mechanism in which an mRNA is produced from an internal open reading frame, which is reverse-transcribed into a cDNA and then integrated into a new position in the genome by an integrase. They can be further divided into LTR (Long Terminal Repeat) retroelements (predominant in plants) and non-LTR elements, and can be autonomous or non-autonomous (the latter requiring enzymes encoded by related autonomous elements). In plants, the most common nonautonomous elements of this kind are the short interspersed nuclear elements (SINEs). Class II transposons move by a "cut-and-paste" mechanism involving a TE-encoded transposase. They also can occur

In some species or particular varieties, such as of maize, TEs move frequently enough so that their mobility can be observed in a small number of generations

as autonomous and non-autonomous elements (often the majority), containing only the inverted terminal repeats, such as the miniature inverted-repeat TEs (MITEs). Finally, Helitrons are thought to transpose via a 'rolling circle' mechanism. This mechanism often involves the capture of flanking host sequences (Lisch, 2013a; Kapitonov and Jurka, 2007).

4.2.2 Occurrence and identification of active TEs

TEs occupy a large part of all genomes. Like in other plant species, in maize, where TEs were discovered by Barbara McClintock (McClintock, 1984), 60-70% of the genome is comprised of LTR retrotransposons, although the less common DNA transposons are more active. Despite the propensity for being silenced, and to degenerate over time, many TE families are still active in diverse species. To limit any detrimental effects on the genome, TEs are generally maintained in a transcriptionally inactive state by methylation and as a result, rarely or never move. Activation of TEs may occur in specific developmental stages or as a result of triggering factors such as stress. In some species or particular varieties, such as in maize, TEs move frequently enough so that their mobility can be observed in a small number of generations. In other cases, transposon mobility can be inferred from transposon insertion polymorphisms (TIPs) between varieties that were identified by genome-resequencing using dedicated algorithms for detection of this variation. Active TEs in any species are defined according to five criteria (in increasing order of importance):

1. Identical terminal repeat sequences and intact internal open reading frames ("intact copies"), all prerequisite for proper activity. This can be inferred from high-quality reference genome sequences
2. Active transcription, either constitutive or induced by stress factors. This can be done by unbiased or dedicated (to a specific transcript) expression analysis
3. Insertion polymorphisms occurring between individuals of the same variety (information about these is rare) or between different varieties or accessions of one species (the majority of available data). This has so far mostly been done using short-read resequencing data. Increasingly this will come from more sophisticated long-read methods. The frequency of activity can only be estimated if it is known how long ago two or more individuals diverged. Within one crop species, this may be any time from a few years and more often, hundreds or thousands of years (the time of domestication).
4. A very recent development is the sequencing of extrachromosomal intermediates of transposition (linear of circular DNA, not part of a chromosome) with homology to transposons: the 'mobilome'
5. Observing TE activity from generation to generation, as was done originally in McClintock's experiments with maize.

In this overview, we have only considered examples that meet the last three criteria. Thus TE activity and its timescale can be inferred from comparison of common and different occurrences between species (evolutionary timescale, millions of years or more) and between individuals of the same species (more recent timescale or currently active TEs), where a difference is assumed to have occurred since the diversion from a common ancestor.

4.2.3 Mechanisms of TE insertion and excision effects

The effects of TE insertions on (nearby) gene functions are summarized in Fig. 4. Insertion into the coding sequence of a gene most often inactivates the gene's function (Fig. 4A), while insertion in regulatory sequences of a gene like silencers or enhancers can change its expression (Fig. 4B), while also the TE itself can introduce regulatory elements like enhancer and silencers (Fig. 4C) or alternative mRNA-splice sites (Fig. 4D) and novel promoters with their transcription start site (Fig. 4E). LTR-retrotransposons, class II DNA transposons and Helitrons can all incorporate and mobilise (parts of) nearby genes (Fig. 4F), which will be discussed in more detail below and in Fig. 5. Finally, insertion downstream of a gene can produce an antisense transcript that negatively regulates the preceding gene (Fig. 4G, and silencing of the inserted transposon may spread to nearby genes, silencing them as well (Fig. 4H) (Lisch, 2013a; Lisch, 2013b).

Class II DNA transposons like the well studied *Ac/Ds* system in maize move by excision and re-insertion. In the process of excision, a DNA double-strand break occurs, which is usually repaired by Non-Homologous

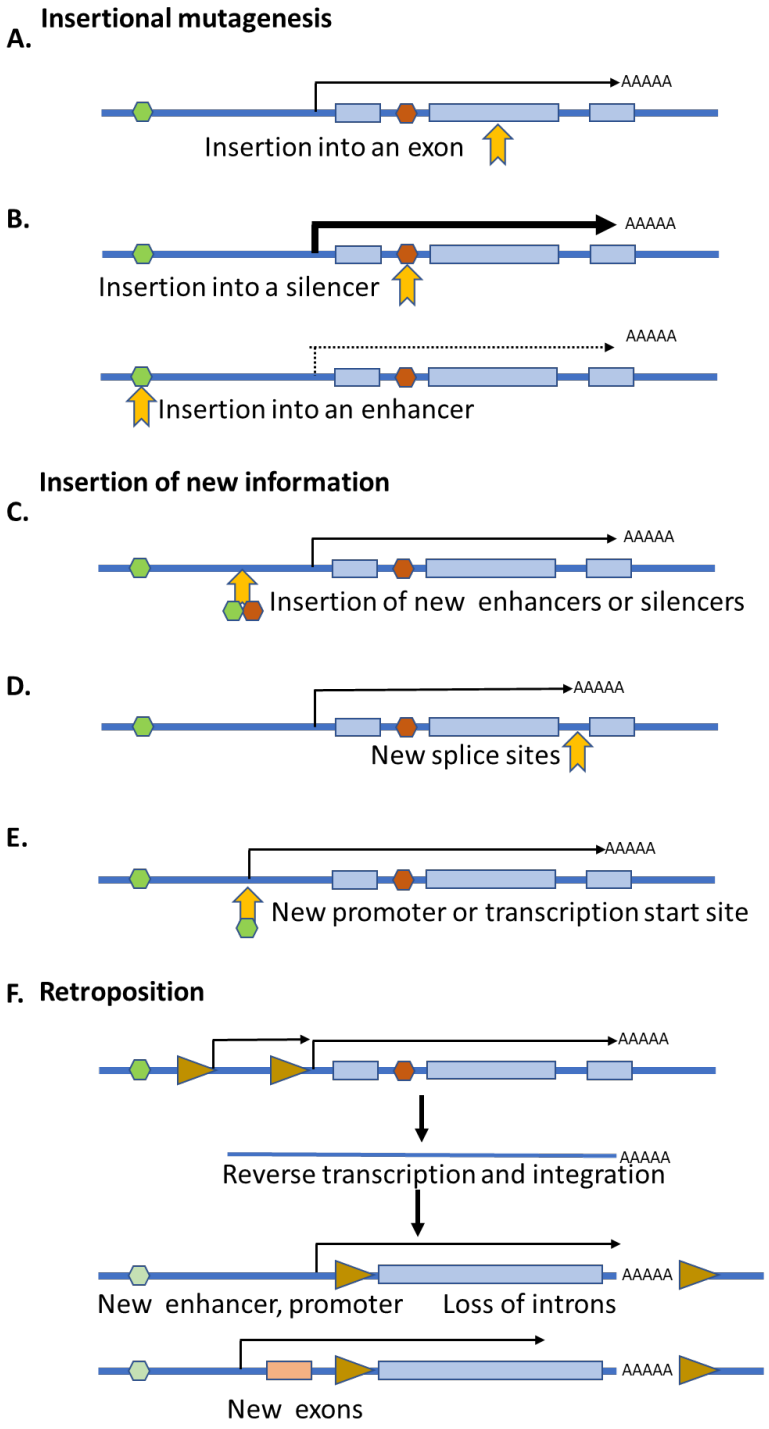

Figure 4. Structural and functional changes that can be caused by the insertion of transposable elements. *Exons are depicted as blue boxes, enhancers as green hexagons, silencers as red hexagons, TEs yellow arrows, transposon terminal repeats as orange triangles. Cm represents 5'-methylated Cytosines, a hallmark of silencing.*

A*. Insertion of a TE into a coding sequence interrupts the reading frame and causes a truncated protein to be produced.*
B*. Insertion into a transcriptional silencer or an enhancer may increase or decrease expression, respectively.*
C*. An inserted TE, possibly containing additional sequence, can itself act as an enhancer or silencer of transcription.*
D*. An inserted TE can include new mRNA splice sites that alter the composition of the mRNA.*
E. *The inserted TE may provide a new promoter to a gene, as well as a new transcription start site.*
F. *Retroposition: transcriptional readthrough from a TE into a neighbouring gene followed by reverse transcription and insertion into a new position produces a new, translocated copy of a gene (without the original introns), possibly under control of a new promoter and having new exons. The structure without the Reverse Transcriptase gene of the original TE is called 'retroposon'. Epigenetic regulation:*
G. *Insertion of TE 3' (downstream)of a gene may lead to transcription of an antisense RNA which can negatively regulate the original mRNA level.*
H*. An TE inserted near a gene may be silenced by (as a first step) DNA (cytosine) methylation, which may spread to the surrounding DNA and thus silence the gene expression as well. Adapted from* (Lisch, 2013a).

End Joining (NHEJ), leaving a 'footprint' of (usually) small rearrangements, mostly deletions, around the break site. However, since a DSB occurs, TE excision may be the initiator of larger rearrangements that are all connected to DSB repair, such as the *Tunicate* maize example discussed in section 6.2. In grasses in general, and specifically in a comparison of two rice species, it was apparent that transposon excisions correlate with increased mutation frequency in 3000 bp flanking the excision site (Wicker *et al.*, 2016), including large rearrangements such as deletions from hundreds to thousands of nucleotides as a result of DSB repair (Roffler and Wicker, 2015). The causative relation between DSB repair and deletions or rearrangements is discussed in more detail in Chapter 5.

4.3 Transposition in other eukaryotes

More than half of all phenotypic laboratory mutants of the fruit fly (*Drosophila melanogaster*) are caused by insertions of TEs. In laboratory mice, TE activity is responsible for 10-15% of heritable mutant phenotypes. Additionally, natural environments may impose stress on organisms that enhance TE activity to above that found in the laboratory.

In humans, many TE insertion polymorphisms cause no noticeable phenotypic difference, but several insertions, including *de novo* examples, lead to genetic disease following some of the mechanisms shown in Fig. 4 (Payer and Burns, 2019).

Approximately 44% of the human genome consists of TEs or related sequences. By pairwise comparing sequences between human individuals, it emerged that three TE families (*LINE*, *Alu*, and *SVA*) have been active in more recent times and some probably still are (Mills *et al.*, 2007). The human (haploid) genome contains an estimated 500,000 copies of the TE *long interspersed element 1* (*LINE 1*), yet all but 0.1% are no longer active due to accumulated mutations that render them incapable of transposition. An estimated 100 intact *LINE 1* copies are present in each person and are still segregating in the population. A small number of these is considered highly active or "hot". Active *LINE 1* transposition has also been demonstrated *in vitro*, supporting the assumption that at least *L1* is still currently active. On average, any two human haploid genomes differ by approximately 1000 TE insertions, most of the *L1* and *Alu* types (Bourque *et al.*, 2018).

Approximately 44% of the human genome consists of TE-like sequences. On average, any two human haploid genomes differ by approximately 1000 TE insertions

4.4 Transposition in plants or crops

4.4.1 Frequencies of transpositions

In crops that have been sequenced, TE or TE-derived sequences comprise from 20% (melon) to 82-84% (barley, wheat, and maize) of the entire genome sequence. For the majority of these crop species, class I TEs (retrotransposons) and in particular, Long terminal repeat (LTR-) retrotransposons are the most prevalent, although this is possibly skewed by the easier identification of this type due to the hallmark LTRs (Vitte *et al.*, 2014). Although the vast majority of TEs in the plant genomes are inactive because of silencing or mutations rendering them incapable of transposition, there is growing evidence from whole genome-resequencing data combined with improved algorithms for

In sequenced crops, TE or TE-derived sequences comprise from 20% (melon) to 82-84% (barley, wheat, and maize) of the total genome sequence

detection of Transposon Insertion Polymorphisms (TIPS) (Xing *et al.*, 2013), that the latter is very prevalent in all studied species. *Intra*species TIPS, differences in insertion events between varieties or accessions of the same species, point to (relatively) recent activity. This is within the time since the crop's domestication, which can still be up to 10,000 years. Only in a few cases, a frequency on a generation-scale could be inferred.

A selection of publications on TIPS in crops species is listed in Table 1.

Table 1. Summary of genome-analysis studies for intraspecies TE insertion polymorphisms (TIPS).

Species	sequence	# varieties[1]	#TIPS	Reference
Maize	assembly	4	400,000	(Anderson *et al.*, 2019)
Soybean	resequenced	31	34,154	(Tian *et al.*, 2012)
Rice	resequenced	3000	50,000	(Carpentier *et al.*, 2019)
Arabidopsis	resequenced	211	2835	(Quadrana *et al.*, 2016)

[1]Arabidopsis is not a crop and therefore does not have varieties, but so-called 'accessions', collected from different locations and then propagated.

4.4.1.1 Arabidopsis

In Arabidopsis, an analysis of 211 diverse re-sequenced accessions at different levels showed that any two of these differ by 200-300 relatively newly transposed TE copies since their divergence. A global study showed that TE insertions affected nearby gene expression in both positive as well as negative ways. DNA methylation of the TE commonly spread to 300 bp on both sides, but for a small number could extend up to 3.5 Kb (Quadrana *et al.*, 2016).

4.4.1.2 Maize

In maize, a comparison of 4 genome assemblies identified approximately 400,000 polymorphic TEs, distributed genome-wide and including deletions of older TEs, but also recent transposition events. Many were in (3070) or near (within 1 kb upstream, 10,500) genes, which amounts to 7% of maize genes. This underlines the importance of TEs in causing variation in genome organisation and gene content (Anderson *et al.*, 2019). Although this implies a significant effect on gene function, this was not assessed in that study. While inference of novel insertions and their time is generally difficult in these comparisons, the comparison of TE insertions in regions that have minimal sequence divergence between the accessions allow this to some extent. There were 29 examples from 19 families, of which 14 showed gene expression, and many have highly similar LTR pairs. The well-known previously detected transposition in maize was due to a small set of class II (DNA) transposons, such as *Ac/Ds*, *Mu*, and *Spm/En* in particular genotypes (Nannas and Dawe, 2015), and these were not found in the current study. Apparently, the type and activity of TEs in maize is highly genotype-dependent. In another study, using 3 of the 4 genotypes from Anderson et al. (2019) and one new genotype, it was shown very recently that many deletions, as well as insertional mutants, can occur by activation of lineage-specific retrotransposons specifically in male gametophytes, and is observable within one generation at a frequency of ~$4x10^{-5}$ at a single locus. Besides, screening for new insertion sites of three retrotransposons in a 1000-seedling pool revealed 18 to 300 new insertion sites (on average 0.4 per seedling), of which more than half in genes (Dooner *et al.*, 2019). The identified retrotransposons included ones described earlier, such as *Hopscotch*, *Magellan*, and *Bs2*; others, like *Foto* and *Focou,* were new.

A study on 3 retrotransposons in maize showed on average 0.4 new insertions per seedling per generation

4.4.1.3 Soybean

During domestication of soybean (*Glycine max*), four recently duplicated *FLOWERING LOCUS T* (*FT*) genes followed different evolutionary paths, where the expression of one gene was attenuated by a TE insertion nearby, and another gene was inactivated by a TE insertion in an exon (Wu *et al.*, 2017).

A currently active CACTA-like element, *Tgm9*, causes variegated (having sections with different colours) flowers in soybean, through excision of an inserted copy from *DFR2*, restoring its activity (and colour) in the process (Xu *et al.*, 2010). This excision is both somatic (leading to variegated flowers) as well as germinal (leading to entirely purple flowers in the progeny) at high frequency and may be followed by reintegration nearby. One *CACTA*-type class II TEs carries additional gene fragments – a *Pack-CACTA*- and is discussed in detail below.

4.4.1.4 Cotton

Common upland cotton (*Gossypium hirsutum*), the predominant cultivated species of cotton, is an allotetraploid derived of *G. arboreum* and *G. raimondii* genomes. Diversity in transposon content and position does exist and has been used, for example for genotyping (Noormohammadi *et al.*, 2016; Noormohammadi *et al.*, 2018), and transposon content in the relatively recently sequenced subgenomes and *G. hirsutum* genome have been characterised (Wang *et al.*, 2016; Lu *et al.*, 2018; Liu *et al.*, 2018; Cheng *et al.*, 2019). However, probably due to the polyploid genome, a proper *G. hirsutum* reference genome has become available only relatively recently (Wang *et al.*, 2019), and genome-wide studies of TIPS are not (yet) available. The recent *de novo* assembly of two upland cotton cultivars and study of SV therein will probably soon be followed by such an analysis (Yang *et al.*, 2019).

4.4.1.5 Rice

Analysis of data from 3000 resequenced rice genomes, tracking TIPS from 32 retrotransposon families with low to very high copy number, detected 50,000 TIPS. A considerable amount of them had distributions and variation in the occurrence that suggested that they have been active, or still are, during agriculture (Carpentier *et al.*, 2019).

4.5 Effects of TE insertion on crop phenotypes

For flowering plants, including crops, we refer to a review from 2013 that lists 65 instances of TE insertions through various mechanisms involved in the development of traits during domestication, breeding, or in wild plants (Oliver *et al.*, 2013) and below we focus on relevant, as well as more recent examples. These are, among others, a retrotransposon insertion in grape leading to colourless fruit skin varieties (Kobayashi *et al.*, 2004), in rose leading to continuous flowering (Iwata *et al.*, 2012), in apple leading to seedless fruit development (Yao *et al.*, 2001), in cauliflower leading to orange or purple curd (Lu *et al.*, 2006; Chiu *et al.*, 2010) and underlying cold-inducible anthocyanin accumulation in Sicilian blood oranges (Butelli *et al.*, 2012). In *Brassica rapa*, low erucic acid cultivars originate from an LTR-retrotransposon insertion in a gene encoding a key enzyme in erucic acid synthesis (Fukai *et al.*, 2019). In *B. rapa* vegetable crops such as Chinese cabbage, resistance to early bolting (flowering) is important for vegetable production. A transposon insertion in a flowering activator *FLOWERING LOCUS T* gene copy (*BrFT2*), knocking out its expression, was found to underlie a late-flowering trait (X., Zhang *et al.*, 2015). In Arabidopsis ecotype L*er*, a *Mutator*-like TE insertion in the first intron of the flowering repressor *FLC* is correlated with lower expression and early flowering (Gazzani *et al.*, 2003).

Cluster regions of the resistance genes of the NLR type (see section 5.2.2) are commonly associated with TEs (Bayer *et al.*, 2019; Van de Weyer *et al.*, 2019), and TEs have been shown to play a role in generating variation. An LTR retrotransposon *Renovator* insertion in the promoter of the rice *Pit* R gene was shown to enhance its expression. The enhanced expression was proven to be important for generating resistance to the fungal pathogen *Magnaporthe grisea* in comparison to the susceptible variant in another rice variety lacking the TE (Hayashi and Yoshida, 2009). A *COPIA-R7* retrotransposon insertion in an intron of the Arabidopsis resistance gene *RPP7* against the downy mildew *Peronospora parasitica* affects gene function through changing epigenetic control. The TE was shown to generate an epigenetic histone mark (H3K9me2) at the *RPP7* locus. In turn, this leads to a change in the processing of the transcript into the functional form for RPP7 resistance (Tsuchiya and Eulgem, 2013).

4.5.1 Maize

The classic *Ac/Ds* elements were the first TE identified in maize by McClintock, and there insertion and excision were observed through their effect on maize kernel colour, variegation in leaves, and on the *wx (waxy)* locus (McClintock, 1950).

The insertion of a *Hopscotch* TE upstream of the maize domestication gene *teosinte branched* (*tb1*) acts as an enhancer of gene expression leading to increased apical dominance when compared to maize's progenitor, teosinte (Studer *et al.*, 2011). In maize, several mutations in the *WAXY* gene exist, and *Waxy* kernels (*wx*) induced by TE insertion-inactivation of a granule-bound starch synthase causes amylose-free starch production (McClintock, 1950). *ZmGE2* contains a TE-disruption of a cytochrome P450-enzyme gene, leading to higher kernel oil contents due to an increased embryo/endosperm volume ratio (Zhang *et al.*, 2012).

4.5.2 Tomato

The occurrence of various mutations caused by new insertions of *Rider* in tomato, which were observed in the past centuries gives an impression of the frequency (if not quantitatively) at which crops phenotypes are modified by TE activity. The *yellow flesh* mutation is a *Rider* insertion into *PHYTOENE SYNTHASE1*, encoding the first committed step in lycopene biosynthesis (Lindstrom, 1925). *Rider*-mediated gene duplication is involved in elongated fruit shape and is estimated to have occurred within the last 200-500 years (Xiao *et al.*, 2008). In the chlorotic *fer* mutant, *Rider* has inserted into a gene encoding a transcription factor (Ling *et al.*, 2002). A *Rider* insertion into an MYB transcription factor-encoding gene underlies the *potato leaf* mutation (Busch *et al.*, 2011), giving the tomato leaves the semblance of the less- dissected potato leaves (Price and Drinkard, 1908). Finally, a *Rider* insertion into a MADS-box gene results in the loss of fruit pedicel abscission zones, the *jointless-2* (*j-2*) mutation (see below). Apart from the *j-2* allele, there are no indications that these events were introgressed from wild relatives. Instead, they occurred in cultivated tomato, but most are older than a century and difficult to trace back to a specific time.

There are many examples of novel traits in crops, caused by TE insertions

4.6 Epigenetic effects of TE insertion on gene expression

Due to the potentially harmful effects of new TE insertions on genes and the genome, most TEs in plant genomes are silenced through a mechanism involving, among others, DNA cytosine methylation and histone modification (reviewed in (Lisch, 2009)). DNA methylation of a TE can be extended to nearby host DNA up to 300 bp outward from the insertion site, indicating that silencing effects may spread to nearby genes when the TE insertion is close enough or within a gene's intron (Fig. 4H). Methylated TE insertions are associated with reduced expression of adjacent genes in a comparison between two Arabidopsis species (Hollister *et al.*, 2011), as well as in an intraspecies comparison of TIPS in three *A. thaliana* accessions (Wang *et al.*, 2013).

TE insertion can lead to silencing of nearby genes.

Variation in epigenetic modification of transposons may lead to phenotypes, as demonstrated by the *jointless-2* (*j-2*) allele in tomato. The insertion of a *Rider* LTR-retrotransposon in the first intron of a transcription factor gene leads to a decrease of its expression through the spread of DNA cytosine methylation from the TE to the neighbouring exon (Roldan *et al.*, 2017).

It is therefore clear that transposons in crops have had a role in the formation of both advantageous as well as detrimental traits. The presence of a *Karma* retrotransposon, in combination with its tissue culture-induced hypomethylation in oil palm, leads to alternative splicing of a floral homeotic gene with the "mantled flower" and reduced oil yield phenotype as a result (Ong-Abdullah *et al.*, 2015).

4.7 New combinations of genetic elements from TE activity and gene capture

In many cases, structural variation can arise from the activity of TEs in conjunction with Double-Strand Break repair (reviewed in (Krasileva, 2019). LTR retrotransposons, in particular, can capture genes during their reverse transcription when a template switch to an mRNA of a different gene occurs (Fig. 4D, and Fig. 5A). This process, retroduplication, was involved in the expansion of a subfamily of immune receptors in pepper (Kim *et al.*, 2017). A prominent example in tomato is the *sun* locus (see chapter 7 and Fig. 8D). LTRs can also align to induce excision or duplication of associated genes by non-allelic homologous recombination (NAHR, Fig. 7). As many TEs are activated by stress, the insertion of a TE upstream of a gene may confer stress-regulation to that gene and be an adaptive advantage.

DNA TEs that can trap genes or gene fragments are Mutator-like transposable elements (MULEs), and Helitrons. MULEs, type II TEs, are called Pack-MULEs when containing gene fragments. Pack-MULEs are most abundant in rice (2,853), followed by maize (276), and Arabidopisis (46). Carrying gene fragments, they preferentially insert in the 5' end of existing genes, and may thus evolve additional exons in these genes (Jiang *et al.*, 2011) (Fig. 5B). In rice, Pack-MULEs, contain fragments derived from 1000 cellular genes, and they often contain pieces of multiple additional loci, fused to form new open reading frames. About 5% of these are represented in cDNAs, meaning that they are also expressed. In soybean two type II-CACTA family TEs called *Tgm-Express1* en *2,* respectively contain insertions of other soybean genes (Zabala and Vodkin, 2005). The former includes 5 different gene fragments, and when inserted in an intron of the soybean *F3H* gene, these lead to an intricate pattern of alternative transcript splicing forms. Some of the additional gene fragments were present as additional exons of the *F3H* transcript (Zabala and Vodkin, 2007). Clearly, an example of new combinations arising from TE insertions. MULEs are widespread in plants, and thus Pack-MULES may be an essential instrument in gene evolution in plants (Jiang *et al.*, 2004).

Pack-MULES are transposable elements that trap genes or gene fragments. They preferentially insert at the start of genes, which produces new combinations of genetic elements. Plants can have hundreds of Pack-MULES

Helitrons are DNA TEs that do not have repeats at their ends (and are therefore more difficult to detect). They can capture several genes and produce chimeric transcripts, possibly through readthrough caused by the loss of downstream terminator or by template-switching to an exogenous gene during double-strand break repair as part of their insertion process (Fig. 5C) (Morgante *et al.*, 2005; Lai *et al.*, 2005; Kapitonov and Jurka, 2007).

Also Helitron TEs can capture genes by producing chimeric transcripts

Finally, DNA transposons can produce duplicated genes by duplicating an adjacent gene when two TEs are positioned close to each other in the same orientation and transposition involves the use of the two outer terminal repeats, forming a mobile hybrid TE (Fig. 5B) (Zhang *et al.*, 2013).

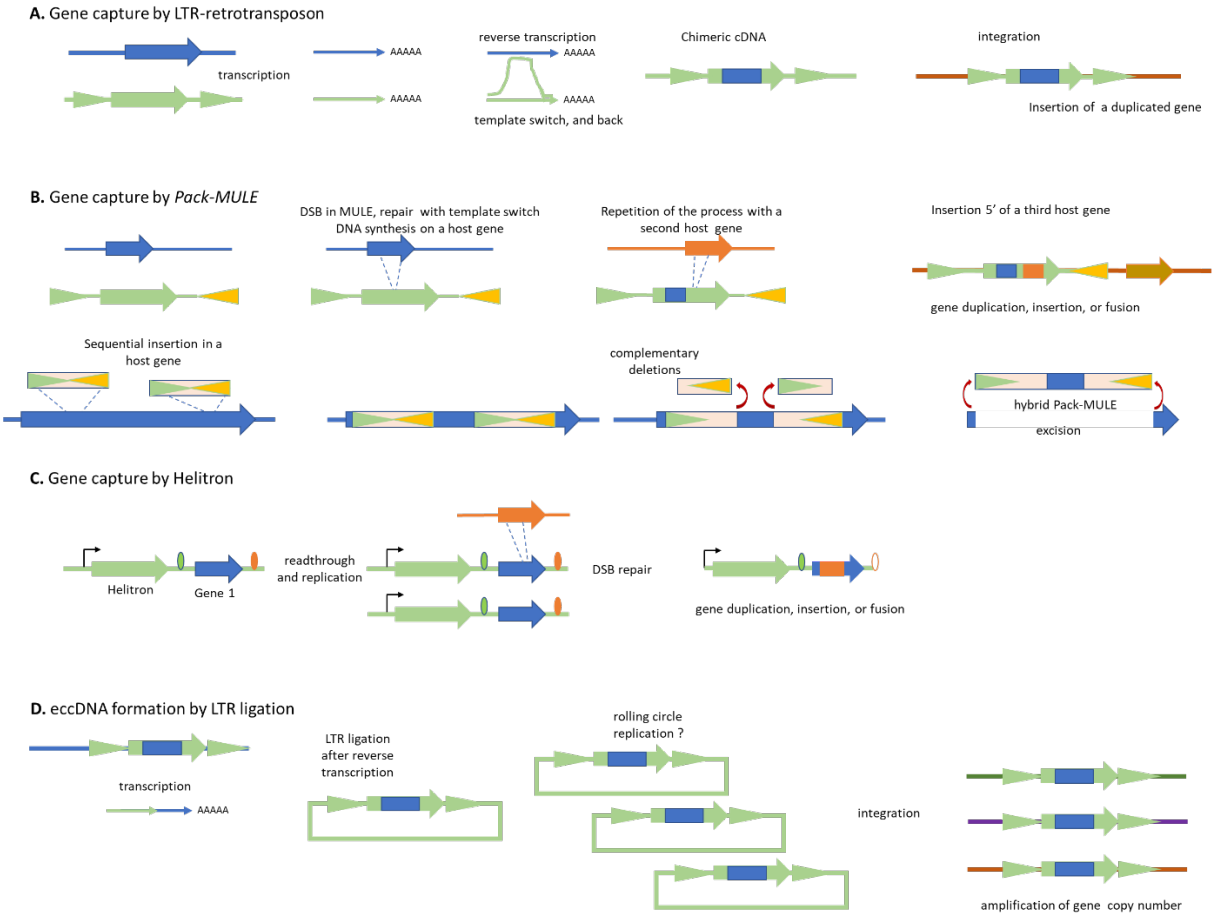

Figure 5. Models for the mechanism with which TE activity can produce new combinations of genes or gene parts. *Mostly adapted from* (Krasileva, 2019). *(Retro-) transposons are depicted by green lines and arrows. Terminal repeats of TEs are represented by green and orange triangles. Host genes are depicted by blue and orange arrows.*
A. *Model for gene capture by retrotransposons. During reverse transcription of the retrotransposon RNA, the TE's reverse transcriptase may switch templates to a host gene mRNA and back again, so generating a chimeric cDNA, which upon new insertion, produces a stable chimeric gene (see also the example for the* sun *duplication in tomato).*
B. *Models for gene capture by Pack-MULEs. New DNA may be incorporated by template-switching to a host gene during DSB repair. In an alternative model for gene capture, two MULE elements insert in the same gene and undergo complementary deletions, after which excision takes place using the outermost repeats, mobilising the middle part of the gene. Adapted from* (Lai *et al.*, 2005).
C. *Gene capture by Helitrons. Readthrough of the Helitron transcription or replication to an alternative inverted repeat (orange oval) may lead to the incorporation of a neighbouring gene, and insertions during DSB repair (see also B) may additionally incorporate parts of unrelated genes. Adapted from* (Kapitonov and Jurka, 2007)
D*. Extrachromosomal circular(ecc)DNA formation by LTR ligation. The reverse-transcribed cDNA of a TE, containing a host gene fragment as in (A), may circularise by NHEJ and produce an eccDNA that can be amplified by rolling circle replication. Amplified copies may then remain free or integrate at unrelated sites in the genome*

DNA transposons can produce duplicated genes, and can lead to high copy numbers of genes.

In addition to other mechanisms, extrachromosomal circular DNA (eccDNA) can originate from end ligation of reverse-transcribed retrotransposon transcripts or of mobile DNA transposon intermediates, possibly followed by amplification, such as through rolling circle replication. This leads to an increase of the copy number of any gene (fragment) that is co-mobilized (Fig. 5D). More insight in these phenomena comes from recent developments in the sequencing of extrachromosomal intermediates of transposition, the 'mobilome',

which also identifies the transposons most likely to be active, either in specific developmental stages, upon hypomethylation, or under stress (Lanciano *et al.*, 2017; Quadrana *et al.*, 2016; Cho *et al.*, 2019). The role of eccDNA in copy number variation, which is associated with herbicide resistance is discussed elsewhere in section 7.3.

4.8 Transposon mutagenesis

The (highly) active TE elements of maize have been extensively used as a tool for mutagenesis in maize (Nannas and Dawe, 2015), as well as through transgenesis in other crops. Yet, most other crops contain no TEs with natural activity high enough to make such mutagenesis feasible. However, activating transposable elements in crops is an attractive strategy for creating genotypic diversity and new traits for plant breeding (Paszkowski, 2015). Recognised early, various biotic and abiotic stresses activate mobility of transposons (McClintock, 1984). Alternatively, transposons which are usually silenced may be activated in mutants (Tsukahara *et al.*, 2009). Two recent independent reports use such lines to detect transgenerational TE activity. Progeny of an Arabidopsis inbred line with a mutation in the methylation gene *met1* mutation (originated in 2009) was sequenced using long-range sequencing. This revealed 9 new insertions of the LTR-retrotransposon *EVADE* and one of DNA transposon *CACTA* (Debladis *et al.*, 2017). Using the same mutation, resequencing of 67 descendants showed that *EVADE* numbers had increased 12 times, and a PACK-type *CACTA1* had increased 7 times (Catoni *et al.*, 2019). Moreover, 50 new insertions of Pack-CACTA were found in 8 lines.

A mutation in a DNA methylation gene in Arabidopsis leads to activation and multiplication of several TEs

In a combination of disruption of silencing and activation by heat stress in Arabidopsis, the LTR-retrotransposon *ONSEN* was activated and increased in copy number, indicating novel insertions. Using this strategy, a potentially stress-tolerant TE insertion mutant was produced in one generation (Ito *et al.*, 2016). These results show that previously silenced TEs can create many new inserts in a few generations when activated. Curiously, activation of endogenous transposons and characterisation of their (semi-) random insertions, such as recently shown for tomato LTR-retrotransposon *Rider* can provide an attractive, non-genetic modification type of mutagenesis (Benoit *et al.*, 2019).

When previously silenced TEs are activated, e.g. by stress, this can create many new inserts within a few generations

4.9 Conclusions

Transposable elements make up a large part of any plant genome. Most of these TEs are not or no longer active because they have degenerated over time, or because their activity is actively suppressed (silenced) by the host. Nonetheless, most of the studied species have some active transposons, and this activity can vary between accessions of one species. Furthermore, TE activity can remain hidden until their activation by stress or by interference with the silencing machinery.

The activity of transposons and the considerable variation in possible outcomes (Fig. 4), as discussed here is very likely to be a major, if not the main cause of structural variation differences between crop varieties or accessions. However, this conclusion awaits support from more quantitative data for SV occurrence between generations (see next chapter). Some estimation of per-generation-frequency may be derived from a comparison of varieties of crop with a known pedigree and timeline, if available. In many cases, this estimation will be confounded by the phenomenon of introgression breeding, where portions of the genome of wild relatives of the crops are introduced by crossing and stabilised by back-crossing, for the purpose of introduction of disease resistance or other beneficial traits. Any such introgression brings along

a set of SVs (including those from transposons) which have evolved independently over a much more extended time since the split from a common ancestor and should be left out of the estimation. One example from this chapter is the *jointless-2* mutation in tomato, which was introgressed to tomato from a wild relative (Rick, 1956). This observation even extends to individuals within one cultivar. The cultivar may contain introgressions from a different variety, and the introgression's size and positions have not been stabilised (fixed) during backcrossing. In such a case, CNV's between individuals may just reflect the different amounts of introgressed DNA, and the CNVs can be traced back to their origin in the two different parents. An example is a variation among individuals of a soybean reference cultivar (Haun *et al.*, 2011).

From the observation of the phenotypic variety derived from TE insertion described here, it can be inferred that TE activity is an essential factor in creating phenotypic variation and that much activity is going unnoticed because it does not result in a visible phenotype. However, a precise estimation of activity on a per-generation timescale is available for only a few examples, and generalisation of these data is tricky, since TE activity may differ greatly among genotypes of the same species. An exciting new addition to the identification tools for active transposons is the sequencing of the 'mobilome'.

TE activity is an important factor in creating phenotypic variation. Much activity is going unnoticed because it does not result in a visible phenotype. A precise estimation of activity on a per-generation timescale is available for only a few examples, and generalisation of these data is tricky, since TE activity may differ greatly among genotypes of the same species

TE activity may make a substantial contribution to the evolution of new genes through their capacity to mobilise (parts of) genes (see the Pack-MULE and Pack-CACTA examples in rice and soybean, respectively) and insert them in new positions, including inside other genes. It is not yet known what the natural generation-to-generation frequency of such events is, because novel insertions were only measured in mutant plants that were compromised in their silencing machinery. Towards the discussion of the "baseline" to which to compare the changes in new crops (see further on in this report), this chapter produces additional questions. The implementation of Transposon mutagenesis strategies that involve increasing the natural background activity of endogenous TEs, as described above, may or may not shift this baseline upwards.

5 Other mechanisms of SV generation

5.1 Introduction

This chapter describes mechanisms underlying structural variation other than from transposable elements (TEs). Because TEs represent a major path to the creation of new combinations of regulatory elements and genes, they warranted a separate discussion (see the previous chapter, 4). The process of change of the structure of a chromosome is usually not observed, but the resulting SVs are. In hindsight, models can be proposed for the mechanisms that have led to the observed differences in structure. Several models for SV creation have been proposed. They can be broadly divided into those based on TEs (see the previous chapter), chromosome replication errors, and those based on double-strand breaks and their repair. While DSB repair discussed in this chapter, it should be realized that TE excision is an important cause of DSB formation and that its 'sloppy' repair as described below, is an important factor in SV creation and the formation of new combinations of genetic elements. Although often first described for other eukaryotes, they have now all been observed in plants as well.

5.2 Errors during DNA replication

During DNA replication a moving replication fork is formed, which progresses through unreplicated DNA, driven by DNA synthesis. The strand that is synthesised continuously in the direction of fork movement in the 5' to 3' direction is the 'leading strand'. The other strand, which is necessarily produced in fragments in the opposite direction, is the 'lagging strand'. In repeat-rich DNA, such as short-interspersed repeats (SSRs, e.g. ATATATAT, or ATG ATG ATG), the fork may stall, leading to slippage of the replication machinery. In that case, repeats may be added or removed from the SSR (Fig. 6A). As mentioned in the Introduction, we will not pay further attention to this frequently occurring phenomenon, as we do not regard this as structural variation, but as small InDels.

Alternatively, when a replication fork is stalled, the elongating lagging strand is 'looking for' a template, and may invade another replication fork nearby that has microhomology with DNA at the original fork. Subsequently, it continues DNA synthesis from this new position (Fig. 6B). This is named Fork Stalling and Template Switching (FoSTeS). Depending on the relative locations of the forks and which strand of the invaded fork is used as a template, insertions, deletions or more complex rearrangements will be the result.

Fork stalling during DNA replication can lead to SV

Further, mechanisms have been proposed that lead to inverted repeats (J.,-M., Chen *et al.*, 2005; Löytynoja and Goldman, 2017). These mechanisms would explain several inverted repeats in humans, that have been causal for genetic diseases. During DNA replication, the two complementary DNA strands are separated, and subsequently, new complementary strands are made. However, when during the synthesis of the complementary strand, replication slippage occurs, e.g. at mononucleotide repeats (such as AAAAAA or TTTTTT), that strand may switch to the other 'mother strand' using that as a template. This template-switching may create an inverted repeat, that can lead to an RNAi-like hairpin later, giving rise to gene silencing.

5.3 Double strand break (DSB) repair

Duplications, deletions, inversions, and chromosomal translocations can also be a consequence of inaccurate repair of DNA double-strand breaks (DSBs). When a DSB occurs, the organism may join the

two broken ends together again, a process called non-homologous end-joining (NHEJ). This process is probably accurate most of the time, or produces small Indels at most (Fig. 6C).

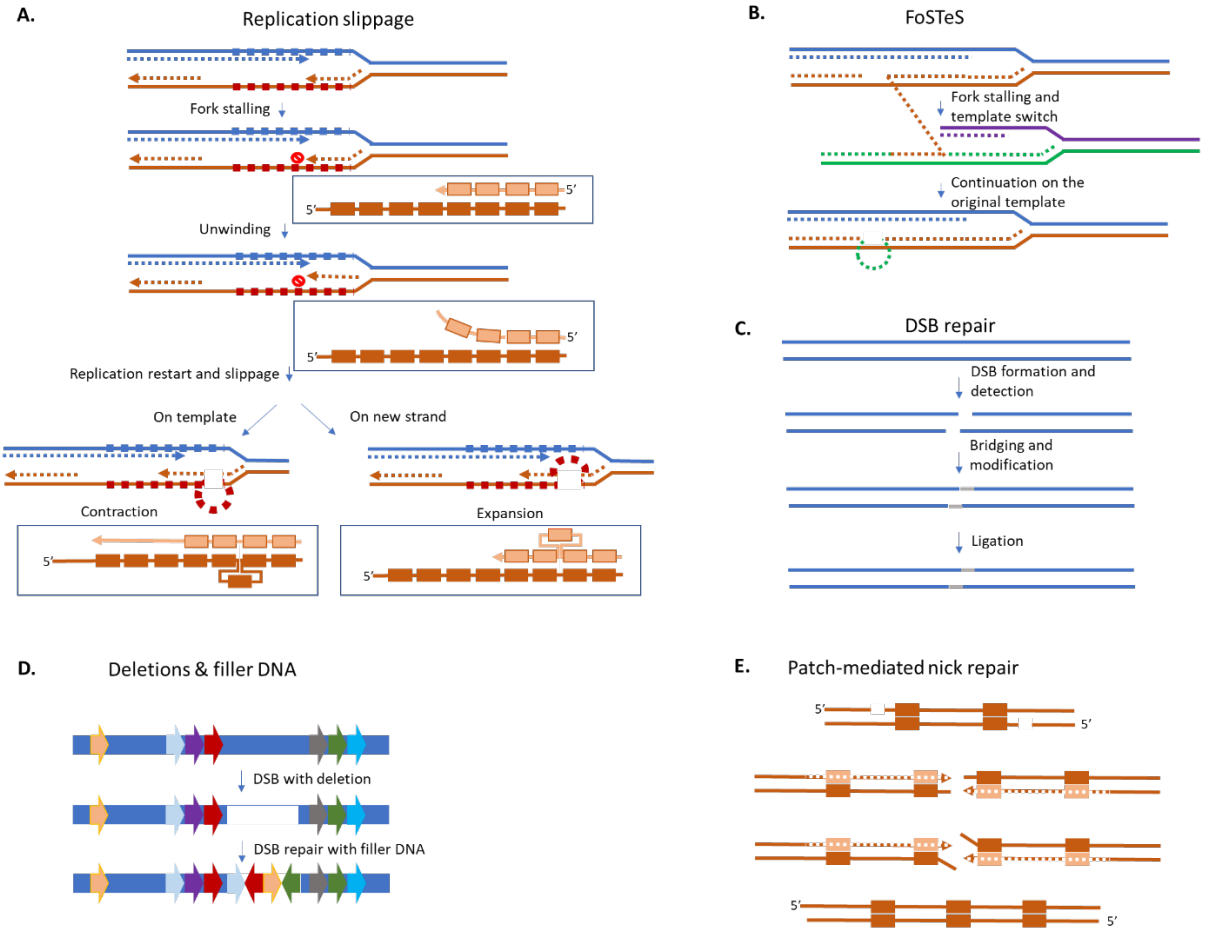

Figure 6. Models of DNA-level mechanisms of SV generation.
A. *Fork stalling and replication slipping during DNA replication in a repetitive genomic region (repeats are depicted by the "beads" on both strands). Upon stalling, the lagging strand disengages and resumes DNA synthesis further up the template, skipping some repeats ("slipping" on the template) leading to a contraction of the repeat region (left). Alternatively, the new strand may slip and thus add some repeats, leading to expansion (right). Boxed inserts show a detail of the lagging strand at each stage.*
B. *Fork Stalling and Template Switching (FoSTeS). A stalled lagging strand may disengage and anneal, by microhomology to another nearby replication fork and continue synthesis there. This may be repeated on another fork before reannealing to the original template, thus incorporating one or more stretches of unrelated sequence (green).*
C. *Double-Strand Break (DSB) repair by non-homologous end joining (NHEJ) is either perfect or involves the production of small indels.*
D. *DSB repair by annealing of microhomologies around but at a distance from the break site may result in deletion of the intervening DNA. Alternatively, when Synthesis-Dependent Strand Annealing (SDSA) is involved in the repair, a free 3′ end of the broken DNA may be extended by DNA synthesis using sequences near the break site, but also entirely unrelated sequences as a template before ligation of the break. This results in a patchwork of apparently randomly inserted sequences at the break site.*
E. *Patch-mediated nick-repair of nearby single-strand breaks (nicks) in a repetitive region. The free 3′ end at each nick is extended by DNA synthesis using the overhanging strand as a template, effectively creating an extra copy of the sequence between the nick positions.*
Adapted from (Guo *et al.*, 2014) *and* (Vaughn and Bennetzen, 2014).

There are several variants described for inaccurate DSB repair. DSB repair can include insertion of "filler" DNA, including insertion in spontaneous deletions caused by DSBs (Fig. 6D). In maize, it was observed that several spontaneous deletions in the *waxy* locus occur in G-C rich regions, which may form secondary structures when single-stranded. The endpoints of these deletions were all separated by 1-131 bp "filler" DNA, usually but not always (see below) derived from within 50 bp of the deletion endpoints, and often a complex of short and/or reiterated sequences (Fig. 6D). These deletions were flanked by repetitive sequences, suggesting a mechanism involving mispairing during DNA replication (Wessler *et al.*, 1990). Filler DNA may be added when Synthesis-Dependent Strand Annealing (SDSA) is involved in the repair. A free 3' end of the break may be extended by DNA synthesis using sequences near the break site, but also entirely unrelated sequences as a template before ligation of the break. This results in a patchwork of apparently randomly inserted sequences at the break site.

DSB repair can include insertion of "filler" DNA

An alternative model for the production of small local tandem duplications as filler DNA is "patch-mediated nick repair" (Fig. 6E). Here, single-strand lesions in opposite strands at distances from 30-80 bp of each other form the starting points of repair strand synthesis. This leads to blunt ends where they meet. Subsequent NHEJ results in a tandem duplication (Vaughn and Bennetzen, 2014).

In a rice study, analysis of indels >9 bp (note that this size is not included in our definition of SVs) in 50 accessions yielded over 65,000 SVs, which were analysed for the presence of filler DNA. Assuming that insertions followed DSB repair, the majority of small insertions (<20 bp) might be explained by the patch-mediated nick-repair model, but 60% of the incorporated sequences >100 bp were ectopic (no match with the local region) (Vaughn and Bennetzen, 2014). Much larger, up to 50 kb regions were apparently moved to DSB sites caused by transposon excision during the evolution of grasses, resulting in the loss of collinearity (common gene order in genomes) between rice, sorghum, and Brachypodium (Wicker *et al.*, 2010). Apparently this a powerful mechanism for reshuffling DNA sequences during DSB repair. The rice study discussed here suggests that this also occurs within a species' evolution timeframe and is possibly a common phenomenon. Moreover, studies on the type of DSB repair of linearised plasmids in tobacco protoplasts showed that 30% of the repaired junctions contained filler DNA in patchworks ranging from 2 bp to 1.2 kb (Gorbunova and Levy, 1997).

For repair of double strand DNA breaks, sequences from nearby or from elsewhere in the genome can be used as a template for repair, leading to new combinations of genetic elements

DSB repair can also proceed in an RNA-mediated fashion. Non-homologous end joining (NHEJ) may use non-homologous cDNA from random RNAs or use RNA as a template, enabled by reverse transcription (RT) (Meers *et al.*, 2016).

5.4 Gene conversion

Instead of ligation of the two broken ends (non-homologous end joining), a process of homology-directed repair (HDR) can occur. For initiation of this process, one strand at the break is removed (resected), and the remaining strand starts 'searching' for a homologous sequence, as a template for repair. That similar sequence can be detected in the sister chromatid, in the homologous chromosome in somatic cells (Molinier *et al.*, 2004) and during meiosis, or even in a different chromosome (see Fig. 7F). The result is that the broken sequence is replaced by a "donor" sequence without the donor sequence being altered. The acceptor sequence and the donor sequence share high levels of homology. This process is called gene conversion. Experimental evidence comes mainly from yeast (Freire-Benéitez *et al.*, 2016), but recently also from tomato. Gene conversion may generate new combinations of sequences, such as promoters and

genes, when the underlying sequences show high homology. The sizes of the conversion tracks (the 'copied DNA stretch') of the gene conversions that frequently occur during meiosis is about a few tens of bps, as observed in tomato (Sander Peters, WUR, pers. Comm.). It is difficult to give exact estimations when the number of polymorphisms between the homologous chromosomes is low, as conversion tracks are only visible where DNA polymorphisms are present. Gene conversion in somatic cells in yeast and plants give rise to far larger conversion tracks and can be up to a 2 kb long (Gorter de Vries *et al.*, 2018; Filler Hayut *et al.*, 2017).

5.5 Non-allelic homologous recombination (NAHR)

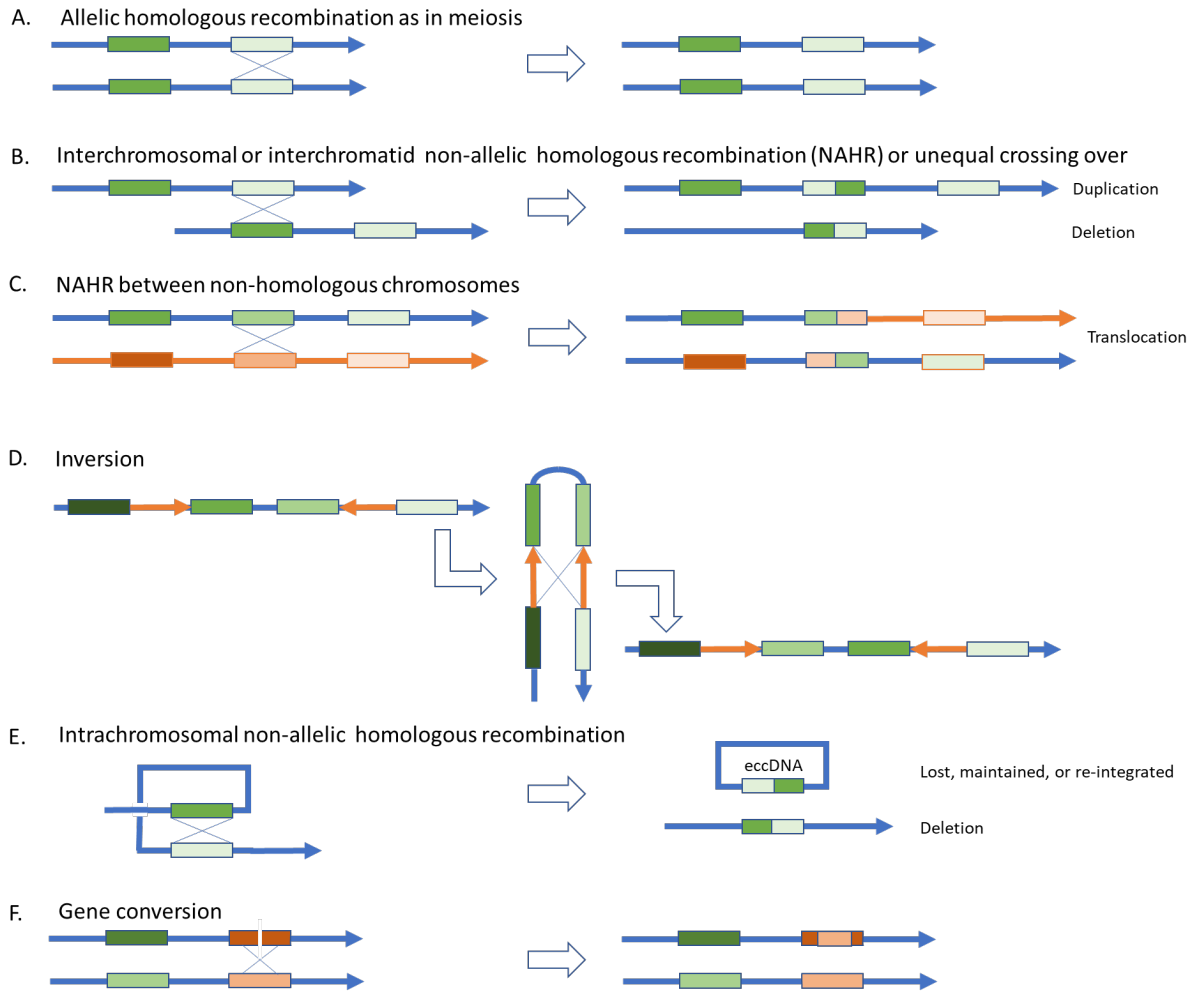

Figure 7. Various modes of recombination between alleles and 'non-alleles'. ***A****. Allelic homologous recombination.* ***B.*** *Non-allelic sequences with high homology, either between homologous chromosomes (inter-), or* ***C****. between non-homologous chromosomes leading to translocation, or* ***D.*** *within one chromosome (intra-), in a tandem inverted repeat, leading to an inversion or* ***E.*** *between tandem direct repeats on the same chromosome, leading to a local deletion. The looped-out DNA may be lost or maintained (and even replicated) as extrachromosomal circular (ecc)DNA, and eventually, insert elsewhere in the genome during DSB repair.* ***F.*** *gene conversion occurring during meiosis between allelic but non-identical sequences. A double-strand break in one chromatid (top) is repaired using a homologous sequence on another chromosome (bottom), resulting in a chimeric gene. Partially, adapted from* (Peng *et al.*, 2015)*.*

Where the other mechanisms involving DSBs, except Gene Conversion, are often based on microhomologies, non-allelic homologous recombination (NAHR) involves recombination between DNA regions with higher levels of homology that are not alleles of the same locus (i.e. they are not in the same physical position on the two parental chromosomes, or even on different chromosomes, Fig. 7). The similarities mostly encompass low-copy repeats (LCR, >1 kb) across the genome. Depending on the orientation of the repeats, NAHR may result in deletions, duplications or inversions. NAHR can happen

between areas on the same chromatid, or between chromatids, or even between chromosomes. (Fig. 7B-E).

In hotspots containing tandemly repeated sequences, non-allelic homologous recombination (NAHR) and unequal crossover will lead to recurrent rearrangements. During meiosis, when the homologous chromosomes pair, the pairing of the genes in the cluster can be imperfect (unequal crossing-over), leading to loss of some alleles in one chromosome, and gain of alleles in the homologous chromosome. Also, new combinations of a coding gene and a promoter can occur, or new combinations of parts of genes, because of the unequal crossing-over. These elements all reside within the cluster, albeit on different, homologous chromosomes.

The most prevalent disease resistance genes in plants encode a nucleotide-binding site and a leucine-rich repeat domain (*NB-LRR*-related genes, *NLR*). Often, these *NLR* genes are clustered in tandem repeats. These clusters are known to be hypervariable. Unequal crossing-over is probably a major driver for these new combinations of genetic elements within the cluster. Such a quickly evolving cluster is important for the creation of new R-like genes, that possibly provide novel resistances to pathogens that have overcome resistances provided by older R-genes (see section 6.7).

5.6 Inversions

Inversions that place duplicated flanking regions in tandem, predispose to higher frequencies of structural variation (a so-called "microdeletion syndrome" in humans) (Acuna-Hidalgo *et al.*, 2016). Furthermore, inversions limit recombination and thus could promote ecotype differentiation in heterogeneous environments. Inversions inhibiting recombination could protect favourable allele combinations and change gene expression near breakpoints. For human inversion 17q21.31, there are indications that it is increasing fertility and has been positively selected for in Europeans. Inversions can also be associated with human disease and be predisposing to other (recurring) rearrangements (Giner-Delgado *et al.*, 2019).

Rearrangements like inversions have caught attention in breeding programs because of problems with introgression of desirable genes from compatible wild relatives. Examples are inversions introduced with virus resistances from wild relatives into tomato (Wolters *et al.*, 2015). These are SVs likely having occurred over longer evolutionary timespan than is the focus of this study.

5.7 Extrachromosomal circular DNA (eccDNA)

EccDNA is prevalent in all eukaryotes (Cohen and Segal, 2009), and in plants was initially associated in particular with satellite DNA repeats in several species, including many repeats with a monomer length over 50 bp and present in up to millions of copies (Navrátilová *et al.*, 2008). EccDNA may be formed by looped-out DNA during intrachromosomal NAHR (Fig. 7E) and, through deletion or reintegration, contribute to copy number variation particularly of tandem repeats. Alternatively, as described previously in section 4.7 and Fig. 5D, they may be post-excision or post-replication intermediates of transposon mobilisation. Their numbers may vary in somatic cells of an individual, and they may be able to replicate autonomously such as by rolling circle replication (Cohen and Segal, 2009). They can also be transmitted through meiosis as they are often found tethered to chromosomes (Koo, Molin, *et al.*, 2018). As described in more detail in section 7.3, eccDNA's are involved in the rapid evolution of herbicide tolerance.

5.8 Aneuploidy and chromothripsis

Under stresses like tissue culture (also see section 8.2) or inducing haploid progeny by hybridization between distantly related species, more extensive rearrangements may occur. Haploids are a favoured breeding method for the fast production of homozygous (inbred) lines by genome doubling of haploids. Aneuploidy, i.e. additional or missing chromosomes, can be a result, but also chromothripsis is a possible outcome. In chromothripsis, chromosomes have undergone disintegration and reassembly, often by

illegitimate rejoining in micronuclei, which has mostly been described for mammalian cancer cells (Yi and Ju, 2018). Usually, chromosomes would be degraded entirely in micronuclei, but sometimes reassembled chromosomes may return into the nucleus (Comai and Tan, 2019). Chromothripsis-like phenomena have also been observed in spontaneous somatic mutants of grape, where one inversion, one intrachromosomal translocation and four interchromosomal translocations, as well as the loss of several chromosome fragments where identified (see also section 7.4)(Carbonell-Bejerano *et al.*, 2017).

5.9 (Synthetic) polyploidy leading to genome rearrangements

Polyploidization (by whole-genome duplication; WGD) is quite commonly occurring in plants compared to mammals or even yeast (Panchy *et al.*, 2016) and can lead to structural variation. A particularly interesting case is the allopolyploids, arising through interspecific hybridisation. In allopolyploids, the parental genomes are recombining as in diploid genomes, maintaining the distinct genomes. As this results in increased copy numbers of genes and new combinations of genes, it leads to tensions in the genome, and most hybridisations will not be viable. This tension needs to be resolved, to keep the balance in gene expression and create a new equilibrium. At the same time, it makes new variation possible, meaning that in subsequent generations restructuring happens. This can often be reproduced to some extent in synthetically reconstituted allopolyploids. It has been done for breeding purposes in several species, e.g. tobacco (*Nicotiana tabacum)*, oilseed rape (*Brassica napus*), bread wheat (*Triticum aestivum*) and Triticale (hybrids of wheat and rye). This is probably a fast track to new combinations of genetic elements.

Polyploidization is probably a fast track to new combinations of genetic elements in plants

Among the structural changes are translocations between subgenomes. In wheat, reproducible loss and gain of DNA fragments occur in synthetic mimics with some resembling the natural allopolyploid. Such downsizing of the genome can be occurring relatively rapidly, i.e. in a few generations, mostly affecting repetitive sequences. Even aneuploidy is reported (i.e. losses and gains of chromosomes, e.g. in wheat). Homogenization of rDNA through recombination-based mechanisms and copy number changes can occur within a few generations as shown in synthetic mimics (e.g. allopolyploids in *Nicotiana*) (Renny-Byfield and Wendel, 2014).

Likewise, homeologous genes (homologous genes originating from different subgenomes) can be homogenised through gene conversion. This may sometimes happen with a bias towards one subgenome (e.g. in cotton). In tetraploid wheat, *Ha* gene deletion in the AABB genome has lead to hard-grained genotypes suitable for pasta; adding an intact *Ha* on the additional DD genome of hexaploid bread wheat has been leading to soft grains for dough. Later varietal development was accompanied by rearrangements at the *Ha* locus, often involving TEs (Renny-Byfield and Wendel, 2014). Also, R gene clusters can become differentially lost between homeologous chromosomes, generating recombinationally isolated loci (Ashfield *et al.*, 2012).

5.10 Conclusions

We have described mechanisms by which SV could be generated. Quite a few mechanisms have been described that vary in specific details. One can broadly distinguish mechanisms based on replication fork stalling and mechanisms based on repair of DNA double-strand breaks (DSBs). With replication fork stalling, a lagging DNA strand can invade other forks and so switch to another strand (template) for further replication. Depending on orientations, this can lead to insertions, deletions or more complex rearrangements. In special cases, inverted repeats may be the outcome. Likewise, DSB repair can produce insertions (duplications), deletions, inversions or translocations. These mechanisms are often facilitated by microhomologies at the sites of sequence exchange, which could be close to each other or more distant in

the genome, up to different chromosomes. DSB repair also regularly includes "filler" sequences. These can derive from close by DNA and also can involve microhomologies, but not necessarily so. The fillers can also be a more or less complex combination of various sequences from further away on the genome, also resulting in new combinations of genetic elements. Special cases are based on larger homologies, such as non-allelic homologous recombination (NAHR), which involves sequence exchanges between various sites on the genome, including unequal crossover in repetitive regions, and gene conversion, where homologous recombination leads to a change at the acceptor site but not at the donor site of the exchanged sequence. These homologous repair-based mechanisms can result in new combinations of genetic elements, such as promoters and genes. Complex rearrangements are also enabled by transposable elements (TEs).

As a result of various stresses and in vegetatively propagated crops, aneuploidy or chromothripsis may occur, where the latter particularly causes chromosomal rearrangements on a large scale. Structural variation may also occur through the production of extrachromosomal circular DNA, which, as will be shown later is involved in the rapid development of herbicide resistance under selective pressure. Polyploidy has occurred multiple times in Angiosperm evolution, allowing structural changes to occur in subgenomes. The synthesis of new polyploids during breeding can lead to similar changes.

6 The occurrence of intraspecies SVs

6.1 Introduction

In this section, an exploration of the occurrence of structural variations in eukaryotes, with an emphasis on plants, will be given. In addition, examples of SVs underlying specific phenotypes will be discussed. These have mostly caught attention in breeding programs because of the desirability of these phenotypes (and sometimes trade-offs with less desirable characteristics) (Gabur *et al.*, 2019). Mainly, for this reason, the molecular-genetic basis of the phenotype has been elucidated. The focus is on plant examples that address SVs developed from domestication onwards, preferably with a known history or at least found within a crop species. Frequencies of events occurring at the timescale of a few generations are addressed in a separate chapter.

We provide references to publications that contain data on the amounts of SV among individuals, varieties or accessions within a species. We will also provide frequencies of the various mechanisms underlying SV in plants or at least give an indication for these. Many studies in the literature quantify SVs that have accumulated over an evolutionary timescale. This ranges from comparisons between species, for instance assessing colinearity of genes between genomes of distantly to closely related species, to comparisons within species. For the purpose of this report, our focus will be on timescales starting from domestication, preferably from the breeding era, which involves studies of intraspecies comparisons. Delimitation is not straightforward, as domestication may be a lengthy process, sometimes involving new species formation or allopolyploids arising from several species. SVs found in a crop species germplasm can have arisen before domestication or somewhere along the road to modern varieties, and the time of origin can often not be pinpointed. The time of selection of a desirable trait for breeders will often be recorded, but the underlying SV mechanism will only have been resolved recently. It will be difficult to derive the time of origin from this; it may be estimated from comparisons between plant lines and a wild progenitor. However, it will only be apparent when it arose from particular selections or manipulations, such as selecting mutants from a pure line cultivar or from a clonally propagated cultivar. A confounding factor is introgression breeding, i.e. introducing desirable traits from wild relatives; along with the underlying gene(s), SV from the other species may come along, which likely has arisen at a longer (evolutionary) timescale than within species (see tomato example in section 4.5). The most relevant to our study is empirical data on generation-to-generation SV frequencies. These will be discussed in a separate chapter 9, although the number of well-studied examples of such very recent structural chances is low.

Table 2. Examples of the most recent reports on intraspecies SV in plants.

Species	***Genome sequence***	***# varieties[1]***	***Reference***
Cotton (allotetraploid)	de novo assembly	2	(Yang *et al.*, 2019)
Maize	de novo assembly	2	(Sun *et al.*, 2018)
Soybean	de novo assembly	2	(Valliyodan *et al.*, 2019)
Brassica oleracea	pangenome/ R genes	10	(Bayer *et al.*, 2019)
Arabidopsis	long read/NLR genes	40	(Van de Weyer *et al.*, 2019)
Rice	resequencing	3000	(Fuentes *et al.*, 2019)
Rice	resequencing	3010	(Wang *et al.*, 2018)
Medicago truncatula	de novo assembly	15	(Zhou *et al.*, 2017)
Arabidopsis	de novo assembly	2	(Zapata *et al.*, 2016)
Rice	resequencing	50	(Bai *et al.*, 2016)
Maize	long read/zein genes	3	(Dong *et al.*, 2016)
Cucumber	resequencing	115	(Z., Zhang *et al.*, 2015)
Potato (diploid)	resequencing	13	(Hardigan *et al.*, 2016)
Melon	resequencing	4	(Sanseverino *et al.*, 2015)

[1]as far as possible, when wild relatives were included in the study, only the cultivars of the same species were counted. Arabidopsis is not a crop, and thus has no varieties; in this context, 'ecotypes' are used.

Several reviews focus on the genome-wide presence of copy number variations and their origins (Zmieńko *et al.*, 2014; Saxena *et al.*, 2014; Zhang *et al.*, 2018) including in domestication (Lye and Purugganan, 2019) and disease resistance (Dolatabadian *et al.*, 2017). **Table 2** gives an overview of the more recent ones on intraspecies variation. Some of these, as indicated, focus specifically on the rich CNV and Presence-Absence variation in NLR genes.

6.2 SVs in Maize

Maize has a surprisingly large amount of SV

Early studies in maize looking at the colinearity of genes on chromosomes with other more or less related species already led to surprising findings of a large amount of structural variation. Such SV did not only break up colinearity of genes between different species (Bennetzen, 2000) but also within a single species. Complete segments containing several genes were found missing between inbred maize lines (Fu and Dooner, 2002). More recent work in maize confirmed this species' large variation, for instance, a 2.6 Mb region with 25 genes was missing in several maize inbred lines. Between two commonly used inbred lines, already 416 CNVs and 1783 PAV segments were identified (Springer *et al.*, 2009).

Though CNVs are naturally most prevalent in repetitive non-coding regions, many of them affect genes as well. Thirty-two per cent of the genes in the maize B73 reference genome was estimated to be impacted by CNVs. Eighty-two per cent of these CNVs were also present in teosinte; therefore a majority of SV was apparently preceding domestication (Zmieńko *et al.*, 2014). In 19 maize inbreds and 14 teosinte accessions, 479 genes with higher and 3410 with lower copy number than the reference genome were identified. Many CNVs were associated with domestication (Dolatabadian *et al.*, 2017). A comparison of two assembled inbred maize genomes revealed large-scale structural variation (Sun *et al.*, 2018). Approximately 10% of all genes were non-syntenic, meaning that they had different genome locations in the two varieties, around 3000-4000 genes in each were affected by structural variation, and each had 50 and 72 genes, respectively, that were not present in the other (Presence-Absence Variation).

An interesting example of an SV in maize is an inversion found to be causative to the Tunicate phenotype in maize (popularly called "pod corn"). Kernels are covered with long glumes instead of being naked as in normal maize ears, and extra branches develop producing irregular seed rows. Moreover, male flowers (tassels) become partly feminised and are developing a few kernels. The 1.8 Mb inversion affects a transcription factor gene, the MADS-box gene *Zmm19*, at the *Tunicate* locus. The inversion brings the *Zmm19* gene close to another promoter, resulting in a change in expression. Instead of in the leaf, the MADS-box gene is expressed in the developing inflorescence, with the aberrant flowering phenotype as a result. The inversion was followed by duplication at the breakpoint, in line with rearrangements' tendency of making genomic regions more prone to further SVs. Also, a Mutator-like TE is present at the breakpoint, which points at its involvement in the rearrangement (see chapter 4 for a more detailed discussion of TEs in SV). The rearrangement must have arisen after the domestication of maize, but pod corn was already known to pre-Columbian cultures (Han *et al.*, 2012).

6.3 SVs in Rice

In rice (*Oryza sativa*), CNVs were detected by comparative genomic hybridisation (CGH) in 20 Asian cultivars of different background: 2886 CNVs were detected that span 2.7% of the genome and overlap with 1321 genes. Specifically enriched areas probably originate from non-allelic homologous recombination (NAHR). Subspecies-specific variants are enriched for rare alleles, indicating their recent emergence (Yu *et al.*, 2013). Another study using WGS in 50 rice accessions found 9196 CNVs, containing 1117 entire genes (Bai *et al.*, 2016). A more recent study using data from 3000 rice genomes identified 63 million SVs

grouping into 1.5 million allelic variants, with long SVs particularly in promoters, and as many as 7% of genes being deleted in some varieties (Fuentes *et al.*, 2019).

Submergence tolerance in rice depends on SV in the *Sub1* locus. A subset of subspecies *indica* and *aus* rice varieties, as opposed to other *indica* varieties and all *japonica* varieties, is tolerant to submergence. These contain a specific copy of an *ETHYLENE RESPONSE FACTOR* gene, *Sub1A-1*, which arose from a tandem duplication of one of two pre-existing genes (*Sub1B*), while a diverged *Sub1A-2* variant does not confer tolerance. This gene duplication and divergence may have occurred either before or after rice domestication (Fukao *et al.*, 2009).

6.4 SVs in soybean, Arabidopsis

Compared to maize, SV appears to be lower in Arabidopsis and soybean. The sequence contained in all SVs among 80 widely distributed accessions of Arabidopsis equalled 2% of the entire reference genome (Zmieńko *et al.*, 2014). Comparison of a de novo assembly of Arabidopsis L*er* to the reference genome (Col-0) revealed 564 transpositions and 47 inversions, as well as 4.1 Mb of non-reference sequence. 105 single-copy genes were present in only one of the two genomes (PAV), and 334 single-copy genes had an additional copy in only one of the genomes (CNV) (Zapata *et al.*, 2016). These illustrate most of the events described in Fig. 3 of this report.

Soybean showed from 188 to 267 CNVs in pairwise comparisons among four diverse soybean genotypes in an array-based comparative genome hybridisation (CGH) analysis, and 672 genes were present in CNVs (McHale *et al.*, 2012).

Another interesting example of CNV was found in soybean cyst nematode (SCN) resistance that was improved by a higher tandem copy numbers of a 31 kb gene cluster at the *Rhg1* locus (up to ten tandem copies). The cluster contains three different genes, an amino acid transporter, an α-SNAP protein and a WI12 wound-inducible domain-containing protein; the resistance was apparently only enhanced when all three genes were expressed together (Cook *et al.*, 2012).

6.5 SVs in cucumber

In cucumber, a resequencing survey of 115 accessions by (Z., Zhang *et al.*, 2015) explored structural variation. They also inferred underlying mechanisms and discussed SVs underlying interesting phenotypes among which one was also mentioned in the interviews. Among the 26,788 SVs identified, were 19,168 deletions ranging from 50 to 16,600 bp, 7337 insertions from 50 to 400 bp, 205 tandem duplications from 340 to 155,800 bp, and 78 inversions from 180 to 40,500 bp, combined comprising about 12 Mb. More than 65% of SVs were present in two or more accessions.

According to the same study 1240 of the SVs involved the coding sequence of 1676 genes, with 130 genes deleted and 509 duplicated, and 27 in inversions. Various types of SV were enriched for categories of genes, e.g. deletions with abiotic stress responses and tandem duplications with reproductive and (a)bio stress. 943 SVs strongly differentiated cultivated from wild cucumber, with 70% fixed mainly in cultivated cucumber, indicating their involvement in its domestication (Z., Zhang *et al.*, 2015).

Mechanisms underlying the SVs were inferred based on certain assumptions about which mechanism causes which type of SV. The largest category was DSB repair by NHEJ, that is, without evidence of the involvement of (micro)homologies or TEs. A large part (about a quarter) was associated with DSB repair based on microhomology (MMBIR), and a relatively small part (0.7%) with NAHR possibly due to the low amount of interspersed repeats and segmental duplications. Another relatively small part (~4%) was associated with inverted repeats. Approximately 20% was associated with TE activity (see Chapter 4).

Interesting for breeding, cucumber fruits lacking spines are more attractive for large scale packaging for the consumer market. They arose from a complete deletion of *TU*, a zinc finger transcription factor gene.

A smaller deletion, in the untranslated part of a kinesin gene, is associated with a change in gene expression that can be related to larger leaf and fruit size as kinesin is involved in regulating cell division. (Z., Zhang *et al.*, 2015).

One particularly interesting CNV is a 30.2 kb duplication at the *Female* (*F*) locus. This locus is responsible for a plant having only female flowers (Z., Zhang *et al.*, 2015). Typically cucumber plants are bearing both male and female flowers. Having only female flowers is greatly increasing yield under high-input conditions in a greenhouse. As cucumber fruits for consumers have no seeds, commercial fruits develop without fertilisation (parthenocarpy), making male flowers superfluous. This *F* locus was bred into modern cultivars and was mentioned in the interviews with breeders. The underlying SV appears to be a duplication of the *ACS1* gene at the *Female* locus, an essential gene for the biosynthesis of ethylene. Strikingly, the additional copy of the *ACS1* gene has a truncated promoter which leads to a high level of expression of the gene (interview). This leads to higher ethylene production and a pronounced shift towards female flowers at the cost of male flowers. This duplication apparently made the region meiotically unstable, and led to further CNVs; for example, spontaneous back-mutations were found missing the duplication.

Duplication of the *FT*-locus in cucumber leads to more flowers. Wild cucumber plants make the first flowers after the formation of the 8th till 13th internode. Domesticated cucumber, however, flowers already after the creation of the first internode, leading to earlier fruit production and higher yield. Molecular analysis showed that this is a result of duplication of the *FLOWERING LOCUS T (FT)* (information from an interview). So, here again, the desired domestication trait arose by gene duplication.

6.6 SV in grapevine

Clonal propagation in grape probably causes the relatively high incidence of somatic mutations observed in this species. Grape is also an excellent example of a clonally propagated perennial crop, which is permanently and highly heterozygous, as opposed to too many inbred, homozygous annual crops for which most data on SV are collected (as in this chapter). A high-quality genome assembly of the Chardonnay cultivar allowed the detection of SVs between the homologous chromosomes (the so-called haplotypes), including of the high level of hemizygosity (the phenomenon of a gene copy being present on only one chromosome). This revealed 19,000 heterozygous SVs of over 50 bp between the haplotypes, being up to 5.3 Mb in length and together constituting 15% of the genome, and comprising 15% of the protein-encoding genes. A similar result was found for cultivar Cabernet Sauvignon, indicating that 1 in 7 genes in grapevine is hemizygous. Between the two cultivars, 60,000 SVs were detected (Zhou *et al.*, 2019). The effect of SV on grape berry colour is further discussed in section 7.4.

6.7 Adaptively advantageous structural variation

In many organisms, specific mechanisms for phenotypic variability have been described that are adaptive mechanisms for response to environmental changes at a timescale of several generations, a timescale between phenotypic plasticity through gene regulation and longer-term (evolutionary timescale) mutational processes. It may come with the disadvantage of having maladapted individuals in each generation. It is particularly occurring in microbes that are subject and more sensitive to variable environments than cells in multicellular organisms, and that may lack sexual cycles (e.g. some fungi). On the other hand, in multicellular organisms (both in animals and plants), genes involved in immunity are also more prone to variation in response to pathogens that may co-evolve.

6.7.1 Plant resistance gene clusters and adaptation

The most important class of resistance genes in plants are the *NLR*s, the nucleotide-binding/leucine-rich repeat genes. These resistance genes usually follow a gene for gene interaction pattern with pathogens, i.e. they respond to specific effectors secreted by the pathogen, such as the fungi and oomycetes mentioned in section 6.7.2 below.

Clusters of resistance genes in plants are hypervariable, making new combinations of genetic elements. This is likely due to unequal cross-overs

In a pangenome analysis on a representative set of **Arabidopsis** accessions, a total of 13,367 different NLRs were found that could be clustered into 474 groups of orthologous genes (genes that have evolved from a common ancestral species gene), with each accession having 167 to 251 genes (Van de Weyer *et al.*, 2019). In ***Brassica oleracea***, this number was 556 in total and 424-440 per line (Bayer *et al.*, 2019). The NLR genes often occur in clusters on the genome, in *B. oleracea* 68% of genes and in Arabidopsis accessions 47-71% of the genes. In ***B. napus***, across 50 lines, a total of 1749 resistance genes were identified, of which 996 are core (present in all) and 753 are variable (Dolatabadian *et al.*, 2019). The high copy number variation in these resistance gene clusters is likely due to unequal cross-over.

In the major cluster of **lettuce**, a part of the NLRs was highly variable due to gene conversion/intragenic unequal crossing-over between paralogous genes (genes that evolved after duplication in the same species), resulting in genes with varying numbers of leucine-rich repeats in these NLRs (Kuang *et al.*, 2004). The leucine-rich repeats are involved in pathogen recognition.

Resistance genes in plants also integrate domains from other genes, again making new combination of genetic elements

Another interesting phenomenon in NLRs is the occurrence of integrated domains (IDs) from other types of genes, such as genes encoding WRKY transcription factors, that may function as a decoy for effectors from pathogens. Usually, ~10% of NLRs have an integrated domain from other types of genes. In an analysis of nine grass species, repeated independent integration of IDs was more frequently observed in an NLR clade, showing 58% of ID-NLRs (Grund *et al.*, 2019). The latter authors also speculate about engineering NLRs with new IDs for broadening plant resistance. This would represent yet another interesting intragenesis approach to plant resistance breeding (further see section 1.5).

6.7.2 Rapidly evolving SV in plant pathogenic fungi and oomycetes

In fungi and oomycetes, rapid SV can be prevalent in regions containing, amongst others, effector genes that are involved in manipulating the host (see e.g. NLR genes in the previous section). The encoded proteins may be recognised by the host plant, leading to disease resistance. Therefore, these effector genes are under strong selection for variability to evade host recognition. These genomic regions are subject to extraordinarily frequent rearrangements and can take the form of complete, accessory (supernumerary) chromosomes, which is popularly described as "two-speed genomes" (Fouché *et al.*, 2018).

6.8 Conclusions

In spite of the limitations imposed by the use of short sequence reads for detection of structural variants, for crop species, some detailed explorations of SV have been performed. It revealed large variation, with a crop like maize having much SV across its genome, whereas a crop like soybean had less SV that was

also mainly found in a part of its chromosomes. Part of the SV could be inferred to have occurred after domestication and also a considerable part of SV affected gene sequences. The last years have seen *de novo* assemblies of several accessions within a species (cotton, maize, soybean), which greatly improves the detection of SVs, their types, and their effect on genes.

Structural variation often affects genes involved in stress responses of plants, a particular example being resistance genes of the nucleotide-binding/leucine-rich repeat (NLR) type. They often occur in tandem clusters that are subject to large variation through unequal recombination or gene conversion, which is functional in creating variation in the continuous arms race with pathogens. Pathogens, like fungi and oomycetes, and animal parasites and yeasts likewise have hypervariable genomic regions, at least in part to adapt to their hosts.

7 New combinations of genetic elements that arose spontaneously during cultivation or breeding

7.1 Introduction

As described in this report, quite a lot of SV has been described, leading to novel combinations of genetic elements, and new phenotypic traits. The far majority of these SVs did arise at an evolutionary time scale. However, there are also examples of SVs that have occurred spontaneously in plant varieties during the last hundred years, so at a breeding time scale.

We are aware that by far, most SV that has arisen since the last 100 years will have been unnoticed. Only a few that have led to striking, or beneficial phenotypic traits, have been studied and characterised at the DNA level. Knowing that we overlook nearly all new SV, we are holding back from substantial conclusions at this phase.

Despite our limited current knowledge about SV that has occurred within the plant breeding time scale, we can still provide several examples.

7.2 A striking effect of retrotransposon activity, about 55 years ago in an apple orchard in Canada

In 1964, a compact shoot was discovered on a McIntosh apple tree in a Canadian orchard. This particular shoot was clonally propagated by grafting. The resulting trees had a 'columnar' growth habit, i.e. with fruit-bearing spurs replacing branches. This led to very compact trees with fruits attached to the stem. This columnar growth of the McIntosh trees allowed a very high tree density, with higher yields, and possibly allowing harvest by robots in the future. Breeders also used this McIntosh mutant as a parent, resulting in several new columnar type apple varieties, that are currently sold around the world. All these columnar varieties harbour the specific mutation from the original branch from the 'McIntosh' tree, discovered in 1964.

The underlying dominant mutation (*Co*) appeared to be an insertion of a non-autonomous 8.2 kb LTR-retrotransposon (Wolters *et al.*, 2013; Otto *et al.*, 2014). Evaluation of the published genome sequence of the apple variety 'Golden Delicious' showed that many copies of similar transposons are present throughout the genome of apple.

Wolters et al. (2013) showed that the retrotransposon was inserted into another transposon. This induced upregulation of a putative 2OG-Fe(II) oxygenase gene, located 15.6 kb downstream the insertion, but surprisingly had no effect on two intervening genes. The mechanism of this upregulation and the molecular function of the upregulated gene giving rise to the columnar growth are yet unknown.

This example illustrates that retrotransposons are still active in commonly grown crops such as apple varieties, also in orchards.

7.3 Multiple gene copies, leading to herbicide tolerance.

In places in the world where Roundup has been used intensely to control weed, weeds that are tolerant to this herbicide have been selected. The active compound of Roundup is glyphosate, which has been used in non-selective applications since the 1970s, and in selective applications in glyphosate-resistant transgenic crops since their introduction in 1996. This compound binds to the plant enzyme 5-enolpyruvylshikimate-3-phosphate synthase (EPSPS), and blocking its activity prevents the synthesis of aromatic amino acids. However, several weeds have developed resistance to glyphosate, through developing different

mechanisms, among which increased *EPSPS* copy number (Gaines *et al.*, 2019). Two types of copy number increase have been observed:

In kochia (*Kochia scoparia*), from 3 to more than 10 extra EPSPS copies are arranged as a tandem gene duplication at one locus (Patterson et al., 2018). Apparently, the exceptionally high selection pressure leads to a high abundance of extra gene copies, leading to very high levels of the EPSPS protein. Presumably, the concentration of herbicide has been insufficient to block that large amount of EPSPS. Likely, this induction of structural variation has occurred only recently, i.e. since the increase of application of roundup following the introduction of GM glyphosate-tolerant crops.

In at least one case the resistance is correlated with the presence of extrachromosomal circular DNA (eccDNAs) containing extra copies of EPSPS (Koo, Jugulam, *et al.*, 2018; Koo, Molin, *et al.*, 2018). In the case of Palmer amaranth (*Amaranthus palmeri*), a section of >300 kb containing *EPSPS* and many other genes has been replicated and "inserted" at new loci throughout the genome, resulting into a significant increase in total genome size. The replicated sequence contained several classes of mobile genetic elements including helitrons, raising the intriguing possibility of extra-chromosomal replication of the *EPSPS*-containing sequence, using rolling-circle amplification (Fig. 5). The eccDNA appears to be tethered to chromosomes at various locations and is transmitted through meiosis and can be inherited.

EPSPS copy number increase to tens of copies, through yet uncharacterised mechanisms, has also been observed in four different grasses (reviewed in (Gaines *et al.*, 2019)

7.4 Dramatic rearrangements in grapevine varieties, turning red berries white

Dramatic rearrangements were shown to be underlying the production of white berries in Tempranillo Blanco, which is a spontaneous somatic variant coming from a bud from the grapevine variety Tempranillo Tinto, first published in 2006. The large-scale unbalanced genome reshuffling was appearing as chromothripsis-like, also described in section 5.8. As a consequence, the variety shows low gamete viability leading to lower fruit production. Berry colour is regulated by two MYB TFs in a tandemly repeated cluster on chromosome 2. The structural change that directly caused the white berry phenotype itself was a hemizygous deletion of 313 genes, i.e. in one of the two haploid chromosome sets. Among these genes is the only functional copy of an MYB transcription factor (TF) regulating the anthocyanin production in berry skin, VviMybA1, present in Tempranillo (Carbonell-Bejerano *et al.*, 2017). White berries in cultivar Chardonnay may also be attributable to a deletion of the same region on one chromosome (as compared to Cabernet Sauvignon), and a simultaneous inversion of the region on the other. In a further comparison of six closely related pairs of white and red berry varieties, a similar inversion was implicated in the difference in berry colour (Zhou *et al.*, 2019).

Retrotransposon insertion in the promoter of *VvMYBA1*, leading to loss of expression, is also implicated in the loss of berry colour in white varieties. In two independent red-skinned bud mutations of two white-skinned grape varieties, *MYBA1* expression was restored by the removal of the retrotransposon, probably through recombination between the LTRs (Kobayashi *et al.*, 2004).

7.5 Structural changes in the tomato genome

Some examples of characterised structural variation leading to new combinations of genetic elements in tomato literature are shown in Fig. 8. Intrachromosomal deletions between neighbouring genes lead to the *rin* (ripening inhibitor) (Robinson and Tomes, 1968) and *Mouse-ear* (Harrison, 1955) mutations. This *rin* mutation is a deletion between two genes with some homology (MADS-box genes *MADS-RIN* and *MADS-MC*) (Vrebalov *et al.*, 2002). A part of the first gene is fused with a part of the second gene, leading to the production of a chimeric protein with a novel function (Li *et al.*, 2018) (Fig. 8A). The homology between the two adjacent genes indicates that repair of a double strands break at this locus may have led to non-

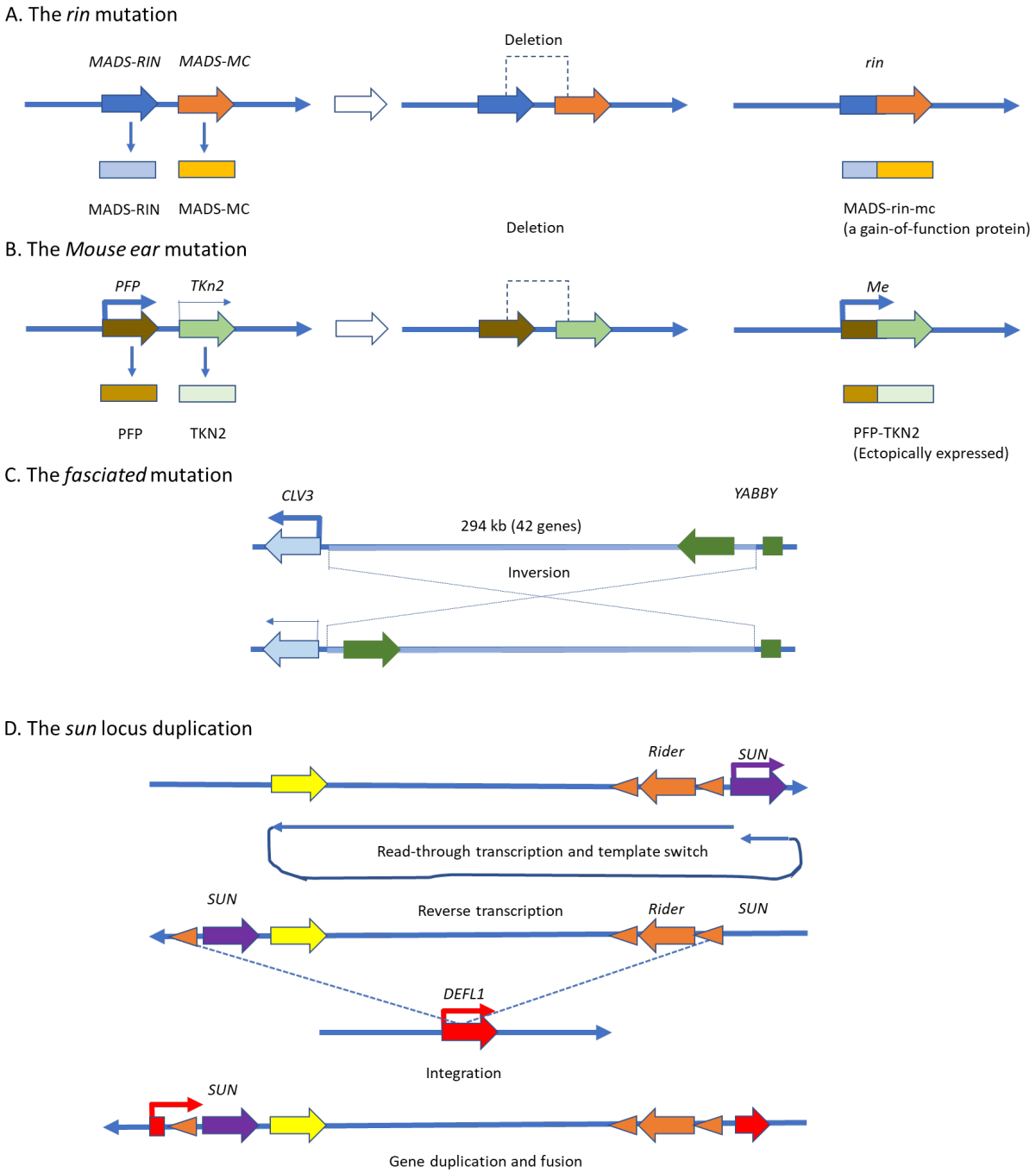

Figure 8. Examples of structural variation leading to new combinations of elements in tomato. A. *The* rin *mutation leading to the production of a gain-of-function fusion protein that actively represses ripening.* ***B.*** *The* Me *mutation leading to a leaf phenotype.* ***C.*** *The inversion in the* fas *mutant leading to multilocular beef tomatoes.* ***D.*** *The* sun *duplication leading to elongated fruit. See the main text for more details. Fig. 8D was adapted from* (Xiao *et al.*, 2008).

allelic homologous recombination (NAHR), although that cannot be established precisely. The *rin* mutation was detected in a cultivar, and must, therefore, be a recent mutation.Also in the case of *Mouse-ear* a deletion occurred between two adjacent genes, leading to the production of a chimeric mRNA. Probably, the second gene still leads to a normal Tkn2 protein, but with an abnormal expression pattern. This ectopic expression leads to deviating leaf shapes (Fig. 8B) (Parnis *et al.*, 1997). This mutation probably occurred in the early 1950's as it was first submitted to TGRC (Tomato Genomics Resource Centre, a seed repository) in 1955.

In tomato the *fasciated* mutation leads to multi-locular, large beef tomatoes. This mutation is characterised by the inversion of a 294 kb genome fragment, which replaces part of the upstream regulatory sequences of *CLAVATA3*, leading to lower expression and an increased apical meristem size (Fig. 8C) (Xu *et al.*, 2015). We do not exactly know when this inversion has occurred.

An example of gene transduction with interchromosomal gene duplication and fusion by an LTR-retrotransposon is shown in Fig. 8D. In the ancestral situation, a copy of LTR-retrotransposon *Rider* sits near the *SUN* gene on chromosome 10. According to the hypothesis, a complex combination of read-through and template-switching following transcription of the *Rider* copy occurred. This was followed by reverse transcription that produced a mobile *Rider*-derived TE incorporating 4 genes, including *SUN*. This subsequently inserted in an intron of the gene *DEFL1* on chromosome 7, thus producing a duplication of SUN into a new position under the regulatory control of *DEFL1*. This results in an elongated fruit phenotype (Xiao *et al.*, 2008). The mutation is not likely to be introgressed from wild germplasm. Therefore, possibly this mutation has arisen during cultivation of tomato, but we cannot pinpoint a time.

7.6 Conclusions

We have provided a few examples of structural changes that occurred during recent cultivation and breeding. Most structural changes will have remained unnoticed. However, the examples we have given here have led to pronounced phenotypes with a sometimes commercial advantage. Therefore, these have been studied in detail, and the underlying rearrangements elucidated.

The diversity of types of structural changes in these examples is striking: In apple a TE was active, leading to columnar growth, in weeds herbicide tolerance developed by means of the sharp increase in the copy number of a gene, and in red grape, berries turned white because of dramatic rearrangements of the genome. In tomato, a deletion led to delayed fruit ripening and another deletion to a change in leaf shape. Further, in tomato, an inversion led to large fruits, and a duplication led to elongated, rather than round fruits. So, a wide variation in types of structural changes, and also a wide variation of phenotypic changes. This underlines the current plasticity of plant genomes. For most examples, it can be traced that the structural change occurred during the last century, but we are not sure about this for all examples.

8 Conventional mutagenesis and tissue culture

8.1 Mutagenesis

8.1.1 Introduction

Plant mutation breeding is the practice of exposing plant parts (most often seeds or pollen) to a mutagenic agent, such as a chemical or radiation. The latter may be either non-ionizing ultraviolet light or, more commonly, **ionising radiation**(radioactivity) (Shu *et al.*, 2012). These mutagenesis methods are frequently used in plant breeding and, considering their long history of safe use are exempt from European regulations for approval of genetically modified plants. It is therefore relevant for the purpose of this report to assess their effects in terms of Structural Variation as a possible contribution, besides natural variation to the baseline for intragenic crops. The impact of the mutagenesis is almost immediate upon treatment with mutagens, and therefore the frequency of SV occurrence per generation is more straightforward to establish than that of naturally caused SV inferred from interspecies or intraspecies comparisons. This section deals with conventional chemical and physical mutagenesis methods and assesses the nature of structural variation caused by these. Although **transposon-mutagenesis** is sometimes considered a form of mutation breeding, we do not include it here but in Chapter 4.

Mutation induction in crop plants started in the 1920s using X-rays, and from the 1930s systematic mutation breeding programs were set up in various places. Gamma radiation has become a popular mutagen since the 1950s, and multiple forms of neutron radiation have been used since the 1960s-70s. In 1964 a Joint Division from the FAO and IAEA for mutagenesis breeding was initiated. As one of its results, the FAO/IAEA Mutant Variety Database now lists 3200 officially released mutant varieties from 214 different species (https://mvd.iaea.org/, accessed November 21, 2019; (Ahloowalia *et al.*, 2004)).

Today, there are >3200 officially released mutant varieties from 214 different species

8.1.2 Agents of mutagenesis and their genomic effects

DNA-alkylating agents such as EMS (ethyl methanesulfonate), ENU (1-ethyl-1-nitrosourea), and MNU (1-methyl-1-nitrosourea) are the mutagens that are most commonly used in chemical mutagenesis. Mutagenic effects include single base conversions, as well as deletions, insertions, inversions and DNA breaks (reviewed in (Leitão, 2012).

Different varieties of non-ionising and ionising radiation have been employed for mutagenesis. Fast neutron and Gamma ray-irradiation are the most frequently used methods used for physical mutagenesis. Although most of the mutation methods produce small mutations, such as nucleotide substitutions and small indels, which are outside the scope of this report, there are also many examples of more substantial structural variations such as deletions, inversions, and translocations. In particular fast neutron mutagenesis yields deletion mutations in plants, as the name "Deleteagene" for one high-throughput gene knockout method suggests (Li and Zhang, 2002). In Arabidopsis these deletions were found to be from a few bp up to 30 kb (ref). In another study deletions in the Arabidopsis genome were shown to be up to 12 kb (Li *et al.*, 2001). A

Mutagenesis has not only been used to induce base substitution, and small or large deletions, but also to translocate a (part of a) chromosome from a wild donor onto a chromosome of commercial varieties

repeat-rich fragile site in chromosome 4 of Arabidopsis is the starting point of deletions of up to 925 kb that are caused by mutagenesis methods, such as fast-neutron irradiation, but also X-rays and even EMS. This suggests that some sites are more vulnerable to recurring deletions caused by mutagenesis and may be associated with karyotype evolution, the rearrangement of chromosomes during evolution of species (DelaPaz *et al.*, 2012).

In plant breeding, radiation has also been applied to promote translocations that stabilise introgressions in complex interspecies hybridisations (in other words, radiation is used to translocate a (part of a) chromosome from another species into a host chromosome, to keep it stable there). This is especially useful for the stable introgression of clusters of disease resistance genes that lie on specific chromosome arms that can be added onto a host chromosome. This approach has yielded several translocations, some of which were useful for commercial applications. In this way, a chromosomal segment with a rust resistance gene was translocated from an *Aegilops umbellulata* chromosome to wheat, leading to a stable resistant wheat line (Jauhar, 2006).

8.1.3 Frequencies of characterised genome-wide structural variation induced by mutagenesis

Although many plant species have been subjected to chemical or radiation mutagenesis, including **maize** (Nannas and Dawe, 2015), soybean, and cotton, few have been subjected to high-throughput whole-genome sequencing or comparative hybridisation to assess the effect on structural variation.

A single recent, informative review summarises the frequencies of varying types of mutations from EMS-mutagenesis (one study in tomato) and radiation mutagenesis (Jo and Kim, 2019). This frequency was based on 11 whole-genome sequencing studies, mostly in Arabidopsis, with two in tomato, and one in rice, and compared to the rate of spontaneous mutation in Arabidopsis. Mutation frequency, measured per base pair of genomic sequence, ranged from 31-91 times that of spontaneous mutations in Arabidopsis for radiation, and 400 times the background for EMS-treated tomato. EMS-mutagenesis in tomato resulted in virtually exclusively (99%) single nucleotide changes (not relevant for this study). Various types and doses of radiation resulted in considerably more deletions and insertions (up to 50%), of which most were deletions smaller than 100 bp, and up to 12% larger deletions and insertions, reflecting the distribution (but not the frequency) of the background mutations. Large-scale DNA rearrangements were analysed in only four studies, probably due to the use of short sequencing reads. The proportion of large-scale SVs varied from 9% in rice to 4-27% in Arabidopsis. The complexity of rearrangements varied and was highest for argon beam irradiations with multiple rearrangement events and occurrence of many filler DNA sequences. When focussing further on large structural rearrangements caused by radiation, using genomic hybridisation or whole-genome resequencing, the information is limited to two publications in Arabidopsis and four in rice (Jo and Kim, 2019).

In rice, resequencing of 1504 mutant lines following fast-neutron irradiation revealed a frequency of 61 mutations per line, with 50% consisting of deletions (of which 10% > 100 bp), insertions, inversions, translocation and tandem duplications.

In **Arabidopsis**, a combination of comparative genomic hybridisation and resequencing was used to characterise radiation-induced mutations. Array-CGH identified 1.9-3.7 deleted areas per genome, with size ranging from 149 bp to 602 kb. Importantly, 81% of the deletions appeared to have genomic rearrangements. Three mutants were resequenced, revealing 22 fragments contributing to chromosomal rearrangements, of which 19 were intrachromosomal, and three included interchromosomal rearrangements (Hirano *et al.*, 2015). In an in-depth analysis of 16 Arabidopsis mutants derived from two types of heavy-ion irradiation experiments, an average of 2.3 and 10 rearrangements (>100 bp) per genome was found (Kazama *et al.*, 2017). These included intrachromosomal deletions, inversions, and as well as interchromosomal translocations. These were apparently produced after DSB and repair and often contained filler DNA with no obvious host genome origin.

A comparison of fast-neutron derived mutants of **soybean** with transgenesis-induced variation or intraspecies natural variation (Anderson *et al.*, 2016) also allowed some inference of radiation-induced change versus already present variation (of 40 analysed soybean cultivars). In the cultivars without mutagenesis, 156 to 362 genes overlapped with deletions and 45 to 124 genes overlapped with duplications in pairwise comparisons. For mutagenised plants, these numbers were 0 to 290 genes overlapping with deletions and 0-2312 overlapping with duplications.

8.1.4 Conclusions

Mutagenesis breeding, and in particular radiation mutagenesis (fast-neutron and heavy-ion irradiation) dramatically increase the genome-wide frequency of large-scale inter- and intrachromosomal rearrangement, (chromosomal translocation) probably through an increase in the incidence of DNA double-strand breaks and rearrangements following repair. This can go as far as stably adding parts of chromosomes from wild donors onto host chromosomes if the two species can be hybridised. Together with the appearance of apparently random filler DNA in the repair junctions, this type of mutagenesis is probably a powerful generator of new combinations of coding sequences, of promoter sequences, or of mixtures of both.

Radiation mutagenesis is probably a powerful generator of new combinations of genetic elements that are already present in a plant

8.2 Tissue culture

8.2.1 Introduction

Somaclonal variation is the term for variation in genotype and phenotype in plants originating from a shared mother plant and propagated in tissue culture or other forms of vegetative propagation. Clonal propagation by tissue culture (also known as micropropagation) is used for fast multiplication of new varieties of crops such as apple, strawberry, papaya, banana, grapes, pineapple, citrus, oil palm and others (reviewed in (Krishna *et al.*, 2016)). This preference may stem from the lack of propagation by seed in parthenocarpic varieties (banana) or the heterozygosity of crops like apple and strawberry. Since micropropagation usually is utilised for mass production of uniform plants, somaclonal variation is considered a nuisance, such as for the "mantled flower" phenotype and resulting yield loss in oil palm (see chapter 4). Alternatively, sometimes somaclonal variation in vegetative propagation yields favourable phenotypes, such as the columnar growth habit in apple or novel grape varieties (chapter 8). Somaclonal variation occurs through either base deletion or substitution, hyper/hypo-methylation of DNA, changes in chromosome number, or chromosomal rearrangements (various forms of structural variation, including through transposon activation) (Krishna *et al.*, 2016).

A large amount of observed variation is attributed to epigenetic modifications (Kaeppler *et al.*, 2000), and indeed the mentioned "mantled flower" phenomenon is one. We consider epigenetic changes leading to altered gene expression outside the scope of this report, except where these result in increased transposon activity.

Somaclonal variation at the DNA level also occurs in human somatic tissues (O'Huallachain *et al.*, 2012) and is prevalent in various human cancers (Yi and Ju, 2018). For plants, some information on small DNA sequence changes is available, but much less information is available on more substantial structural variation as a result of tissue culture.

8.2.2 Structural variation resulting from plant tissue culture

High-throughput sequences technologies, let alone the newest long-read technologies, have so far only scarcely been applied to somaclonal variation from tissue culture. Fossi et al. (2019) investigated large-scale genome changes in potato plants regenerated from protoplasts or from stem explants and compared those with plants propagated from cuttings by whole-genome resequencing and inspecting genome-wide sequence coverage (Fossi *et al.*, 2019). In contrast to the latter (cuttings), both types of regeneration (protoplasts more so than explants) resulted in large scale chromosome changes, such as segmental deletions and duplications, from one to many per plant, and even so the phenotype was apparently normal. Control plants propagated by cutting showed none of these chromosome changes. The authors speculate that this high frequency may be tolerated in regenerated potato, which is tetraploid and where large scale deletions may be buffered by the available additional chromosome copies. This hypothesis awaits confirmation by a more comprehensive study of such rearrangements in tissue culture.

8.2.3 Transposon activation by tissue culture

TEs may be activated by stress, which includes the stress derived from tissue culture, as was first discovered in maize (McClintock, 1984; Wessler, 1996). See Chapter 4 (Transposons).

8.2.4 Conclusions

Data on Structural Variation caused by tissue culture are scarce. This phenomenon may be especially relevant for tissue culture of polyploid crops as these could be more tolerant to the loss of large parts of one subgenome where the remaining subgenome copy compensates for this loss.

9 Frequency of spontaneously occurring SV

9.1 Introduction

This chapter will try to summarise the available literature that reports frequencies of Structural Variation occurrence on a generation time scale in crops. As we will show below, the amount of available, reliable data is not very high. We speculate that there are three reasons for this:

1. The frequency of spontaneously occurring structural changes is usually low unless radiation is applied.
2. For estimations of reliable per-generation frequencies, plant lines which are genotypically uniform at the start should be grown over several generations in a traceable manner. Such lines are called **Mutation accumulation (MA) lines**. These projects are challenging to perform with plant species that have a long generation time, as most projects will last just a few years. Thus it is not surprising that most of this type of data comes from Arabidopsis MA lines.
3. Current detection strategies for SVs are based primarily either on Comparative Genome Hybridization to determine Copy Number Variation or on whole-genome resequencing for comparison of intraspecies variation. The first method has a low resolution and does not give information about the type of SV, or its borders ("breakpoints"). Thus it may also report multiplication or loss of TE sequences. The second method uses short sequence reads and is prone to high false-positive and false-negative results (see chapter 3).

This chapter focusses on structural variation in general or by TE activity in particular detected in one or a few generations. Frequencies of induced mutations, either through Transposon mutagenesis (Chapter 4) or conventional mutagenesis with chemicals or radiation (section 9.1) are discussed elsewhere in this report.

9.2 Quantitative data

The technology for detecting and comparing structural variation is continuously being improved, with the human genome and cancer studies taking the lead. However, even there most information comes from the comparison of multiple unrelated individuals, such as in the 1000 Genomes Project, and few compare parents with children to estimate per-generation frequencies. Previously the rate of the TE *Alu de novo* insertions was calculated to be one new per 20 births (Cordaux *et al.*, 2006). In a later study, comparing an African parent-child trio, no new insertions were found with more strict algorithms although previous algorithms predicted 410 or 20 *de novo* insertions, respectively for different algorithms (Hormozdiari *et al.*, 2011). This clearly highlights the difficulty of accurate prediction of novel events purely based on short sequence reads. Comparison of a European parent-child trio in the same study for deletions (>100 bp, and < 1Mbp) with the most conservative prediction tool gave 39 *de novo* events for the child.

9.2.1 Mutation accumulation lines

Species with short life cycles, such as Arabidopsis, or even more so, the flagellate unicellular green alga *Chlamydomonas reinhardtii*, are especially useful for the study of Mutation Accumulation (MA) lines. This has been done for Arabidopsis in several studies, as well as in *C. reinhardtii*.

CNVs that emerged in Arabidopsis over 5 generations starting from siblings subsequently undergoing different temperature (16 or 28 vs 22°C as control) or stress (Salicylic Acid vs control at 22°C) were assessed with CGH. 38 repetitive (occurring in 2 out of 3 siblings) CNVs (26 unique) distributed over all 5 chromosomes were identified at the end of the experiment (Debolt, 2010). These were all microdeletion events and mostly shared, in 16 vs. 22 (control) degrees and SA vs control, but much more unique and variable between siblings and also involving duplications at 28 vs 22°C. The number of genes contained in

such CNVs ranged from 2 to 128 (3 to 300 kb), and a total of 292 to 402 genes detected per growth regime.

From 5 resequenced descendants of a mutation accumulation line of 30 generations, several insertions and deletions were identified with an estimate of $0.5x10^{-9}$ per generation for deletions larger than 3 bp, with an average of 800 bp (with high variation) deletions per event. The sequencing information from this paper can probably yield more precise and validated estimations of different events, but these were not produced here (Ossowski *et al.*, 2010).

A very recent report of a 25-generation MA line and sequencing data from 107 descendants reported 7 deletions larger than 100 bp, of which four overlap open reading frames, and no insertions larger than 100 bp (Weng *et al.*, 2019). Although not strictly an MA line, the observation that North America was colonised by a single Arabidopsis accession 400 years ago, and the availability of both historic (herbarium) as well as modern samples allowed to set up a pedigree of these lines (Exposito-Alonso *et al.*, 2018). Although in this publication only single-nucleotide polymorphisms and indels were analysed, a future analysis may be able to say more about the SV between these lines and allow a (rough) estimate of the per-generation frequency of SV generation.

9.2.2 CNV creation during meiosis

The reshuffling of copy number variants of two parents during meiosis in their offspring may create new CNVs. In humans**,** the frequency of *de novo* creation of large CNVs (>100 kb) is about $1.2\ x10^{-2}$ per genome per generation (Itsara et al. 2010), and another study reported that most of over 4000 CNVs analysed had individual rates of approximately 10^{-5} per generation (Fu et al. 2010).

A particular case is the study of genome-wide variation by resequencing of the parents, and the progeny of a hybrid Columbia(Col) x Landsberg *erecta* (L*er*) of Arabidopsis cross using the *qrt1* mutation to produce progeny with undetached pollen from single meiosis events (tetrads) (Lu *et al.*, 2012). Although primarily used for detection and characterisation of meiotic crossovers, analysis of CNV between the two parents as well as the creation of new CNVs during recombination in the hybrid could be observed. The F-box and (as to be expected) NLR families were highly enriched (22/656 and 8/147 genes, respectively) in the 316 large deletions or insertions. After meiosis, the two sets of four meiotic products showed 21 and 32 new CNVs, respectively. Since this was comparing parents of the hybrid with its progeny, these numbers are the sums for two generations.

9.2.3 Transposon activity

Maize

In the original publication on mutable loci in maize (McClintock, 1950), McClintock observed a sharp increase in mutable loci in the progeny of maize lines subjected to X-rays, leading to a rearranged chromosome 9 (McClintock, 1941). The "controlling elements" causing the mutations consisted of the autonomous *Ac* (*Activator*) and the non-autonomous, *Ac*-dependent *Ds* (Dissociation), where the latter is an internal deletion mutation of *Ac*. Although according to the publications mentioned here and those following them, the frequency of de novo insertion and excision is high, few exact data are available from the earlier studies. *Ac/Ds* is a relatively low-copy number TE system, and about 2-4% of progeny receive a newly transposed *Ac* or *Ds* copy (Brutnell and Dellaporta, 1994).

Mutator is a relatively high-copy number, active TE system leading to high amounts of new mutations (from 10-50% of all ears of the progeny of selfing containing visible mutations) (Walbot, 1991). Furthermore, the use of *Mutator* for the creation of a mutagenic inbred maize population led to 5-9% independent seed mutations per generation (McCarty *et al.*, 2005)

As was reported in section 4.4.1., in maize (*Zea mays*), it was shown very recently that many deletions, as well as insertional mutants, can occur by activation of lineage-specific retrotransposons specifically in male gametophytes of some varieties, and is observable within one generation at a frequency of $\sim 4x10^{-5}$

at the single *Bronze* (*Bz*) locus. Besides, screening for new insertion sites of three retrotransposons in a 1000-seedling pool revealed 18 to 300 new insertion sites (on average 0.4 per seedling), of which more than half in genes (Dooner *et al.*, 2019). This publication also links "sloppy" retrotransposition to the occurrence of deletions. Apparently, in maize, the early identified DNA transposons are the most active, but only in some lines and not, for example, in the 4 lines of the Dooner study. There, new retrotransposon activity was detected. The former study also exemplifies the differences in the more easier observable Bz phenotype giving a transposition frequency for this specific locus, and the more reliable but resource-demanding whole-genome sequencing approach leading to an overall frequency. This also nicely shows that a lot of SV creation goes unnoticed when not readily observed by visible phenotype.

9.2.4 LIS-1 in flax

In flax, nutrient stress-induced structural variation occurs in the form of a lower copy number of ribosomal DNA (rDNA). Also, insertion events in all 15 chromosomes, among which a 5.8 kb insertion called LIS-1 (Linum Insertion Sequence 1) were found at the same position in five independent lines (Y., Chen *et al.*, 2005). LIS-1 consisted mainly of a complex set of short repeats in two orientations, and these repeats were already present in the flax genome. Furthermore, at both breakpoints, a TCC duplication was present, and 3792 bp of flanking sequence contained 129 SNPs and indels as compared to the progenitor; otherwise, no characteristics for TEs were identified. We suggest that this may also be an example of DSB repair with filler (in this case LIS-1 repeats).

9.2.5 NLR resistance gene clusters

Concerning the highly variable NLR genes (resistance genes) described above, 11 of 12 losses of a particular NLR (*Dm3*) in the lettuce major cluster (RGC2) that was found among 11000 S2 and 16500 F1 plants analysed was due to deletions at the RGC2 locus. Also, 4 of 167 recombinations identified from ~2220 F2 were leading to CNVs in the RCG2 region (Chin *et al.*, 2001; Kuang *et al.*, 2004). Therefore, although recombination frequency was relatively low at the RCG2 cluster, such events may occur in each generation as lettuce plants produce thousands of seeds. Also, every 3-4 plants would be creating one chimeric gene every generation (Kuang *et al.*, 2004).

9.3 Conclusions

Reliable data on SV occurrence over one or several generations are scarce, as is shown in this chapter. There is probably much more SV generation going on than we can currently see, due to the paucity of data and the still imperfect characterisation methods. Nonetheless, in a synthesis of all reported data from intraspecies SV (Chapter 6), Transposon activity (Chapter 4) and per-generation frequency, with all plant data combined and observations from other eukaryotes as an indicator, we believe that it is possible to make a putative ranking of SV types and their frequency, from high to low. At the top and most frequent, probably occurring in large part of the individuals of a generation is non-allelic homologous recombination (NAHR), such as unequal crossovers, especially in highly repetitive clusters of NLR resistance genes (section 6.6). If this recombination occurs between NLR genes, this produces new combinations and possibly new resistance genes. If it occurs between intervening TE sequences, it will change copy numbers but less likely create unique combinations. NAHR outside these clusters, which because of long partial homology stretches (tandemly organised NLR genes) may align easier, is probably much less frequent but may have more dramatic effects in terms of combining different elements. Transposon mobility may have similar or slightly lower frequencies than NAHR in NLR clusters, but a far higher rate than the next mechanisms, deletions and inversions involving DNA double-strand breaks. The impact is probably highest where it concerns knocking out genes by transposon insertion, followed by expression changes. Combination of new elements through the mobilisation of genes has been observed with relevant phenotypes (see the example of the *SUN* locus in section 8.5) is probably less frequent. Combinations of exons by Pack-mules may take even longer to produce expressed new genes.

It should be noted that TE activity may vary significantly between accessions of the same species (Chapter 4). This is based on whether an individual inherited a copy of an active, low copy TE or possibly whether silencing mechanisms have been activated or inactivated through a random process or stress. This synthesis produced the hierarchy of relative frequencies as shown in Chapter 8.

Two methods of mutagenesis, one relatively old (conventional mutagenesis) and one new (transposon mutagenesis) can overturn this hierarchy. As we have seen in section 9.1, particularly fast neutron and ion particle radiation types dramatically increase large deletions, inversions, and translocations. This is probably through their high DNA-breaking capacity, producing this SV at a far higher rate than natural occurrences, and within one or a few generations. This type of SV probably has the highest impact through production of relatively random combinations of genetic elements on the same chromosome or through interchromosomal translocations. It might be argued that such dramatic SV will be negatively selected for during propagation through seeds, but this is true for the more dramatic natural occurrences as well. Transposon mutagenesis through activation of usually silent TEs or increasing TE mobility by applying stress can dramatically increase the incidence of insertions and even of mobilisation of endogenous gene sequences too. And so, the baseline moves upwards.

10 Concluding remarks on frequencies of structural variation

10.1 Introduction

In the current chapter, we try to describe a baseline for structural variation. Ideally, the structural variation that arose spontaneously during the past decades would be the appropriate basis for defining the baseline. However, we are aware that the far majority of recent (< 50 years) structural changes in genomes remain unknown thus far. To slightly compensate for this underestimation of the current structural dynamics of genomes, we also consider to some extent the structural variation that is present within species and arose during a more extended period of several thousands of years.

The scarcity of information about frequencies of different types of structural changes that have arisen spontaneously during the last ~50 years, makes us aware that it is not feasible to describe a very precise baseline. Clear limits of new combinations of genetic elements that are below this baseline, and of those that are above this baseline, and therefore are very unlikely to arise spontaneously during conventional breeding or in nature, are difficult to determine. However, having stated this caveat, below, we describe a series of structural changes that occur in living organisms, especially for plants, which can be relevant for intragenic crops. We describe these changes in a deliberate order, starting with new combinations of genetic elements that are more likely to occur, and continuing with structural changes that are more and more unlikely to occur spontaneously.

The chapter finishes with a schematic overview of these structural changes, and their positions relative to the baseline. Exact positions cannot be determined, but we give indications. The list below is not exhaustive. As explained in the previous chapters, far more kinds of rearrangements did occur in nature and in plant varieties. However, the list below summarises the most predominant structural changes, and their position compared to the baseline.

10.2 Types of SV, and their positions regarding the baseline

1. Exchange of allelic genetic elements

Especially during meiosis, when homologous chromosomes pair, crossovers and gene conversions between alleles occur, leading to new combinations of genetic elements. For example, consider a gene *A* that has two alleles, i.e. *A1* and *A2*. During meiosis, recombination may occur between the promoter and the coding sequence, leading to a new combination of genetic elements, such as the promoter from *A2* with the coding sequence of *A1*. This also holds for coding regions, e.g. leading to protein domain swaps. As the order of these elements stays the same, this is not a structural change. In spite of this, we still mention this frequently occurring phenomenon, as it is relevant for the description of the baseline of intragenics. This type of new combinations of allelic elements frequently occurs during breeding and in nature and therefore can certainly be regarded as belonging below the baseline (Fig. 9).

2. Exchange of genetic elements within a cluster of tandem repeats

During meiosis, homologous chromosomes pair. When clusters of tandemly repeated genetic elements pair, one chromosome may shift a bit compared to the other chromosome, leading to the pairing of homologous but non-allelic elements of the cluster. If a crossover or a gene conversion occurs there, this causes a so-called 'unequal crossover' or 'non-allelic homologous recombination' (NAHR), as explained earlier. This leads to a new combination of genetic elements, such as a promoter of gene *A* with a coding sequence of a downstream or upstream gene *B.* Different kinds of new combinations of genetic elements within such a cluster can occur spontaneously. This phenomenon is especially known for clusters of disease resistance genes but holds also for other tandem repeats (paralogous genes, transposable elements). The order and orientation of the elements remain the same, but the number of elements may grow or shrink, leading to copy-number variation (CNV). Also, domains of these elements may be exchanged, even if they

are not strictly allelic. This type of new combinations of genetic elements can still be regarded as being below the baseline of structural variation.

3. Deletions

As described in the previous chapters, deletions frequently occur during breeding and in nature, including large deletions (>100 bp). This may lead to a new combination of for instance a promoter of gene *A* and a coding sequence of a downstream gene *B*. It may also lead to a chimeric gene, such as the first part of gene *A* merged to a remaining part of the downstream gene *B*. A restriction on this new combination of genetic elements is that the elements were already present on the same chromosome. Moreover, they stay in the same order on that chromosome. This type of new combinations of elements occurs rather frequently during breeding and in nature (Fig. 9).

Repair of deletions may also be accompanied by the inclusion of genomic sequence from the vicinity of the breakpoint and occasionally from unrelated regions of the genome in a seemingly disordered mosaic, so-called 'filler DNA'. Although its consequences have been observed in extant plant accessions, we have no information on the per-generation frequency of this phenomenon. Therefore it is not possible to say how this affects the 'baseline'.

4. Translocations of transposable elements (TEs)

Transposable elements (TEs) are probably most often translocated to new positions in the genome (Chapter 4). This transposition leads, by definition, to new combinations. The far majority of TEs in plants is no longer active, but as we have shown in chapter 4, some TEs are currently active or become so under stress conditions. TEs have caused very prominent phenotypes during breeding and agriculture in the last few decades (Chapters 4 and 8). We expect that far more TE-activity will be detected in the coming years when long-range sequencing methods will be more adopted. Therefore, we regard the translocations of TEs as belonging to the baseline.

The position of TE activity becomes less clear when some additional discoveries are considered:

- Genomic DNA, including exons or whole genes in the vicinity of LTR-retrotransposons and Helitrons, may "hitchhike" with the transposable element. This happens in case of transcriptional readthrough or template switching, leading to a transcript that does not only contain the TE but also other DNA.
- Pack-MULE type of TEs may incorporate exons or whole genes (see section 4.7) Such gene (parts) can then be re-integrated together with the TE at a new genomic location. When inserted into a gene, these elements may be included in the transcript of the gene at that location, effectively creating a new gene.
- Excision of class II transposons (the cut & paste types) may cause DSBs that can lead to genome rearrangements, such as inversions, translocations and large deletions, especially when two neighbouring transposons 'jump' simultaneously. Repair of the DSB or deletion may also lead to the inclusion of filler DNA from elsewhere of the genome.

5. Inverted repeats

In the examples of intragenic plants described in the introductory chapter, RNAi constructs harbouring one or more inverted repeats were inserted. The DNA elements included were taken from the host plant. When such an inverted repeat is transcribed into RNA, the transcript can fold into double-stranded RNA with a hairpin structure. This double-stranded RNA will subsequently be processed into 21 nt siRNAs by the plant, triggering the degradation of the transcript of the targeted gene. We wondered if such inverted repeats would frequently arise in conventional breeding and in nature. We did not find documented examples of such inverted repeats that arose spontaneously during the last few decades in plants. This indicates that this structural change is not occurring frequently or that there is a strong purifying selection (negative selection) for it (at least for those with homology to any mRNA) as it would also silence the expression of the original and other homologous gene copies. However, inverted repeats are present in genomes, as appears from a study in soybean by (Tuteja and Vodkin, 2008). This indicates that they are created at an

evolutionary timescale, anyhow. Given the commonly occurring inverted repeats in eukaryotes, and the underlying proposed mechanisms, the generation of new inverted repeats may be regarded as a natural process, that may also occur during breeding. However, the frequency is likely to be lower compared to the other structural changes mentioned above. Therefore, we are not sure if the creation of inverted repeats should be regarded as belonging to the baseline of spontaneously arising SV during the relatively short timeframe of breeding. During evolution, they may have been created abundantly, and according to one hypothesis have given rise to gene expression regulation involving siRNAs or miRNAs (Allen *et al.*, 2004).

6. Translocations of non-TEs

Although we mentioned that translocations of TEs do belong below the baseline, this does not imply that translocation of any arbitrarily chosen DNA fragment belongs under the baseline as well. Translocation of DNA that is not a transposable element (non-TE) is far less likely to happen spontaneously. Therefore, this should probably not be regarded as being under the baseline of conventional breeding and nature.

However, there are examples of translocations of non-TE DNA using irradiation that did occur during conventional mutagenesis breeding. Mutation breeding in wheat led to translocation of a pest resistance gene in wheat varieties that are commonly grown and used in food (section 9.1). If mutagenesis is regarded as a method of conventional breeding, then translocations of non-TEs can be regarded as being below the baseline. However, if conventional mutagenesis is not regarded as being a traditional plant breeding method, then the creation of new translocations of non-TEs likely belongs above the baseline. We fully realise that translocations have occurred abundantly during evolution, and therefore translocation events of non-TE do happen in nature and in cross-breeding, albeit at a very low frequency, as far as we know.

7. Inversions

New inversions do occur at a low frequency. At an evolutionary timescale, inversions can be discovered in many genome positions. However, the frequency of spontaneously arising inversions is too low for regarding the creation of new inversions as belonging to the baseline of cross-breeding and nature.

Conventional mutagenesis by irradiation, causing many DNA breaks and repairs, may result in inversions in a short timespan, i.e. directly after the treatment. Recent reports have shown that if two DNA breaks occur in a chromosome, the DNA fragment between these two breaks may be ligated in the opposite orientation, leading to inversion at a frequency of a few percents (Schmidt *et al.*, 2019; Zhang *et al.*, 2016). If mutagenesis is regarded as a conventional plant breeding method, then the creation of inversions can be regarded as being below the baseline.

8. Complex rearrangements in intragenic example plants

In Chapter 1, we describe two examples of intragenic plants which contain complex DNA-constructs. E.g. the inserted construct in the 'Innate' potato from Simplot encompasses 22 genetic elements, taken from all over the potato genome, and from a wild relative of potato. We regard it as extremely unlikely, although not impossible given what is written earlier in the report, that such an event would occur spontaneously during plant breeding.

The virus-resistant tomato lines from Australia have fewer genetic elements, yet the elements were taken from more than 10 different loci of the tomato genome. For these lines, it is very unlikely as well that such complex rearrangements would occur spontaneously. In spite of this, these lines were not covered by the mandate for GMO regulation of APHIS. The reason for this was the absence of DNA sequences from pests. It is a stated policy by APHIS that anything that could also be developed by breeding is outside of purview. More specifically, introducing only sequences from compatible relatives is not considered by APHIS, provided that they are not from pests or introduced by a plant pest. This, rather than a likelihood of this event occurring in nature or breeding, is the trigger for an assessment by APHIS. It should be noted that assessment by APHIS is only one of the steps in approval GMOs in the United States.

Fig. 9 shows a schematic summary of the SVs and their position with respect to the baseline. Again, we want to emphasise that the border between 'below baseline' and 'above baseline' is not sharp, but fuzzy.

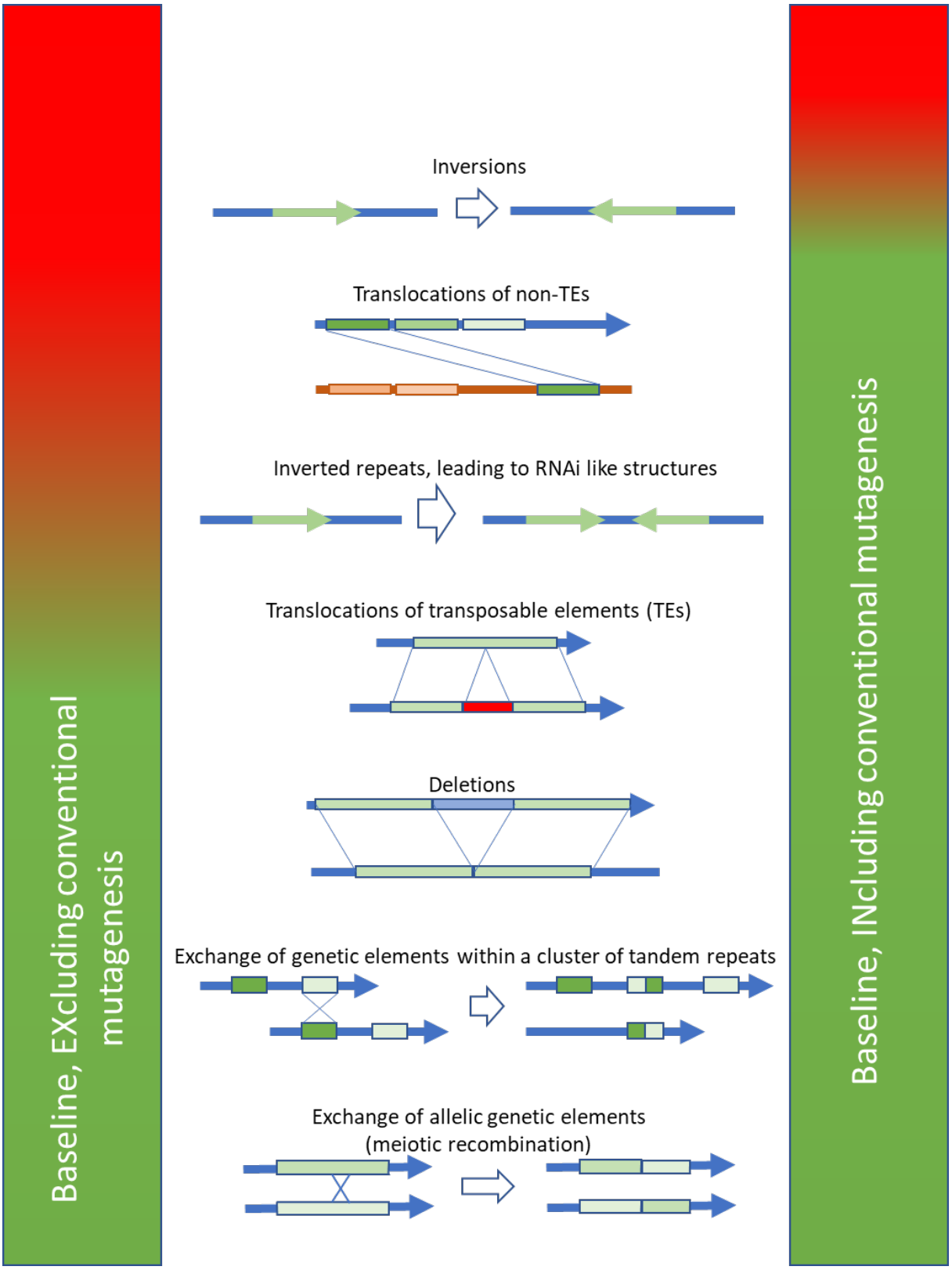

Figure 9. Different types of structural variation ordered from bottom to top in a decreasing likeliness to occur in nature or conventional breeding. *At the left-hand side, the baseline is shown in green, for cross-breeding and nature. The red part is considered as being above this baseline. At the right-hand side, the baseline*

is shown again, this time not only referring to nature and cross-breeding but also including conventional mutagenesis, in case this technology is regarded as a conventional plant breeding method.

The wording 'baseline' suggests a line. However, for SVs we rather regard it as a 'base zone' with a fuzzy border. The main reason for this is the scarcity of information on the frequency of spontaneously occurring structural changes. A second issue here is the discussion on whether mutagenesis, which may lead to a lot of structural changes in genomes, is regarded as belonging to conventional breeding methods, or as an (exempted) GM-method. Therefore, in Fig. 9, we show two columns: The right column regards mutagenesis as a method of conventional plant breeding, whereas the left column does not.

10.3 Final remarks

We conclude that many types of structural changes can occur spontaneously during cultivation and breeding, especially during conventional mutagenesis. However, apart from induction by mutagenesis, most structural changes occur at a low frequency. At the moment, there is insufficient information about spontaneously arising SV for defining a sharp baseline. Interviewed professionals of breeding companies had the opinion that the far majority of spontaneously arising SVs are not being noticed. Only SV that leads to beneficial traits is selected, investigated, and identified as such.

Although this report provides an interesting overview, we have to mention a few limitations:

- This report does not provide a comprehensive overview of all described SV in eukaryotes or crops. Instead, we have intended to describe all types of SVs, and have illustrated these with examples.
- The causal mechanisms that led to SV usually are not certainly known, as we have detailed descriptions of the resulting SVs, but we have not witnessed the processes themselves that led to these SVs. Instead, models have been proposed that give reasonable mechanisms that likely have been causal for the observed SV.
- The far majority of SV has remained unnoticed. The commonly used sequencing methods that produce short reads are not very suitable for detection of SV, and frequently give rise to false-positive and false-negative results. In crops, only the SVs that have led to striking phenotypes with a commercial or scientific value have been investigated in detail. Some SVs, such as inversions, have been studied as well since they had another impact on plant breeding, such as leading to recombination suppression, and therefore a tight genetic linkage between desired genes (alleles) and undesired linkage drag. As by far, most SVs have not been discovered yet, we cannot give reliable estimations of the frequencies of the different types of spontaneous structural changes, that might occur 'on a daily basis'. However, thanks to emerging and advancing long-range sequencing methods, we expect that the insight in frequencies in SVs will increase rapidly in the near future.

In spite of these limitations, we provide an overview of structural changes, and their tentative positions relative to a baseline.

References

Acuna-Hidalgo, R., Veltman, J.A. and Hoischen, A. (2016) New insights into the generation and role of de novo mutations in health and disease. *Genome Biol.*, **17**, 1–19.

Ahloowalia, B.S., Maluszynski, M. and Nichterlein, K. (2004) Global impact of mutation-derived varieties. *Euphytica*, **135**, 187–204.

Alkan, C., Coe, B.P. and Eichler, E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.

Allen, E., Xie, Z., Gustafson, A.M., Sung, G.-H., Spatafora, J.W. and Carrington, J.C. (2004) Evolution of microRNA genes by inverted duplication of target gene sequences in *Arabidopsis thaliana*. *Nat. Genet.*, **36**, 1282–1290.

Anderson, J.E., Michno, J.M., Kono, T.J.Y., Stec, A.O., Campbell, B.W., Curtin, S.J. and Stupar, R.M. (2016) Genomic variation and DNA repair associated with soybean transgenesis: A comparison to cultivars and mutagenized plants. *BMC Biotechnol.*, **16**, 1–13.

Anderson, S.N., Stitzer, M.C., Brohammer, A.B., et al. (2019) Transposable elements contribute to dynamic genome content in maize. *Plant J.*, **100**, 1052–1065.

Ashfield, T., Egan, A.N., Pfeil, B.E., et al. (2012) Evolution of a complex disease resistance gene cluster in diploid phaseolus and tetraploid glycine. *Plant Physiol.*, **159**, 336–354.

Audano, P.A., Sulovari, A., Graves-Lindsay, T.A., et al. (2019) Characterizing the major Structural Variant alleles of the human genome. *Cell*, **176**, 663–675.

Bai, Z., Chen, J., Liao, Y., Wang, M., Liu, R., Ge, S., Wing, R.A. and Chen, M. (2016) The impact and origin of copy number variations in the *Oryza* species. *BMC Genomics*, **17**, 1–12.

Bayer, P.E., Golicz, A.A., Tirnaz, S., Chan, C.K.K., Edwards, D. and Batley, J. (2019) Variation in abundance of predicted resistance genes in the *Brassica oleracea* pangenome. *Plant Biotechnol. J.*, **17**, 789–800.

Bennetzen, J.L. (2000) Comparative sequence analysis of plant nuclear genomes: Microcolinearity and its many exceptions. *Plant Cell*, **12**, 1021–1029.

Benoit, M., Drost, H.-G., Catoni, M., Gouil, Q., Lopez-Gomollon, S., Baulcombe, D. and Paszkowski, J. (2019) Environmental and epigenetic regulation of *Rider* retrotransposons in tomato O. Mittelsten Scheid, ed. *PLOS Genet.*, **15**, e1008370.

Bourque, G., Burns, K.H., Gehring, M., et al. (2018) Ten things you should know about transposable elements. *Genome Biol.*, **19**, 199.

Brutnell, T.P. and Dellaporta, S.L. (1994) Somatic inactivation and reactivation of Ac associated with changes in cytosine methylation and transposase expression. *Genetics*, **138**, 213–225.

Busch, B.L., Schmitz, G., Rossmann, S., Piron, F., Ding, J., Bendahmane, A. and Theres, K. (2011) Shoot branching and leaf dissection in tomato are regulated by homologous gene modules. *Plant Cell*, **23**, 3595–3609.

Butelli, E., Licciardello, C., Zhang, Y., Liu, J., Mackay, S., Bailey, P., Reforgiato-Recupero, G. and Martin, C. (2012) Retrotransposons control fruit-specific, cold-dependent accumulation of anthocyanins in blood oranges. *Plant Cell*, **24**, 1242–1255.

Carbonell-Bejerano, P., Royo, C., Torres-Pérez, R., et al. (2017) Catastrophic unbalanced genome rearrangements cause somatic loss of berry color in grapevine. *Plant Physiol.*, **175**, 786–801.

Carpentier, M.-C., Manfroi, E., Wei, F.-J., et al. (2019) Retrotranspositional landscape of Asian rice revealed by 3000 genomes. *Nat. Commun.*, **10**, 24.

Catoni, M., Jonesman, T., Cerruti, E. and Paszkowski, J. (2019) Mobilization of Pack-CACTA transposons in Arabidopsis suggests the mechanism of gene shuffling. *Nucleic Acids Res.*, **47**, 1311–1320.

Chaisson, M.J.P., Sanders, A.D., Zhao, X., et al. (2019) Multi-platform discovery of haplotype-resolved structural variation in human genomes. *Nat. Commun.*, **10**, 1–16.

Chen, J.-M., Chuzhanova, N., Stenson, P.D., Férec, C. and Cooper, D.N. (2005) Intrachromosomal serial replication slippage in trans gives rise to diverse genomic rearrangements involving inversions. *Hum.*

Mutat., **26**, 362–273.

Chen, Y., Schneeberger, R.G. and Cullis, C.A. (2005) A site-specific insertion sequence in flax genotrophs induced by environment. *New Phytol.*, **167**, 171–180.

Cheng, H., Sun, G., He, S., Gong, W., Peng, Z., Wang, R., Lin, Z. and Du, X. (2019) Comparative effect of allopolyploidy on transposable element composition and gene expression between *Gossypium hirsutum* and its two diploid progenitors. *J. Integr. Plant Biol.*, **61**, 45–59.

Chin, D.B., Arroyo-Garcia, R., Ochoa, O.E., Kesseli, R. V., Lavelle, D.O. and Michelmore, R.W. (2001) Recombination and spontaneous mutation at the major cluster of resistance genes in lettuce (*Lactuca sativa*). *Genetics*, **157**, 831–849.

Chiu, L.-W.W., Zhou, X., Burke, S., Wu, X., Prior, R.L. and Li, L. (2010) The purple cauliflower arises from activation of a MYB transcription factor. *Plant Physiol.*, **154**, 1470–1480.

Cho, J., Benoit, M., Catoni, M., Drost, H.-G., Brestovitsky, A., Oosterbeek, M. and Paszkowski, J. (2019) Sensitive detection of pre-integration intermediates of long terminal repeat retrotransposons in crop plants. *Nat. plants*, **5**, 26–33.

Cohen, S. and Segal, D. (2009) Extrachromosomal circular DNA in eukaryotes: Possible involvement in the plasticity of tandem repeats. *Cytogenet. Genome Res.*, **124**, 327–338.

Comai, L. and Tan, E.H. (2019) Haploid induction and genome instability. *Trends Genet.*, **35**, 791–803.

Cook, D.E., Lee, T.G., Guo, X., et al. (2012) Copy number variation of multiple genes at *Rhg1* mediates nematode resistance in soybean. *Science*, **338**, 1206–1209.

Cordaux, R., Hedges, D.J., Herke, S.W. and Batzer, M.A. (2006) Estimating the retrotransposition rate of human *Alu* elements. *Gene*, **373**, 134–137.

Debladis, E., Llauro, C., Carpentier, M.C., Mirouze, M. and Panaud, O. (2017) Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore Sequencing technology. *BMC Genomics*, **18**, 1–8.

Debolt, S. (2010) Copy number variation shapes genome diversity in Arabidopsis over immediate family generational scales. *Genome Biol. Evol.*, **2**, 441–453.

DelaPaz, J.S., Stronghill, P.E., Douglas, S.J., Saravia, S., Hasenkampf, C.A. and Riggs, C.D. (2012) Chromosome fragile sites in Arabidopsis harbor Matrix Attachment Regions that may be associated with ancestral chromosome rearrangement events. *PLoS Genet.*, **8**, e1003136.

Dolatabadian, A., Bayer, P.E., Tirnaz, S., Hurgobin, B., Edwards, D. and Batley, J. (2019) Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant Biotechnol. J.*, in press.

Dolatabadian, A., Patel, D.A., Edwards, D. and Batley, J. (2017) Copy number variation and disease resistance in plants. *Theor. Appl. Genet.*, **130**, 2479–2490.

Dong, J., Feng, Y., Kumar, D., Zhang, W., Zhu, T., Luo, M. and Messing, J. (2016) Analysis of tandem gene copies in maize chromosomal regions reconstructed from long sequence reads. *Proc. Natl. Acad. Sci. U.S.A*, **113**, 7949–7956.

Dooner, H.K., Wang, Q., Huang, J.T., Li, Y., He, L., Xiong, W. and Du, C. (2019) Spontaneous mutations in maize pollen are frequent in some lines and arise mainly from retrotranspositions and deletions. *Proc. Natl. Acad. Sci. U. S. A.*, **166**, 10734–10743.

EFSA Panel on Genetically Modified Organisms (GMO) (2012) Scientific opinion addressing the safety assessment of plants developed through cisgenesis and intragenesis. *EFSA J.*, **10**, 2561.

Exposito-Alonso, M., Becker, C., Schuenemann, V.J., et al. (2018) The rate and potential relevance of new mutations in a colonizing plant lineage. *PLoS Genet.*, **14**, 1–21.

Ferguson-Smith, M.A. (2015) History and evolution of cytogenetics. *Mol. Cytogenet.*, **8**, 19.

Feuk, L., Carson, A.R. and Scherer, S.W. (2006) Structural variation in the human genome. *Nat. Rev. Genet.*, **7**, 85–97.

Feuk, L., Marshall, C.R., Wintle, R.F. and Scherer, S.W. (2006) Structural variants: changing the landscape of chromosomes and design of disease studies. *Hum. Mol. Genet.*, **15 Spec No**, 57–66.

Filler Hayut, S., Melamed Bessudo, C., Levy, Avraham A., Zhang, D. and Levy, A. A. (2017) Targeted recombination between homologous chromosomes for precise breeding in tomato. *Nat. Commun.*, **8**, 15605.

Firko, M. (2019) Letter from USDA to Nexgen Plants. Available at: https://www.aphis.usda.gov/biotechnology/downloads/reg_loi/19-095-02_air_response_signed.pdf [Accessed December 3, 2019].

Fossi, M., Amundson, K.R., Kuppu, S., Britt, A.B. and Comai, L. (2019) Regeneration of *Solanum tuberosum* plants from protoplasts induces widespread genome instability. *Plant Physiol.*, **180**, 78–86.

Fouché, S., Plissonneau, C. and Croll, D. (2018) The birth and death of effectors in rapidly evolving filamentous pathogen genomes. *Curr. Opin. Microbiol.*, **46**, 34–42.

Freire-Benéitez, V., Gourlay, S., Berman, J. and Buscaino, A. (2016) Sir2 regulates stability of repetitive domains differentially in the human fungal pathogen *Candida albicans*. *Nucleic Acids Res.*, **44**, 9166–9179.

Fu, H. and Dooner, H.K. (2002) Intraspecific violation of genetic colinearity and its implications in maize. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 9573–9578.

Fuentes, R.R., Chebotarov, D., Duitama, J., et al. (2019) Structural variants in 3000 rice genomes. *Genome Res.*, **29**, 870–880.

Fukai, E., Karim, M.M., Shea, D.J., Tonu, N.N., Falk, K.C., Funaki, T. and Okazaki, K. (2019) An LTR retrotransposon insertion was the cause of world's first low erucic acid *Brassica rapa* oilseed cultivar. *Mol. Breed.*, **39**, 15.

Fukao, T., Harris, T. and Bailey-Serres, J. (2009) Evolutionary analysis of the Sub1 gene cluster that confers submergence tolerance to domesticated rice. *Ann. Bot.*, **103**, 143–150.

Gabur, I., Chawla, H.S., Snowdon, R.J. and Parkin, I.A.P. (2019) Connecting genome structural variation with complex traits in crop plants. *Theor. Appl. Genet.*, **132**, 733–750.

Gaines, T.A., Patterson, E.L. and Neve, P. (2019) Molecular mechanisms of adaptive evolution revealed by global selection for glyphosate resistance. *New Phytol.*, **223**, 1770–1775.

Gao, L., Gonda, I., Sun, H., et al. (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat. Genet.*, **51**, 1044–1051.

Gazzani, S., Gendall, A.R., Lister, C. and Dean, C. (2003) Analysis of the Molecular Basis of Flowering Time Variation in Arabidopsis Accessions. *Plant Physiol.*, **132**, 1107–1114.

Giner-Delgado, C., Villatoro, S., Lerga-Jaso, J., et al. (2019) Functional and evolutionary impact of polymorphic inversions in the human genome. *Nat. Commun.*, **10**, 4222.

Gorbunova, V. and Levy, A. (1997) Non-homologous DNA end joining in plant cells is associated with deletions and filler DNA insertions. *Nucleic Acids Res.*, **25**, 4650–4657.

Gorter de Vries, A.R., Couwenberg, L.G.F., Broek, M. van den, la Torre Cortés, P. de, Horst, J. Ter, Pronk, J.T. and Daran, J.-M.G. (2018) Allele-specific genome editing using CRISPR-Cas9 is associated with loss of heterozygosity in diploid yeast. *Nucleic Acids Res.*, **47**, 1362–1372.

Grund, E., Tremousaygue, D. and Deslandes, L. (2019) Plant NLRS with integrated domains: Unity makes strength. *Plant Physiol.*, **179**, 1227–1235.

Guo, H., Wang, X., Gundlach, H., Mayer, K.F.X., Peterson, D.G., Scheffler, B.E., Chee, P.W. and Paterson, A.H. (2014) Extensive and biased intergenomic nonreciprocal DNA exchanges shaped a nascent polyploid genome, *Gossypium* (Cotton). *Genetics*, **197**, 1153–1163.

Han, J.J., Jackson, D. and Martienssen, R. (2012) Pod corn is caused by rearrangement at the *Tunicate1* locus. *Plant Cell*, **24**, 2733–2744.

Hardigan, M.A., Crisovan, E., Hamilton, J.P., et al. (2016) Genome reduction uncovers a large dispensable genome and adaptive role for copy number variation in asexually propagated *Solanum tuberosum*. *Plant Cell*, **28**, 388–405.

Harrison, A.L. (1955) New "mouse-eared" mutant from var. Rutgers. Tomato Genet. Coop. Rep. 5, 18. *Tomato Genet. Coop. Rep.*, **5**, 18.

Haun, W.J., Hyten, D.L., Xu, W.W., et al. (2011) The composition and origins of genomic variation among individuals of the soybean reference cultivar williams 82. *Plant Physiol.*, **155**, 645–655.

Hayashi, K. and Yoshida, H. (2009) Refunctionalization of the ancient rice blast disease resistance gene Pit by the recruitment of a retrotransposon as a promoter. *Plant J.*, **57**, 413–425.

Hirano, T., Kazama, Y., Ishii, K., Ohbu, S., Shirakawa, Y. and Abe, T. (2015) Comprehensive identification of mutations induced by heavy-ion beam irradiation in *Arabidopsis thaliana*. *Plant J.*, **82**, 93–104.

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

Hormozdiari, F., Hajirasouliha, I., McPherson, A., Eichler, E.E. and Cenk Sahinalp, S. (2011) Simultaneous structural variation discovery among multiple paired-end sequenced genomes. *Genome Res.*, **21**, 2203–2212.

ISAAA (2015) USDA Approves Simplot's GM Potato. Available at: http://www.isaaa.org/kc/cropbiotechupdate/article/default.asp?ID=13703 [Accessed December 3, 2019].

Ito, H., Kim, J.M., Matsunaga, W., et al. (2016) A stress-activated transposon in Arabidopsis induces transgenerational abscisic acid insensitivity. *Sci. Rep.*, **6**, 1–12.

Iwata, H., Gaston, A., Remay, A., et al. (2012) The TFL1 homologue KSN is a regulator of continuous flowering in rose and strawberry. *Plant J.*, **69**, 116–125.

Jauhar, P.P. (2006) Modern biotechnology as an integral supplement to conventional plant breeding: The prospects and challenges. *Crop Sci.*, **46**, 1841–1859.

Jiang, N., Bao, Z., Zhang, X., Eddy, S.R. and Wessler, S.R. (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature*, **431**, 569–573.

Jiang, N., Ferguson, A.A., Slotkin, R.K. and Lisch, D. (2011) Pack-Mutator-like transposable elements (Pack-MULEs) induce directional modification of genes through biased insertion and DNA acquisition. *Proc. Natl. Acad. Sci. U.S.A*, **108**, 1537–1542.

Jo, Y.D. and Kim, J.-B. (2019) Frequency and spectrum of radiation-induced mutations revealed by whole-genome sequencing analyses of plants. *Quantum Beam Sci.*, **3**, 7.

JR Simplot Company (2015) JR Simplot Company petition (14-093-01p) for determination of nonregulated status for InnateTM potatoes with Late Blight resistance, low acrylamide potential, reduced Black Spot and lowered reducing Sugars: Russet Burbank event W8. Available at: https://gatesopenresearch.org/documents/3-971.

Kaeppler, S.M., Kaeppler, H.F. and Rhee, Y. (2000) Epigenetic aspects of somaclonal variation in plants. *Plant Mol. Biol.*, **43**, 179–188.

Kapitonov, V. V. and Jurka, J. (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet.*, **23**, 521–529.

Kazama, Y., Ishii, K., Hirano, T., Wakana, T., Yamada, M., Ohbu, S. and Abe, T. (2017) Different mutational function of low- and high-linear energy transfer heavy-ion irradiation demonstrated by whole-genome resequencing of Arabidopsis mutants. *Plant J.*, **92**, 1020–1030.

Khan, A.W., Garg, V., Roorkiwal, M., Golicz, A.A., Edwards, D. and Varshney, R.K. (2019) Super-Pangenome by Integrating the Wild Side of a Species for Accelerated Crop Improvement. *Trends Plant Sci.*, in press.

Kim, S., Park, J., Yeom, S.I., et al. (2017) New reference genome sequences of hot pepper reveal the massive evolution of plant disease-resistance genes by retroduplication. *Genome Biol.*, **18**, 1–11.

Kobayashi, S., Goto-Yamamoto, N. and Hirochika, H. (2004) Retrotransposon-induced mutations in grape skin color. *Science*, **304**, 982.

Koo, D.H., Jugulam, M., Putta, K., Cuvaca, I.B., Peterson, D.E., Currie, R.S., Friebe, B. and Gill, B.S. (2018) Gene duplication and aneuploidy trigger rapid evolution of herbicide resistance in common waterhemp. *Plant Physiol.*, **176**, 1932–1938.

Koo, D.H., Molin, W.T., Saski, C.A., Jiang, J., Putta, K., Jugulam, M., Friebe, B. and Gill, B.S. (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.

Krasileva, K. V. (2019) The role of transposable elements and DNA damage repair mechanisms in gene duplications and gene fusions in plant genomes. *Curr. Opin. Plant Biol.*, **48**, 18–25.

Krishna, H., Alizadeh, M., Singh, D., Singh, U., Chauhan, N., Eftekhari, M. and Sadh, R.K. (2016) Somaclonal variations and their applications in horticultural crops improvement. *3 Biotech*, **6**, 1–18.

Kuang, H., Woo, S.S., Meyers, B.C., Nevo, E. and Michelmore, R.W. (2004) Multiple genetic processes result in heterogeneous rates of evolution within the major cluster disease resistance genes in lettuce. *Plant Cell*, **16**, 2870–2894.

Lai, J., Li, Y., Messing, J. and Dooner, H.K. (2005) Gene movement by Helitron transposons contributes to the haplotype variability of maize. *Proc. Natl. Acad. Sci.*, **102**, 9068–9073.

Lanciano, S., Carpentier, M.C., Llauro, C., Jobet, E., Robakowska-Hyzorek, D., Lasserre, E., Ghesquière, A., Panaud, O. and Mirouze, M. (2017) Sequencing the extrachromosomal circular mobilome reveals retrotransposon activity in plants. *PLoS Genet.*, **13**, 1–20.

Leitão, J.M. (2012) Chemical mutagenesis. In Q.-Y. Shu, B. Forster, and H. Nakagawa, eds. *Plant mutation breeding and biotechnology*. Vienna: FAO, pp. 135–158.

Li, S., Xu, H., Ju, Z., et al. (2018) The *RIN-MC* fusion of MADS-Box transcription factors has transcriptional activity and modulates expression of many ripening genes. *Plant Physiol.*, **176**, 891–909.

Li, X., Song, Y., Century, K., Straight, S., Ronald, P., Dong, X., Lassner, M. and Zhang, Y. (2001) A fast neutron deletion mutagenesis-based reverse genetics system for plants. *Plant J.*, **27**, 235–242.

Li, X. and Zhang, Y. (2002) Reverse genetics by fast neutron mutagenesis in higher plants. *Funct. Integr. Genomics*, **2**, 254–258.

Lindstrom, W. (1925) Inheritance in tomatoes. *Genetics*, **10**, 305–317.

Ling, H.Q., Bauer, P., Bereczky, Z., Keller, B. and Ganal, M. (2002) The tomato *fer* gene encoding a bHLH protein controls iron-uptake responses in roots. *Proc. Natl. Acad. Sci. U. S. A.*, **99**, 13938–13943.

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

Lisch, D.R. (2013a) How important are transposons for plant evolution? *Nat. Rev. Genet.*, **14**, 49–61.

Lisch, D.R. (2013b) Transposons in plant gene regulation. In N. Fedoroff, ed. *Plant Transposons and Genome Dynamics in Evolution*. Oxford: Wiley-Blackwell, pp. 93–116.

Liu, Z., Liu, Y., Liu, F., Zhang, S., Wang, X., Lu, Q., Wang, K., Zhang, B. and Peng, R. (2018) Genome-wide survey and comparative analysis of long terminal repeat (LTR) retrotransposon families in four gossypium species. *Sci. Rep.*, **8**, 2–11.

Löytynoja, A. and Goldman, N. (2017) Short template switch events explain mutation clusters in the human genome. *Genome Res.*, **27**, 1039–1049.

Lu, H., Cui, X., Liu, Z., et al. (2018) Discovery and annotation of a novel transposable element family in *Gossypium*. *BMC Plant Biol.*, **18**, 1–10.

Lu, P., Han, X., Qi, J., Yang, J., Wijeratne, A.J., Li, T. and Ma, H. (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.

Lu, S., Eck, J. Van, Zhou, X., et al. (2006) The cauliflower Or gene encodes a DnaJ cysteine-rich domain-containing protein that mediates high levels of β-carotene accumulation. *Plant Cell*, **18**, 3594–3605.

Lye, Z.N. and Purugganan, M.D. (2019) Copy number variation in domestication. *Trends Plant Sci.*, **24**, 352–365.

MacDonald, J.R., Ziman, R., Yuen, R.K.C., Feuk, L. and Scherer, S.W. (2014) The database of genomic variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.*, **42**, 986–992.

McCarty, D.R., Mark Settles, A., Suzuki, M., et al. (2005) Steady-state transposon mutagenesis in inbred maize. *Plant J.*, **44**, 52–61.

McClintock, B. (1942) The fusion of broken ends of chromosomes following nuclear fusion. *Proc. Natl. Acad. Sci. U.S.A.*, **28**, 458–463.

McClintock, B. (1950) The origin and behavior of mutable loci in maize. *Proc. Natl. Acad. Sci. U.S.A.*, **36**, 344–355.

McClintock, B. (1984) The significance of responses of the genome to challenge. *Science*, **226**, 792–801.

McClintock, B. (1941) The stability of broken ends of chromosomes in Zea mays. *Genetics*, **26**, 234–282.

McHale, L.K., Haun, W.J., Xu, W.W., Bhaskar, P.B., Anderson, J.E., Hyten, D.L., Gerhardt, D.J., Jeddeloh, J.A. and Stupar, R.M. (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. *Plant Physiol.*, **159**, 1295–1308.

Meers, C., Keskin, H. and Storici, F. (2016) DNA repair by RNA: Templated, or not templated, that is the question. *DNA Repair (Amst).*, **44**, 17–21.

Mills, R.E., Bennett, E.A., Iskow, R.C. and Devine, S.E. (2007) Which transposable elements are active in the human genome? *Trends Genet.*, **23**, 183–191.

Molinier, J., Ries, G., Bonhoeffer, S. and Hohn, B. (2004) Interchromatid and interhomolog recombination in Arabidopsis thaliana. *Plant Cell*, **16**, 342–52.

Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A. and Rafalski, A. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat. Genet.*, **37**, 997–1002.

Murat, F., Armero, A., Pont, C., Klopp, C. and Salse, J. (2017) Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.*, **49**, 490–496.

Nannas, N.J. and Dawe, R.K. (2015) Genetic and genomic toolbox of *Zea mays*. *Genetics*, **199**, 655–669.

Navrátilová, A., Koblížková, A. and Macas, J. (2008) Survey of extrachromosomal circular DNA derived from plant satellite repeats. *BMC Plant Biol.*, **8**, 1–13.

Noormohammadi, Z., Ibrahim-Khalili, N., Ghasemzadeh-Baraki, S., Sheidai, M., Alishah, O., Ibrahim Khalili, N. and Ghasemzadeh Baraki, S. (2016) Genetic screening of diploid and tetraploid cotton cultivars based on retrotransposon microsatellite amplified polymorphism markers (REMAP). *An. Biol.*, 123–132.

Noormohammadi, Z., Ibrahim-Khalili, N., Sheidai, M. and Alishah, O. (2018) Genetic fingerprinting of diploid and tetraploid cotton cultivars by retrotransposon-based markers. *Nucl.*, **61**, 137–143.

O'Huallachain, M., Karczewski, K.J., Weissman, S.M., Urban, A.E. and Snyder, M.P. (2012) Extensive genetic variation in somatic human tissues. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 18018–18023.

Oliver, K.R., McComb, J.A. and Greene, W.K. (2013) Transposable elements: Powerful contributors to angiosperm evolution and diversity. *Genome Biol. Evol.*, **5**, 1886–1901.

Ong-Abdullah, M., Ordway, J.M., Jiang, N., et al. (2015) Loss of *Karma* transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, **525**, 533–537.

Ossowski, S., Schneeberger, K., Lucas-Lledó, J.I., Warthmann, N., Clark, R.M., Shaw, R.G., Weigel, D. and Lynch, M. (2010) The rate and molecular spectrum of spontaneous mutations in *Arabidopsis thaliana*. *Science*, **327**, 92–94.

Otto, D., Petersen, R., Brauksiepe, B., Braun, P. and Schmidt, E.R. (2014) The columnar mutation ("Co gene") of apple (*Malus* × *domestica*) is associated with an integration of a Gypsy-like retrotransposon. *Mol. Breed.*, **33**, 863–880.

Panchy, N., Lehti-Shiu, M. and Shiu, S.H. (2016) Evolution of gene duplication in plants. *Plant Physiol.*, **171**, 2294–2316.

Parnis, A., Cohen, O., Gutfinger, T., Hareven, D., Zamir, D. and Lifschitz, E. (1997) The dominant developmental mutants of tomato, *Mouse-ear* and *Curl*, are associated with distinct modes of abnormal transcriptional regulation of a *Knotted* gene. *Plant Cell*, **9**, 2143–2158.

Paszkowski, J. (2015) Controlled activation of retrotransposition for plant breeding. *Curr. Opin. Biotechnol.*, **32C**, 200–206.

Payer, L.M. and Burns, K.H. (2019) Transposable elements in human genetic disease. *Nat. Rev. Genet.*, **20**, 760–772.

Peng, Z., Zhou, W., Fu, W., Du, R., Jin, L. and Zhang, F. (2015) Correlation between frequency of non-allelic homologous recombination and homology properties: evidence from homology-mediated CNV mutations in the human genome. *Hum. Mol. Genet.*, **24**, 1225–1233.

Price, H.L. and Drinkard, A.W. (1908) Inheritance in tomato hybrids. *Mol. Gen. Genet.*, **1**, 402–403.

Quadrana, L., Silveira, A.B., Mayhew, G.F., LeBlanc, C., Martienssen, R.A., Jeddeloh, J.A. and Colot, V. (2016) The *Arabidopsis thaliana* mobilome and its impact at the species level. *Elife*, **5**, 1–25.

Renny-Byfield, S. and Wendel, J.F. (2014) Doubling down on genomes: Polyploidy and crop plants. *Am. J. Bot.*, **101**, 1711–1725.

Rick, C.M. (1956) Genetic and systematic studies on accessions of *Lycospersicon* from the Galapagos islands. *Am. J. Bot.*, **43**, 687.

Robinson, R. and Tomes, M. (1968) *Ripening inhibitor*: a gene with multiple effect on ripening. *Tomato Genet. Coop.*, **18**, 36-37.

Roffler, S. and Wicker, T. (2015) Genome-wide comparison of Asian and African rice reveals high recent activity of DNA transposons. *Mob. DNA*, **6**, 1–14.

Roldan, M.V.G., Périlleux, C., Morin, H., Huerga-Fernandez, S., Latrasse, D., Benhamed, M. and Bendahmane, A. (2017) Natural and induced loss of function mutations in *SlMBP21* MADS-box gene led to *jointless-2* phenotype in tomato. *Sci. Rep.*, **7**, 4402.

Salse, J. (2016) Ancestors of modern plant crops. *Curr. Opin. Plant Biol.*, **30**, 134–142.

Sanseverino, W., Hénaff, E., Vives, C., Pinosio, S., Burgos-Paz, W., Morgante, M., Ramos-Onsins, S.E., Garcia-Mas, J. and Casacuberta, J.M. (2015) Transposon insertion, structural variations and SNPs contribute to the evolution of the melon genome. *Mol. Biol. Evol.*, **32**, 2760–2774.

Saxena, R.K., Edwards, D. and Varshney, R.K. (2014) Structural variations in plant genomes. *Briefings Funct. Genomics Proteomics*, **13**.

Schenk, P. and Hervé, P. (2019) Letter from NexGen Plants to Aphis. Available at: https://www.aphis.usda.gov/biotechnology/downloads/reg_loi/19-095-02_air_inquiry_cbidel_a1.pdf [Accessed December 3, 2019].

Schenk, P. and Nowak, E. (2017) Construct and vector for intragenic plant transformation. , **1**.

Schmidt, C., Pacher, M. and Puchta, H. (2019) Efficient induction of heritable inversions in plant genomes using the CRISPR/Cas system. *Plant J.*, **98**, 577–589.

Shu, Q.Y., Forster, B.P. and Nakagawa, H. (2012) *Plant mutation breeding and biotechnology* Q. Y. Shu, B. P. Forster, and H. Nakagawa, eds., Vienna: International Atomic Energy Agency.

Soltis, P.S., Marchant, D.B., Peer, Y. Van de and Soltis, D.E. (2015) Polyploidy and genome evolution in plants. *Curr. Opin. Genet. Dev.*, **35**, 119–125.

Soltis, P.S. and Soltis, D.E. (2016) Ancient WGD events as drivers of key innovations in angiosperms. *Curr. Opin. Plant Biol.*, **30**, 159–165.

Springer, N.M., Ying, K., Fu, Y., et al. (2009) Maize inbreds exhibit high levels of copy number variation (CNV) and presence/absence variation (PAV) in genome content. *PLoS Genet.*, **5**, e1000734.

Sticklen, M. (2015) Transgenic, cisgenic, intragenic and subgenic Crops. *Adv. Crop Sci. Technol.*, **03**, 2–3.

Studer, A., Zhao, Q., Ross-Ibarra, J. and Doebley, J. (2011) Identification of a functional transposon insertion in the maize domestication gene *tb1*. *Nat. Genet.*, **43**, 1160–1163.

Sun, S., Zhou, Y., Chen, J., et al. (2018) Extensive intraspecific gene order and gene structural variations between Mo17 and other maize genomes. *Nat. Genet.*, **50**, 1289–1295.

Tian, Z., Zhao, M., She, M., et al. (2012) Genome-wide characterization of nonreference transposons reveals evolutionary propensities of transposons in soybean. *Plant Cell*, **24**, 4422–4436.

Tsuchiya, T. and Eulgem, T. (2013) An alternative polyadenylation mechanism coopted to the Arabidopsis RPP7 gene through intronic retrotransposon domestication. *Proc. Natl. Acad. Sci.*, **110**, E3535–E3543.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, **461**, 423–426.

Tuteja, J.H. and Vodkin, L.O. (2008) Structural features of the endogenous silencing and target loci in the soybean genome. *Crop Sci.*, **48**, S-49.

Valliyodan, B., Cannon, S.B., Bayer, P.E., et al. (2019) Construction and comparison of three reference-quality genome assemblies for soybean. *Plant J.*, 1066–1082.

Vaughn, J.N. and Bennetzen, J.L. (2014) Natural insertions in rice commonly form tandem duplications indicative of patch-mediated double-strand break induction and repair. *Proc. Natl. Acad. Sci. U.S.A.*, **111**, 6684–6689.

Vitte, C., Fustier, M.A., Alix, K. and Tenaillon, M.I. (2014) The bright side of transposons in crop evolution. *Briefings Funct. Genomics Proteomics*, **13**, 276–295.

Vrebalov, J., Ruezinsky, D., Padmanabhan, V., White, R., Medrano, D., Drake, R., Schuch, W. and Giovannoni, J. (2002) A MADS-box gene necessary for fruit ripening at the tomato *ripening-inhibitor* (*rin*) locus. *Science*, **296**, 343–346.

Walbot, V. (1991) The *Mutator* transposable element family of maize. *Genet. Eng. (N. Y).*, **13**, 1–37.

Wang, K., Huang, G. and Zhu, Y. (2016) Transposable elements play an important role during cotton genome evolution and fiber cell development. *Sci. China Life Sci.*, **59**, 112–121.

Wang, M., Tu, L., Yuan, D., et al. (2019) Reference genome sequences of two cultivated allotetraploid cottons, *Gossypium hirsutum* and *Gossypium barbadense*. *Nat. Genet.*, **51**, 224–229.

Wang, W., Mauleon, R., Hu, Z., et al. (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. *Nature*, **557**, 43–49.

Wang, X., Weigel, D. and Smith, L.M. (2013) Transposon variants and their effects on gene expression in Arabidopsis. *PLoS Genet.*, **9**, e1003255.

Wellenreuther, M., Mérot, C., Berdan, E. and Bernatchez, L. (2019) Going beyond SNPs: The role of structural genomic variants in adaptive evolution and species diversification. *Mol. Ecol.*, **28**, 1203–1209.

Weng, M.L., Becker, C., Hildebrandt, J., Neumann, M., Rutter, M.T., Shaw, R.G., Weigel, D. and Fenster, C.B. (2019) Fine-grained analysis of spontaneous mutation spectrum and frequency in *Arabidopsis thaliana*. *Genetics*, **211**, 703–714.

Wessler, S., Tarpley, A., Purugganan, M., Spell, M. and Okagaki, R. (1990) Filler DNA is associated with spontaneous deletions in maize. *Proc. Natl. Acad. Sci. U. S. A.*, **87**, 8731–8735.

Wessler, S.R. (1996) Plant retrotransposons: Turned on by stress. *Curr. Biol.*, **6**, 959–961.

Weyer, A.-L. Van de, Monteiro, F., Furzer, O.J., et al. (2019) A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell*, **178**, 1260-1272.e14.

Wicker, T., Buchmann, J.P. and Keller, B. (2010) Patching gaps in plant genomes results in gene movement and erosion of colinearity. *Genome Res.*, **20**, 1229–1237.

Wicker, T., Yu, Y., Haberer, G., et al. (2016) DNA transposon activity is associated with increased mutation rates in genes of rice and other grasses. *Nat. Commun.*, **7**, 1–9.

Wolters, A.M.A., Caro, M., Dong, S., Finkers, R., Gao, J., Visser, R.G.F., Wang, X., Du, Y. and Bai, Y. (2015) Detection of an inversion in the *Ty-2* region between *S. lycopersicum* and *S. habrochaites* by a combination of de novo genome assembly and BAC cloning. *Theor. Appl. Genet.*, **128**, 1987–1997.

Wolters, P.J., Schouten, H.J., Velasco, R., Si-Ammour, A. and Baldi, P. (2013) Evidence for regulation of columnar habit in apple by a putative 2OG-Fe(II) oxygenase. *New Phytol.*, **200**, 993–999.

Wu, F., Sedivy, E.J., Price, W.B., Haider, W. and Hanzawa, Y. (2017) Evolutionary trajectories of duplicated FT homologues and their roles in soybean domestication. *Plant J.*, 941–953.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E.J. and Knaap, E. van der (2008) A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science*, **319**, 1527–1530.

Xing, J., Witherspoon, D.J. and Jorde, L.B. (2013) Mobile element biology: new possibilities with high-throughput sequencing. *Trends Genet.*, **29**, 280–289.

Xu, C., Liberatore, K.L., MacAlister, C. a, et al. (2015) A cascade of arabinosyltransferases controls shoot meristem size in tomato. *Nat. Genet.*, **47**, 784–792.

Xu, M., Brar, H.K., Grosic, S., Palmer, R.G. and Bhattacharyya, M.K. (2010) Excision of an active CACTA-like transposable element from *DFR2* causes variegated flowers in soybean [*Glycine max* (L.) Merr.]. *Genetics*, **184**, 53–63.

Yang, Zhaoen, Ge, X., Yang, Zuoren, et al. (2019) Extensive intraspecific gene order and gene structural variations in upland cotton cultivars. *Nat. Commun.*, **10**, 2989.

Yao, J.L., Dong, Y.H. and Morris, B.A.M. (2001) Parthenocarpic apple fruit production conferred by transposon insertion mutations in a MADS-box transcription factor. *Proc. Natl. Acad. Sci. U. S. A.*, **98**, 1306–1311.

Yi, K. and Ju, Y.S. (2018) Patterns and mechanisms of structural variations in human cancer. *Exp. Mol. Med.*, **50**, 98.

Yu, P., Wang, C.H., Xu, Q., Feng, Y., Yuan, X.P., Yu, H.Y., Wang, Y.P., Tang, S.X. and Wei, X.H. (2013) Genome-wide copy number variations in Oryza sativa L. *BMC Genomics*, **14**.

Zabala, G. and Vodkin, L. (2007) Novel exon combinations generated by alternative splicing of gene fragments mobilized by a CACTA transposon in *Glycine max*. *BMC Plant Biol.*, **7**, 1–9.

Zabala, G. and Vodkin, L.O. (2005) The *wp* mutation of *Glycine max* carries a gene-fragment-rich transposon of the CACTA superfamily. *Plant Cell*, **17**, 2619–2632.

Zapata, L., Ding, J., Willing, E., et al. (2016) Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proc. Natl. Acad. Sci. U. S. A.*, **113**, E4052-60.

Zhang, C., Liu, C., Weng, J., Cheng, B., Liu, F., Li, X. and Xie, C. (2016) Creation of targeted inversion mutations in plants using a RNA-guided endonuclease. *Crop J.*, **5**, 83–88.

Zhang, J., Zuo, T. and Peterson, T. (2013) Generation of tandem direct duplications by reversed-Ends transposition of maize Ac elements C. Feschotte, ed. *PLoS Genet.*, **9**, e1003691.

Zhang, P., Allen, W.B., Nagasawa, N., et al. (2012) A transposable element insertion within *ZmGE2* gene is associated with increase in embryo to endosperm ratio in maize. *Theor. Appl. Genet.*, **125**, 1463–1471.

Zhang, X., Chen, X., Liang, P. and Tang, H. (2018) Cataloguing plant genome structural variations. *Curr. Issues Mol. Biol.*, **27**, 181–193.

Zhang, X., Meng, L., Liu, B., Hu, Y., Cheng, F., Liang, J., Aarts, M.G.M., Wang, X. and Wu, J. (2015) A transposon insertion in *FLOWERING LOCUS T* is associated with delayed flowering in *Brassica rapa*. *Plant Sci.*, **241**, 211–220.

Zhang, Z., Mao, L., Chen, H., et al. (2015) Genome-wide mapping of structural variations reveals a Copy Number Variant that determines reproductive morphology in cucumber. *Plant Cell*, **27**, 1595–1604.

Zhou, P., Silverstein, K.A.T., Ramaraj, T., et al. (2017) Exploring structural variation and gene family architecture with *De Novo* assemblies of 15 *Medicago* genomes. *BMC Genomics*, **18**, 1–14.

Zhou, Y., Minio, A., Massonnet, M., Solares, E., Lv, Y., Beridze, T., Cantu, D. and Gaut, B.S. (2019) The population genetics of structural variants in grapevine domestication. *Nat. plants*, **5**, 965–979.

Zmieńko, A., Samelak, A., Kozłowski, P. and Figlerowicz, M. (2014) Copy number polymorphism in plant genomes. *Theor. Appl. Genet.*, **127**, 1–18.

List of Abbreviations

CNV Copy Number Variation

cv. cultivar

CGH Comparative Genome Hybridization

DSB Double-Strand Break (in DNA)

FoSTeS Fork Stalling Template Switching

GC Gene Conversion

HTP High ThroughPut

IR Inverted Repeat

LCR Low Copy Repeat

LRR Leucine-Rich Repeat

LTR Long Terminal Repeat

MA Mutation Accumulation (lines)

MITE Miniature Inverted-repeat TE

MULE Mutator–like transposable element

NAHR Non-Allelic Homologous Recombination

NB-LRR nucleotide-binding domain (NB) and a leucine-rich repeat (LRR) domain

NHEJ Non-Homologous End Joining

NLR NB-LRR-related gene

PAV Presence/Absence Variation

RNAi (RNA interference)

SINE Short interspersed nuclear elements

RT Reverse Transcription

SDN Site-Directed Nuclease(s)

SDSA Synthesis-Dependent Strand Annealing

SV Structural variation

TE Transposable Element (Transposon)

TIPS Transposon Insertion Polymorphisms

WGD Whole Genome Duplication

WGS Whole Genome (Re-)Sequencing