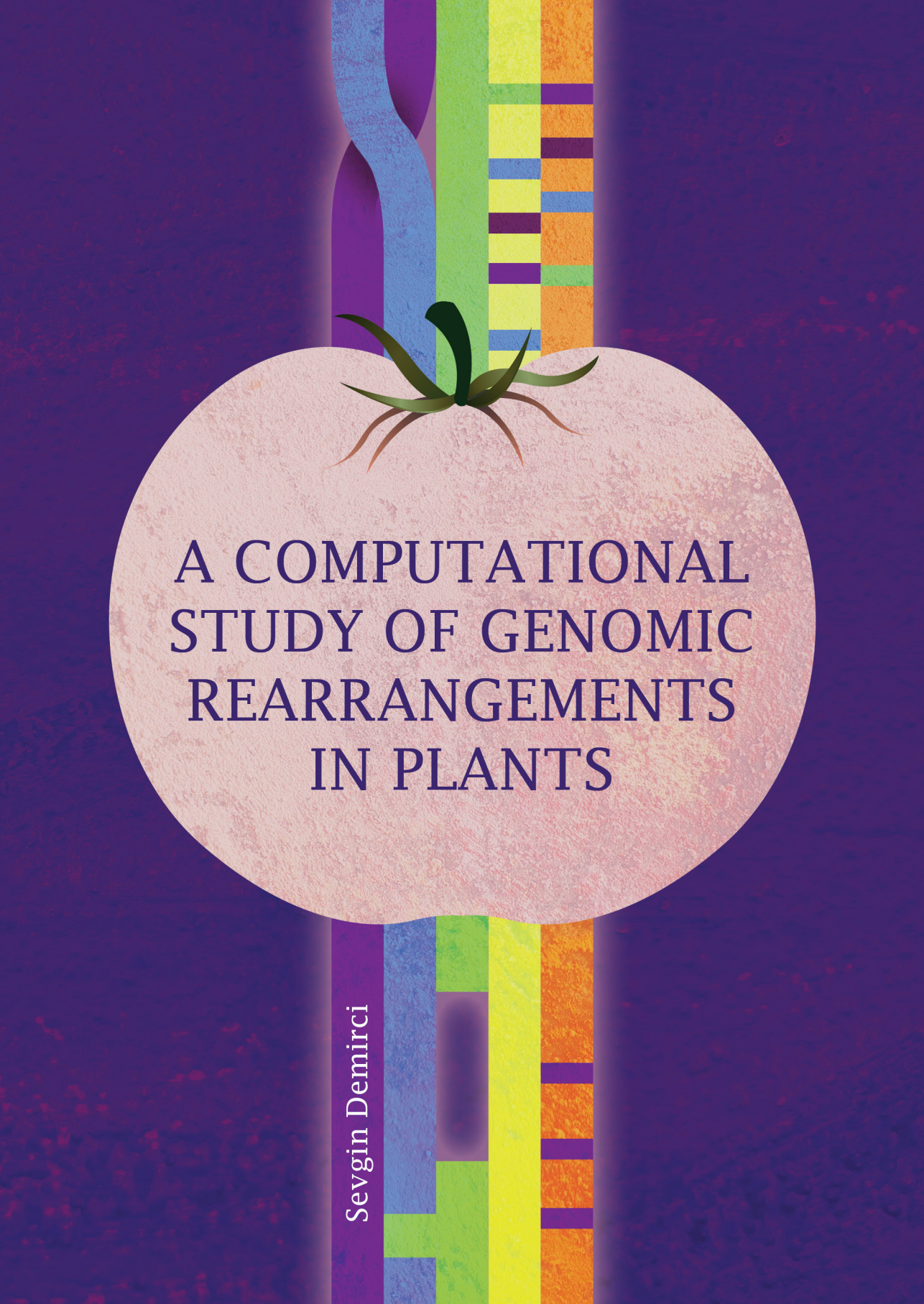
A COMPUTATIONAL STUDY OF GENOMIC REARRANGEMENTS IN PLANTS
Sevgin Demirci

Propositions

1. Due to breeding, *Solanum lycopersicum* and *S. pimpinellifolium* are no longer distinct species.
   (this thesis)

2. Machine learning algorithms can uniquely uncover related features underlying recombination.
   (this thesis)

3. Rewriting existing software code for repurposing is less efficient than starting from scratch.

4. In life science curricula, data-driven analysis should receive as much emphasis as classical hypothesis-based approaches.

5. The lesson of the Covid-19 crisis is that employees should be able to choose to work where they are most productive.

6. Tea bags are a disgrace to the spirit of ancient tea ceremonies.

Propositions belonging to the thesis, entitled

A computational study of genomic rearrangements in plants

Sevgin Demirci
Wageningen, 2 November 2021

A computational study of genomic rearrangements in plants

Sevgin Demirci

Thesis committee

Promotor
Prof. Dr D. de Ridder
Professor of Bioinformatics
Wageningen University & Research

Co-promoters
Dr S.A. Peters
Senior scientist, Bioscience
Wageningen University & Research

Dr A.D.J. van Dijk
Assistant Professor, Bioinformatics Group
Wageningen University & Research

Other members
Prof. Dr Y. Bai, Wageningen University & Research
Dr P.F. Fransz, University of Amsterdam
Prof. Dr K. Schneeberger, Ludwig-Maximilians-Universität München
Dr R.H.G. Dirks, Managerial Genetics BVBA, Maaseik

This research was conducted under the auspices of the Graduate School Experimental Plant Sciences

A computational study of genomic rearrangements in plants

Sevgin Demirci

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus,
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Tuesday 2 November 2021
at 11 a.m. in the Aula.

Sevgin Demirci
A computational study of genomic rearrangements in plants, 131 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2021)
With references, with summary in English

ISBN 978-94-6395-929-2
DOI https://doi.org/10.18174/551238

Table of Contents

Chapter 1

Introduction

In nature, genetic diversity helps species survive in ever changing environmental conditions. Species adapt to these conditions through natural selection which acts on the phenotype controlled by many factors including the gene pool of the population. Plants also adapt in the same way. Over thousands of years, humans intervened in this process and domesticated plants via human directed adaptation, which created crops suitable for growth in predictable and controlled environments. During domestication, the genetic diversity in crops is reduced compared to their wild relatives as we breed and select for a few desirable lines (Tanksley and McCouch, 1997). With reduced genetic diversity, crops lose some natural defence mechanisms compared to wild populations and become less resilient to environmental changes. This is especially important now as we are facing an imminent threat of global warming on Earth: for example, 22% of major cities are expected to face a major change in climate by 2050 (Bastin *et al.*, 2019). Similar changes will affect the rural areas where the majority of crops are grown. In addition to future environment changes, crop production is facing a threat from pests. Pests evolve and adapt, putting plant breeders in a simultaneous race with environment changes and pest adaptation. Given the reduced genetic diversity due to domestication, adaptation to environmental changes in crops is much harder than in wild plants. In trying to control the environment of crops, we make them vulnerable to potential contaminants, environmental changes and evolved or migrated pests.

To make crops resilient to the aforementioned threats, plant breeding approaches such as introgression breeding aim to bring valuable genes from wild species to the crops by crossing. To aid introgression breeding, we require the knowledge of the current genetic composition of crops (cultivars) as well as wild related species. We can only achieve this with fundamental research on genetics. Comparative genomics studies will help to unravel the differences in genetic diversity between crops that have been bred for specific traits and their wild relatives who have successfully survived in nature for thousands/millions of years. Genetic studies also provide an understanding of the evolutionary history of plants, and the genetic consequences of domestication. This knowledge is thus valuable to bridge the genetic gap between wild varieties and crops through introgression breeding. It can also help to increase the genetic diversity of crops through other forms of breeding, for instance precision breeding, i.e. genetic modification using tools (Doebley *et al.,* 2006) such as CRISPR/Cas (Maagd *et al.*, 2020), gene editing, gene knockout, etc. However, introgression breeding uses natural ways to modify the genetic content of the crops and is thus not affected by the various GMO regulations around the world.

Genetic diversity in plants is generated through several natural processes, including meiotic recombination, transposon activity, whole-genome duplication, chromosomal breaks, horizontal gene transfer and random mutations. Among these, meiotic recombination is heavily used in plant breeding and yet its mechanism is largely unknown. Genetic diversity is most often measured in terms of local variation such as SNPs and small indels. However, genomes also change at larger scales, mostly due to recombination, structural variation and transpositions. Such variations and their effects were neglected until recently, mostly due to unavailability of a reliable technology to detect them (Sedlazeck *et al.*, 2018). Once detected, large-scale variation could be associated with phenotypes and certain traits which have direct effect on plant breeding strategies. In this thesis, we study the genetic diversity in crops and their wild relatives at larger genomic scales. We focus on meiotic recombination, structural variation and transposable elements.

1.1 Genomic variation

1.1.1 (Meiotic) Recombination

Recombination is the main source of genetic diversity in organisms. It occurs in mitosis in somatic cells and in meiosis in germ cells. In plants, research is mostly focused on heritable traits, and because recombination in germ cells is inherited by new generations we focus on meiotic recombination. Meiosis, a cell division process, results in haploid cells. In dividing cells, during the interphase stage just before meiosis, genome duplication takes place (see Figure 1.1). The (homologous) chromosomes double and create sister chromatids (identical copies of a chromosome). In the prophase I stage of meiosis, highly syntenic parts of homologous chromosomes align to each other (synapsis). Subsequently, double strand breaks (DSBs) are induced by a protein called SPO11 (Osman *et al.*, 2011). The open end of the strand invades the non-sister chromatid of the homologous chromosome and a D-loop is formed in the invaded chromatid. In this way, a connection between chromatids is established (chiasmata). After this invasion, several outcomes are possible depending on the repair pathways involved. There are at least three known pathways (Figure 1.1). In the synthesis dependent strand annealing (SDSA) pathway, the invasion is rejected and the chromatids are repaired via DNA synthesis, using the sister chromatid as a template. The result is called non-crossover (NCO) recombination, sometimes accompanied by gene conversion, that is small genetic material exchange. In the ZMM pathway, after a successful invasion, the second open end is captured to form a double Holliday junction (dHj). The dHj is then resolved in a crossover (CO), where the rest of the genetic material in homologous chromosomes is switched. The CO formed via the ZMM pathway is interfering (type I), which means that the occurrence of a CO impedes the formation of nearby CO. In the *Mus81* protein dependent pathway, although the intermediates are unknown (in plants) they result in either CO or dissolve and form an NCO via SDSA. The COs formed via the *Mus81*-dependent pathway are non-interfering (type II), which means that CO formation does not interfere with the formation of other CO. Figure 1.1 shows the scheme summarizing CO and NCO formation from DSB (adapted from Wang *et al.* (2018) and Mercier *et al.* (2015)).

Although the major global recombination mechanisms are known, we still lack insight into why recombination occurs where it does. Understanding the preferred locations of meiotic recombination, if any, will not only contribute to our fundamental knowledge of how species evolve via meiosis but also provide breeders with the opportunity to make better use of genetic variability in crops as well as wild relatives of crops via introgression breeding. To understand how the recombination mechanism introduces CO in certain locations, the first step is to locate the products of recombination. Tree types of analysis are used to locate recombination sites on chromosomes. The first measures the elongated chromatin via electron microscopy. This allows scientists to visualize so-called recombination nodules of homologous chromosomes (protein complexes in chiasmata where recombination occurs) as dark spots in two-dimensional spreads. The location and distance of these nodules on the chromatin can be measured in micrometers and evaluated per unit length of spread chromatin (Anderson and Stack, 2005). A second and more precise analysis method, replacing the visualization technique, is based on chromatin immunoprecipitation sequencing (ChIP-Seq, Chouaref *et al.,* 2018) or immunofluorescent *in situ* hybridization (FISH, Sepsi *et al.*, 2018) of recombination-specific proteins bound to chromosomes. By using specific antibodies, the protein complexes together with the DNA that they bind are extracted and subsequently sequenced. Then, these sequences are mapped to the genome to locate the recombination or CO positions. FISH and ChIP-Seq methods are laborious, low -throughput, and limited by the selection and identification of proper proteins and antibodies. Finally, a third type of analysis to detect

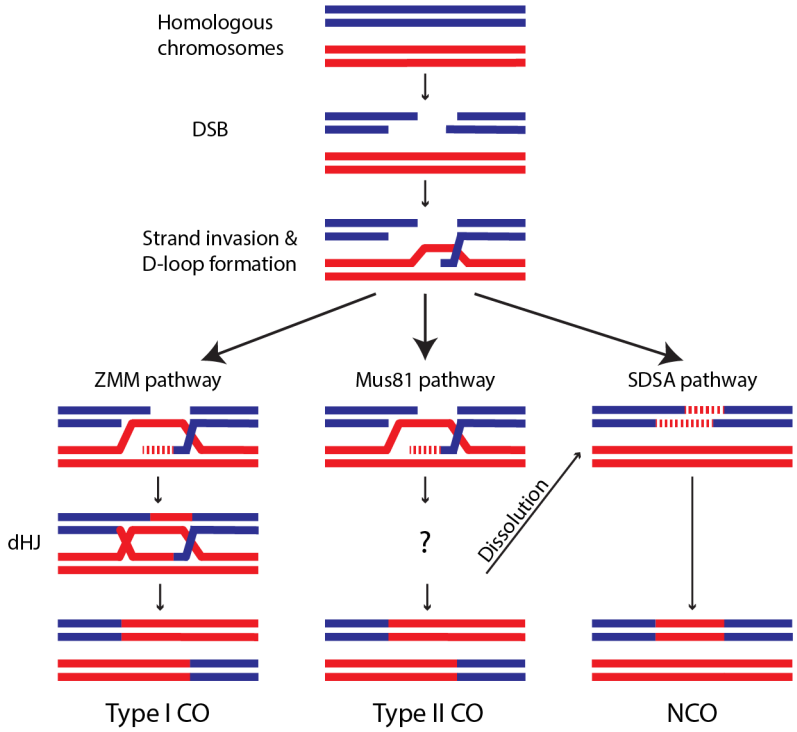

Figure 1.1 A schematic model for meiotic recombination adapted from Wang *et al.* (2018) and Mercier *et al.* (2015). Meiotic recombination starts with alignment of homologous chromosomes (depicted in red and blue). Then double-strand breaks (DSBs) form and invade a non-sister chromatid to form a D-loop. In the ZMM pathway, DNA synthesis (dashed lines) and ligation lead to the formation of double Holliday junctions (dHJ), which are primarily resolved as interfering (type I) crossovers (CO). In the Mus81 pathway, unknown recombination intermediates result in either non-interfering (type II) CO or they are dissolved into non-crossovers (NCO) via the third pathway, Synthesis-dependent strand annealing (SDSA). This pathway also results in NCO from the rejected strands after D-loop formation.

recombination is based on growing a hybrid and sequencing its whole genome. In plant studies, the most common experimental design is crossing different plants with known genetic differences. The hybrid offspring possess a mosaic of genetic material of their parents. The breakpoints in the mosaic can be easily identified since the genetic blocks in the mosaics can be traced back to each parent. Several mating designs are available to generate hybrids: recombinant inbred line (RIL) populations, intermated recombinant inbred line (IRIL) populations and multiparent advanced generation intercross populations (MAGIC, Kover *et al.*, 2009). Among these, RILs provide the most direct way of finding breakpoints as the resulting hybrids have two parents providing distinct markers. Moreover, advanced generations produced by selfing a prior generation become more and more homozygous, allowing a clear separation of parental markers. The widespread use of sequencing over the last two decades has increasingly allowed to detect COs (and gene conversions) by marker-based identification of recombination breakpoints. The precision of detecting the location of recombination breakpoints depends on the density of the markers. If COs are inferred from quantitative trait locus (QTL) maps (Huang *et al.*, 2011) the resolution is on average at the megabase level. It is possible to reach kilobase level resolution if CO borders are based on SNP arrays (Higgins *et al.*, 2018) or, even higher, base pair level resolution if markers are derived from whole genome sequence (WGS) data (Wijnker *et al.*, 2013).

Identifying recombination breakpoints creates the opportunity to investigate the distribution of COs over the genome and a chance to understand the patterns of CO from the aspect of genomic and genetic features. In the early 20th century, it was postulated that the distribution of COs follows a Poisson distribution over the chromosomes, under the assumption of no interference between COs (Haldane, 1919). Later, it was found that the majority of COs was subject to interference (reviewed in Otto and Payseur, 2019) In Arabidopsis, 85% of COs was of the interfering type (Osman *et al.*, 2011). The remaining 15% of the COs (non-interfering type) followed the Poisson distribution. Overall, COs were found to be mostly in gene-rich euchromatic regions of chromosomes (lightly packed chromatin) at the distal ends (sub-telomeric) and clustered in certain regions of the genome, which were called recombination hotspots. A few CO were found in the heterochromatin (densely packed, epigenetically silenced regions) and CO are generally suppressed in centromeres, as recombination in that area can cause segregation errors and result in non-viable gametes (Nambiar and Smith, 2016). Analysing CO hotspots revealed that COs were associated with certain genomic and genetic features. By 2016, based on numerous studies performed in Arabidopsis, it became clear that recombination breakpoints are correlated with promoter regions of genes (Choi *et al.*, 2013), specific sequence motifs (Wijnker *et al.*, 2013; Choi *et al.*, 2013), certain transposable elements (DNA transposons and retrotransposons) and epigenetic modifications (Yelina *et al.*, 2015; Colomé-Tatché *et al.*, 2012; Mirouze *et al.*, 2012) such as DNA methylation and histone methylations. Moreover, similar features were found to be associated with recombination consistently in different maize lines, suggesting that recombination is predictable (Rodgers-Melnick *et al.*, 2015).

The increasing knowledge on crossover breakpoints generates novel opportunities to control recombination in order to improve crops. There are several ways in plant breeding to control recombination such as elevation of breakpoints. Several strategies were developed to increase the recombination rate by increasing the number of DSBs via e.g. environmental stress (de Storme and Geelen, 2014) and UV treatment (Ries *et al.*, 2000). Another way to control recombination is through selection of plants to be crossed. For instance, introgression breeding attempts to introduce favourable genes from wild relatives to elite crops. This is particularly hard when there is strong linkage, i.e. when these favourable genes can only be inherited together with other undesired genes. A way to break this linkage is to select an intermediate species as a bridge in crossing, so that genomic differences between species are reduced in steps, which allows alignment of chromatids creating opportunity for recombination in chromosomes (Canady *et al.*, 2006).

To sum up, meiotic recombination causes large genomic changes such as chromosomal rearrangements. They arise from crossovers between homologous chromosomes, that can also lead to small genomic/genetic changes such as gene conversions.

1.1.2 Structural variations

Structural variations (SVs) are regions differing between genomes, generally defined as sequences longer than 50bp. Note that small insertions and deletions (depending on the read length) can be detected by standard variation calling pipelines, such as GATK, next to single nucleotide variations. There is no upper limit on the size of an SV; the main limitation lies in the SV calling algorithms and the sequencing technology used to detect them. SVs are created via various cellular mechanisms such as recombination errors, DNA break repair errors, and replication errors during meiotic or mitotic cell division (reviewed in Gabur *et al.*, 2019). Genome duplications, chromosome breaks and chromosome fusions contribute to generation of large SVs. Contrary to most SVs, mobile elements have their own way of generating variation

within the genome (see section 1.1.3). Different SV types (Alkan *et al.*, 2011) can be distinguished based on how they affect orientation, relative position, or presence/absence of sequences in the genome. Inversions affect the orientation of sequences, mostly comprising multiple genes. Presence/absence variation (PAV) of sequences is due to generic insertions (INS) and/or deletions (DEL), mobile element insertions (MEI) and T-DNA insertions. Duplications can be observed as tandem (two or more copies next to each other) or interspersed (copies located further away from each other), comprising repeats and transposable elements. Translocations are observed as chromosomal regions that move partially or fully within the same or to another chromosome. Similar to translocations, interchromosomal (homeologous) exchanges seen during meiosis in polyploid organisms are also regarded as structural variations (Schiessl *et al.*, 2019). Next to basic SV types, variations can also be categorized using the term copy number variation (CNV), which represents deletion or duplication of genes or repetitive elements (Girirajan *et al.*, 2011).

Several methods are used to detect SVs. These include genomic hybridization, multiple PCR analyses, SNP arrays and NGS data. SV detection from NGS data involves mapping sequenced reads or genome maps to a reference genome. Many tools are developed for short read mapping followed by SV detection from read alignments (Kosugi *et al.*, 2019), since short read data is more readily available for a genome of interest than a complete assembly or long read data. Most short read-based tools are benchmarked on a single species and developed for a specific sequencing technology (read type). Moreover, most tools can only detect a subset of SV types rather than all. As a result, performance depends on the species, sequencing technology and particular SV type. Due to the different limitations of each tool, it is advised to use multiple tools to obtain reliable results and evaluate the results on the grounds of the research question at hand (Wijfjes *et al.*, 2019; Zarate *et al.*, 2020; Kosugi *et al.*, 2019).

So far, several SV studies on different plants have been published. In Arabidopsis, a PAV study revealed 2407 new genes in 80 Arabidopsis accessions on top of the 27416 protein coding genes listed in the reference genome annotation, TAIR10 (Tan *et al.*, 2012). In asexual diploid potato, around 30% of genes, covering 30.2% of the potato genome (219.8 Mb of ~700 Mb), were affected by at least partial duplication or deletion in 12 samples, revealing the highly heterogeneous nature of the potato genome (Hardigan *et al.*, 2016). In rice, ~10% of protein-coding genes were genome specific in 3 samples (Schatz *et al.*, 2014). In barley, CNVs affected 9.5% of coding genes in 14 different accessions (cultivars and wild barley) spanning 14.9% of the 4.8 Gb sized genome (Muñoz-Amatriaín *et al.*, 2013). More studies revealing the number of SVs, especially CNV and PAV, are reviewed in Zhang *et al.* (2018). Overall, the number of SVs observed in plant genomes depends on i) the number of samples studied, ii) the sample diversity, such as whether wild relatives of crops or landraces are included in the study and iii) the properties of the studied plant species, for instance whether it is inbred or out-crossed, and its genetic diversity prior to domestication. Literature thus suggests that each plant is unique based on its genomic composition and the variation in it needs to be extensively studied to catalogue rare variants that may contribute to adaptation to environmental changes.

Besides being a measure of genetic diversity, SVs are important because they have various effects on phenotypes. Differences in phenotype can be caused by an SV affecting the expression of key genes or regulators. For instance, an increase in the copy number of *Ppd-B1* was found to be associated with early flowering in wheat (Díaz *et al.*, 2012). SVs can also play a role in environmental adaptation, e.g. through regulation of stress response genes and disease resistance genes. For instance, biotic and abiotic stress adaptation was observed to be caused by CNVs in soybean (McHale *et al.*, 2012), Arabidopsis (DeBolt, 2010) and potato (Hardigan *et al.*, 2016). In addi-

tion to gene regulation, SVs have an effect on diversification. SVs are associated with domestication related genes in sorghum (Mace *et al.*, 2013), population differentiation in rice (Yu *et al.*, 2013) and sex determination in cucumber (Zhang *et al.*, 2015). More comprehensive examples of plant SVs and associated traits are found in Gabur *et al.* (2019). Due to their various effects on phenotypes, SVs are of interest to plant breeders.

1.1.3 (Retro)transposons

Transposable elements (TEs), also called mobile elements, are sequences containing genes and repetitive elements which can move throughout the genome. TEs are classified into categories based on their mode of transposition. There are two main groups, DNA transposons and retrotransposons, which use DNA and RNA intermediate to transpose, respectively (Wicker *et al.*, 2007). Briefly, transposition through DNA occurs in "cut-paste" mode: the TE cuts itself from the genome, localizes to a new position and inserts itself. Transposition through RNA intermediates occurs in "copy-paste" mode and involves transcription, encapsulation out of the nucleus, reverse transcription, entering the nucleus and inserting cDNA to the new position in the genome (Schulman, 2013). TEs are an important source of genetic diversity and determinants of genomic structure. Since free movement of TEs is a threat to genome stability, transposon activity is regulated by epigenetic modifications such as DNA methylation and chromatin remodelling (Lisch, 2009; Bucher *et al.,* 2012).

TEs are further divided into superfamilies and families based on the structure and content of the TE (for a review, see Makalowski *et al.,* 2019). Basically, the structure of a TE describes how its repeats and genetic content are organized. For instance, LTR retrotransposons are structured as symmetric, identical long terminal repeats flanking the essential genes needed for retrotransposition. TEs having the same structure can be further grouped into different families based on their sequence identity. The families can have different sizes, where the number of members in a single family can be as little as ten or as many as thousands. TE families can have various sequence lengths, from a few hundred bp to tens of Kb.

TEs are identified and classified in the genome using several strategies (Goerner-Potvin and Bourque, 2018). Using whole genome assemblies, novel TEs can be detected and then grouped into classes and families based on their structure, by searching for distinct parts of a TE such as repetitive regions with specific genes in between. If a TE reference sequence is known, homology based methods are used to find similar sequences in the new genomes via a sequence match algorithm. With the availability of reference genomes and reference TE sequences, TEs can be found in multiple samples in a population in WGS data. Presence/absence of TEs in a population, called TE insertion polymorphisms (TIPs), can be identified from WGS data with strategies similar to identification of insertions. Given that these TE insertions are repetitive, like duplications, or transpose along the genome, common SV identification tools are not suitable to distinguish TE from duplications/insertions/translocations. Therefore, specialized tools have been developed to detect TE insertions. However, the performance of such tools differs for TE families and species under study (Vendrell-Mir *et al.*, 2019).

TE sequences form a large proportion of genomes in most plants (Chen *et al.*, 2018), but the TE content of plant genomes differs from species to species. For example, in Arabidopsis, only 5-10 % of the genome is covered by TE, whereas in tomato and maize, 70% resp. 80% of the genome is composed of TE. Besides the overall number of TEs, the TE content of the genome also differs in type. In Arabidopsis and tomato, most TEs are *Copia* or *Gypsy* type of retrotransposons whereas DNA transposons are the main type of TE in maize. Despite this varying content and type per species, DNA

transposons and *Copia* type retrotransposons are generally observed near genes and in gene-rich euchromatic regions of the genome while *Gypsy* type retrotransposons are generally packed in heterochromatin, which is positioned in pericentromeric regions in plant genomes (reviewed in Galindo-González *et al.*, 2017). In heterochromatin, retrotransposons are found in methylated and thus inactive form. These accumulate mutations and diverge from the original TE, contributing to non-functional genomic diversity.

TE insertions show variation at the population level. Sequencing of 80 Arabidopsis accessions from eight populations showed that ~80% of TEs were partially or completely absent from the genomes of at least one of the individuals (Cao *et al.*, 2011). In 201 *A. thaliana* genomes, 2,311 polymorphic TEs were found (Li *et al.*, 2018). In 3,000 rice cultivar genomes, 50,000 TIPs were found in 32 families of retrotransposons (Carpentier *et al.*, 2019). In 602 tomato cultivars, landraces and wild relatives, 6,906 TIPs were found in 337 TE families (Domínguez *et al.*, 2020). In only four maize genomes, 400,000 TIPs were identified in 1.6 Gb of variable TE sequence, mostly composed of LTR retrotransposons (Anderson *et al.*, 2019).

TE insertions have various effects on genomic diversity and phenotypes. TEs can jump into exons and promoters to interrupt the function of genes or bring new promoters or enhancers to alter gene expression (Hirsch and Springer, 2017). TE mobility also contributes to generating other types of SV, such as CNV or translocations (Lisch, 2013). Moreover, TE insertions contribute to adaptive genome evolution, for instance flowering time variation in Arabidopsis (Strange *et al.*, 2011) and agricultural traits such as tomato ripening (Jouffroy *et al.*, 2016) and the jointless trait (jointless fruit pedicels) in tomato (Soyk *et al.*, 2017; Alonge *et al.*, 2020). There are at least fifty published plant phenotypic variations affected by TEs (see Wei and Cao, 2016). Since TEs have substantial effect on many traits, knowing TE polymorphisms helps towards understanding variation in traits and selecting for relevant traits in plant breeding programs.

1.2 Sequencing technologies

Sequencing technology has come a long way since the introduction of Sanger sequencing of DNA fragments in 1977 (Shendure *et al.*, 2017). In Sanger sequencing, the length of DNA sequence was limited to approx. 1kb, while base calling accuracy was high, with only 0.05% error rate. Next, sequencing technology moved to the era of high throughput sequencing, starting with the introduction of 96-capillary AB 370 and 377 sequencers, which was used heavily in the Human Genome Project (1990-2003) which sequence the first reference human genome over 10 years at a cost of $2.7 billion. Over the subsequent decades, various companies released technologies with different operating principles that all aimed to generate more data in a single run - thus reducing the cost and increasing the length of sequences. Among these, Illumina dominated the market. Over the years, the sequence length of Illumina short reads increased from 35 bp single-reads to 2 x 250 bp paired-end short reads with insert sizes of 500-750 bp and error rates of 0.1-0.25% per nucleotide (Pfeiffer *et al.*, 2018). Paired-end technology is popular as it sequences both ends of a larger DNA fragment (of a certain insert size), thus providing somewhat longer-range contiguity information. Overall, DNA sequencing costs dropped from $10,000,000 to $0.01 per Mb between 2001 to 2020 (Wetterstrand, 2020).

Although short-read technology is preferable for certain applications such as massive resequencing of highly similar genomes, its use is limited in generating new reference genomes for more diverse and complex fungal, animal and plant species, and in detecting large SVs. Third generation sequencing was developed to provide contiguity over larger stretches of DNA, reaching over 1Mb with median length of

~20 kb, but with a higher error rate than short-read technology. The error rate of nanopore sequencing as developed by Oxford Nanopore Technologies is currently ~3% (https://nanoporetech.com/accuracy) and the error rate of continuous long read (CLR) data obtained using single molecule real-time (SMRT) sequencing developed by PacBio is 11-15% (Rhoads and Au, 2015). A recent development of PacBio, which uses circular consensus sequencing (CCS) to generate high fidelity reads (HiFi reads), has error rates less than 1% (Wenger *et al.*, 2019) at slightly shorter read lengths (10-20 Kb) than CLR. Although the cost of long read sequencing was much higher than that of the already established short-read technology when it was first introduced, it is rapidly declining. It is now possible to *de novo* assemble large plant genomes as large as the coast redwood tree, with a genome size of 26.5 Gb (Redwood Genome Project), with long reads.

Novel technologies like synthetic long reads, i.e. short reads produced from a single molecule (up to 100 kb long), promise both high-accuracy base pair data as well as long contiguous sequences. An example of this technology was developed by 10X Genomics, which tags reads coming from single DNA fragments with unique molecular barcodes (reviewed in Goodwin *et al.*, 2016). Reads with the same barcode can then be assembled into a long DNA fragment. Similar linked-read technologies are promising good quality in long-range sequencing, for example single-tube transposase enzyme linked long-read sequencing (TELL-Seq, Chen *et al.*, 2020), and single tube long fragment read (stLFR, Wang *et al.*, 2019). Other technologies such as Hi-C and optical mapping serve to determine chromosome organization by focusing on the location of specific sequences on the chromosome or on a given fragment, without sequencing the full overall sequence. Hi-C couples proximity-based ligation and high-throughput sequencing. Proximal sequences which are close to each other on the folded chromatin can be identified with the Hi-C method (Lieberman-Aiden *et al.*, 2009). In optical mapping technology, the order of specific sequences -nicked by restriction enzymes- are determined under light microscopy (reviewed in Pyle *et al.*, 2018) These methods are mostly used as supplementary to scaffold the contiguous sequences or contigs obtained by assembling short or long reads, to construct genomes *de novo* or to detect very large structural variations (Burton *et al.*, 2013; Tang *et al.*, 2015).

There are many varieties of DNA sequencers which can generate different types and amounts of data. The choice of DNA sequencer(s) depends on the research question at hand. The experimental design is based on i) the size of the candidate genome; ii) the genome coverage needed, based on the research question; and iii) the number of accessions needed to answer the research question. For larger genomes (i.e. up to and over Gb length), a sequencer with higher output is desired. High coverage is needed for a *de novo* assembly (>50x), whereas medium coverage is enough for SV detection (10-20x) and even lower coverages (<5x) are sufficient for SNP identification, for example via genotyping-by-sequencing (GBS, Malmberg *et al.*, 2018). For this low coverage to be effective, usually accurate Illumina reads are chosen. In plants, one accession is generally used to generate a reference genome, but many should be sequenced for population level analyses, often influencing the choice of sequencing technology given a limited budget.

Currently we are in a shift from using short reads to using long reads for many applications, such as large variation detection and sequencing of new genomes. Since short read data became widely available in the mid-2002, costs have dramatically reduced and a large number of specialized algorithms handling short reads have been developed. These are of course still advantages to short read data, such as the low cost which is essential in population level studies and the accuracy which is helpful in resolving genomes at the base pair level. While long read data is essential to resolve large or complex structural variations and repetitive elements, until it becomes

accessible and affordable for population level studies, we will continue to use short reads to unlock the genetic and structural differences within and between related species.

1.3 Contribution of this thesis to the field

The aim of this thesis is to catalogue and study different aspects of large-scale genomic variation in plants. This may help to improve our understanding of the mechanisms that generate this variation, and may provide leads for breeders to improve crops by classical or precision breeding. Because meiotic recombination is the main source of genetic variation, it is important to gain insights on recombination-prone genomic regions. Also, due to the recently discovered effects on agricultural traits, genetic variation such as SVs and TEs are attracting geneticists' attention. The presence and influence of these variations are still not well studied in many species, and studying them is interesting because they contain traces of crop evolution and history.

In chapter 2, to understand preferences for meiotic recombination at certain genomic locations, we identify and locate homeologous COs in an interspecific cross of tomato, specifically based on WGS data of a RIL population. This results in a first set of high-resolution breakpoint locations for interspecific meiotic recombinations in plants. We investigate what genomic features are associated with recombination breakpoints.

In chapter 3, to further understand the underlying mechanism of CO formation, we will reveal the relationship (correlations) between homeologous CO breakpoints and various genomic features. Also, we will show that the relationship between CO and these features are conserved among homologous chromosomes in various plant species. We take a unique approach to reveal these connections by predicting the breakpoints through machine learning.

In chapter 4, we study SVs because of their effect on regulation of gene expression causing phenotype diversity and their contribution to large-scale chromosome evolution. We identify SVs in a relatively large melon population (100 genomes) compromising wild species. We unravel the breeding history by studying the genetic composition of melon groups using SVs instead of the previously used SNPs. We investigate the phylogenetic relations in melon groups and the functional differences in these groups based solely on SVs.

In chapter 5, we turn our focus to transposable elements due to their ability to translocate genomic information, thus increasing genetic diversity, changing gene expression and potentially generating new phenotypes. We attempt to discover the genetic diversity from a transposable element insertion polymorphism point of view and inventory retrotransposon positions in a population of 60 tomato cultivars. We also look for active retrotransposons which may have effects on the phenotype. This will aid the breeding community, as active retrotransposons may serve as a basis for technologies aiming to manipulate genetic diversity thus generating favoured phenotypes in plant breeding. We detect retrotransposon insertions based on known TE reference sequences using WGS data.

Finally, in the last chapter, I discuss decisions we have made at the time of the studies, current developments and my opinion on what the future holds for genome biology research.

1.4 References

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

Alonge, M., Wang, X., Benoit, M., *et al.* (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**, 145-161.e23.

Anderson, L.K. and Stack, S.M. (2005) Recombination nodules in plants. *Cytogenet. Genome Res.*, **109**, 198–204.

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.

Bastin, J.-F., Clark, E., Elliott, T., *et al.* (2019) Understanding climate change from a global analysis of city analogues J. A. Añel, ed. *PLoS One*, **14**, e0217592.

Bucher, E., Reinders, J. and Mirouze, M. (2012) Epigenetic control of transposon transcription and mobility in Arabidopsis. Curr. Opin. Plant Biol., **15**, 503–510.

Burton, J.N., Adey, A., Patwardhan, R.P., Qiu, R., Kitzman, J.O. and Shendure, J. (2013) Chromosome-scale scaffolding of de novo genome assemblies based on chromatin interactions. Nat. Biotechnol., **31**, 1119–25.

Canady, M. a., Ji, Y. and Chetelat, R.T. (2006) Homeologous recombination in Solanum lycopersicoides introgression lines of cultivated tomato. *Genetics*, **174**, 1775–1788.

Cao, J., Schneeberger, K., Ossowski, S., *et al.* (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat. Genet.*, **43**, 956–963.

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

Chen, F., Dong, W., Zhang, J., Guo, X., Chen, J., Wang, Z., Lin, Z., Tang, H. and Zhang, L. (2018) The Sequenced Angiosperm Genomes and Genome Databases. *Front. Plant Sci.*, **9**.

Chen, Z., Pham, L., Wu, T.-C., *et al.* (2020) Ultralow-input single-tube linked-read library method enables short-read second-generation sequencing systems to routinely generate highly accurate and economical long-range sequencing information. Genome Res., **30**, 898–909.

Choi, K., Zhao, X., Kelly, K. a, *et al.* (2013) Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet.*, **45**, 1327–36.

Chouaref, J., Boer, E. de, Fransz, P. and Stam, M. (2018) Protocol for Chromatin Immunoprecipitation of Meiotic-Stage-Specific Tomato Anthers. Curr. Protoc. Plant Biol., **3**, e20074.

Colomé-Tatché, M., Cortijo, S., Wardenaar, R., *et al.* (2012) Features of the Arabidopsis recombination landscape resulting from the combined loss of sequence variation and DNA methylation. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 16240–5.

DeBolt, S. (2010) Copy Number Variation Shapes Genome Diversity in Arabidopsis Over Immediate Family Generational Scales. *Genome Biol. Evol.*, **2**, 441–453.

Díaz, A., Zikhali, M., Turner, A.S., Isaac, P. and Laurie, D.A. (2012) Copy Number Variation Affecting the Photoperiod-B1 and Vernalization-A1 Genes Is Associated with Altered Flowering Time in Wheat (Triticum aestivum) S. P. Hazen, ed. *PLoS One*, **7**, e33234.

Doebley, J.F., Gaut, B.S. and Smith, B.D. (2006) The Molecular Genetics of Crop Domestication. *Cell*, **127**, 1309–1321.

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J.M., Colot, V. and Quadrana, L. (2020) The impact of transposable elements on tomato diversity. *Nat. Commun.*, **11**, 4058.

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.

Galindo-González, L., Mhiri, C., Deyholos, M.K. and Grandbastien, M.-A. (2017) LTR-retrotransposons in plants: Engines of evolution. Gene, **626**, 14–25.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J. and Anderson, L.K. (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.*, **8**, 77–84.

Girirajan, S., Campbell, C.D. and Eichler, E.E. (2011) Human Copy Number Variation and Complex Genetic Disease. *Annu. Rev. Genet.*, **45**, 203–226.

Goodwin, S., McPherson, J.D. and McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. Nat. Rev. Genet., **17**, 333–351.

Goerner-Potvin, P. and Bourque, G. (2018) Computational tools to unmask transposable elements. Nat. Rev. Genet., **19**, 688–704.

Haldane, J.B.S. (1919) The combination of linkage values and the calculation of distances between the loci of linked factors. *J Genet*, **8**, 299–309.

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.

Higgins, E.E., Clarke, W.E., Howell, E.C., Armstrong, S.J. and Parkin, I.A.P. (2018) Detecting de Novo Homoeologous Recombination Events in Cultivated Brassica napus Using a Genome-Wide SNP Array. *G3 (Bethesda).*, **8**, 2673–2683.

Hirsch, C.D. and Springer, N.M. (2017) Transposable element influences on gene expression in plants. Biochim. Biophys. Acta - Gene Regul. Mech., **1860**, 157–165.

Huang, X., Paulo, M.-J., Boer, M., Effgen, S., Keizer, P., Koornneef, M. and Eeuwijk, F.A. van (2011) Analysis of natural allelic variation in Arabidopsis using a multiparent recombinant inbred line population. *Proc. Natl. Acad. Sci.*, **108**, 4488–4493.

Jiao, Y., Peluso, P., Shi, J., *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524–527.

Jouffroy, O., Saha, S., Mueller, L., *et al.* (2016) Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics*, **17**, 624.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. and Kamatani, Y. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, **20**, 8–11.

Kover, P.X., Valdar, W., Trakalo, J., Scarcelli, N., Ehrenreich, I.M., Purugganan, M.D., Durrant, C. and Mott, R. (2009) A Multiparent Advanced Generation Inter-Cross to Fine-Map Quantitative Traits in Arabidopsis thaliana R. Mauricio, ed. PLoS Genet., **5**, e1000551.

Li, Z., Hou, X., Chen, J., Xu, Y., Wu, Q., González, J. and Guo, Y.-L. (2018) Transposable Elements Contribute to the Adaptation of Arabidopsis thaliana. *Genome Biol. Evol.*, **10**, 2140–2150.

Lieberman-Aiden, E., Berkum, N.L. van, Williams, L., *et al.* (2009) Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. Science (80-.)., **326**, 289–293.

Lisch, D. (2009) Epigenetic Regulation of Transposable Elements in Plants. Annu. Rev. Plant Biol., **60**, 43–66.

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

Mace, E.S., Tai, S., Gilding, E.K., *et al.* (2013) Whole-genome sequencing reveals untapped genetic potential in Africa's indigenous cereal crop sorghum. *Nat. Commun.*, **4**, 2320.

Makałowski, W., Gotea, V., Pande, A. and Makałowska, I. (2019) Transposable Elements: Classification, Identification, and Their Use As a Tool For Comparative Genomics. In M. Anisimova, ed. Evolutionary Genomics. Methods in Molecular Biology, vol 1910. New York, NY: Humana, pp. 177–207.

Malmberg, M.M., Barbulescu, D.M., Drayton, M.C., Shinozuka, M., Thakur, P., Ogaji, Y.O., Spangenberg, G.C., Daetwyler, H.D. and Cogan, N.O.I. (2018) Evaluation and Recommendations for Routine Genotyping Using Skim Whole Genome Re-sequencing in Canola. *Front. Plant Sci.*, **9**, 1809.

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.

Maagd, R.A., Loonen, A., Chouaref, J., Pelé, A., Meijer-Dekens, F., Fransz, P. and Bai, Y. (2020) CRISPR/Cas inactivation of RECQ4 increases homeologous crossovers in an interspecific tomato hybrid. Plant Biotechnol. J., **18**, 805–813.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. and Grelon, M. (2015) The Molecular Biology of Meiosis in Plants. *Annu. Rev. Plant Biol.*, **66**, 297–327.

Mirouze, M., Lieberman-Lazarovich, M., Aversano, R., Bucher, E., Nicolet, J., Reinders, J. and Paszkowski, J. (2012) Loss of DNA methylation affects the recombination landscape in Arabidopsis. *Proc. Natl. Acad. Sci.*, **109**, 5880–5885.

Muñoz-Amatriaín, M., Eichten, S.R., Wicker, T., *et al.* (2013) Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol.*, **14**, R58.

Nambiar, M. and Smith, G.R. (2016) Repression of harmful meiotic recombination in centromeric regions. Semin. Cell Dev. Biol., **54**, 188–197.

Osman, K., Higgins, J.D., Sanchez-Moran, E., Armstrong, S.J. and Franklin, F.C.H. (2011) Pathways to meiotic recombination in Arabidopsis thaliana. *New Phytol.*, **190**, 523–544.

Otto, S.P. and Payseur, B.A. (2019) Crossover Interference: Shedding Light on the Evolution of Recombination. Annu. Rev. Genet., **53**, 19–44.

Pfeiffer, F., Gröber, C., Blank, M., Händler, K., Beyer, M., Schultze, J.L. and Mayer, G. (2018) Systematic evaluation of error rates and causes in short samples in next-generation sequencing. *Sci. Rep.*, **8**, 10950.

Pyle, J.R., Sy Piecco, K.W.E., Vicente, J.R. and Chen, J. (2018) Optical Genome Mapping. Encycl. Chem. Process., 1–30. DOI: 10.1081/E-ECHP-140000148

Piégu, B., Bire, S., Arensburger, P. and Bigot, Y. (2015) A survey of transposable element classification systems – A call for a fundamental update to meet the challenge of their diversity and complexity. Mol. Phylogenet. Evol., **86**, 90–109.

Redwood Genome Project. Available at: https://nealelab.ucdavis.edu/redwood-genome-project-rgp/ [Accessed January 5, 2021].

Rhoads, A. and Au, K.F. (2015) PacBio Sequencing and Its Applications. Genomics, Proteomics Bioinforma., **13**, 278–289.

Ries, G., Buchholz, G., Frohnmeyer, H. and Hohn, B. (2000) UV-damage-mediated induction of homologous recombination in Arabidopsis is dependent on photosynthetically active radiation. *Proc. Natl. Acad. Sci.*, **97**, 13425–13429.

Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y. and Buckler, E.S. (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 3823–8.

Schatz, M.C., Maron, L.G., Stein, J.C., *et al.* (2014) Whole genome de novo assemblies of three divergent strains of rice, Oryza sativa, document novel gene space of aus and indica. *Genome Biol.*, **15**, 506.

Schiessl, S.-V., Katche, E., Ihien, E., Chawla, H.S. and Mason, A.S. (2019) The role of genomic structural variation in the genetic improvement of polyploid crops. *Crop J.*, **7**, 127–140.

Schulman, A.H. (2013) Retrotransposon replication in plants. Curr. Opin. Virol., **3**, 604–614.

Sedlazeck, F.J., Rescheneder, P., Smolka, M., Fang, H., Nattestad, M., Haeseler, A. von and Schatz, M.C. (2018) Accurate detection of complex structural variations using single-molecule sequencing. *Nat. Methods*, **15**, 461–468.

Sepsi, A., Fábián, A., Jäger, K., Heslop-Harrison, J.S. and Schwarzacher, T. (2018) ImmunoFISH: Simultaneous Visualisation of Proteins and DNA Sequences Gives Insight Into Meiotic Processes in Nuclei of Grasses. Front. Plant Sci., **9**.

Shendure, J., Balasubramanian, S., Church, G.M., Gilbert, W., Rogers, J., Schloss, J.A. and Waterston, R.H. (2017) DNA sequencing at 40: past, present and future. *Nature*, **550**, 345–353.

Soyk, S., Lemmon, Z.H., Oved, M., *et al.* (2017) Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. Cell, **169**, 1142–1155.

de Storme, N. and Geelen, D. (2014) The impact of environmental stress on male reproductive development in plants: biological processes and molecular mechanisms. *Plant. Cell Environ.*, **37**, 1–18.

Strange, A., Li, P., Lister, C., Anderson, J., Warthmann, N., Shindo, C., Irwin, J., Nordborg, M. and Dean, C. (2011) Major-Effect Alleles at Relatively Few Loci Underlie Distinct Vernalization and Flowering Variation in Arabidopsis Accessions M. Tsiantis, ed. PLoS One, **6**, e19949.

Tan, S., Zhong, Y., Hou, H., Yang, S. and Tian, D. (2012) Variation of presence/absence genes among Arabidopsis populations. *BMC Evol. Biol.*, **12**, 86.

Tang, H., Lyons, E. and Town, C.D. (2015) Optical mapping in plant comparative genomics. Gigascience, **4**, 3.

Tanksley, S.D. and McCouch, S.R. (1997) Seed Banks and Molecular Maps: Unlocking Genetic Potential from the Wild. *Science (80-.).*, **277**, 1063–1066.

Vendrell-Mir, P., Barteri, F., Merenciano, M., González, J., Casacuberta, J.M. and Castanera, R. (2019) A benchmark of transposon insertion detection tools using real data. Mob. DNA, **10**, 53.

Wang, O., Chin, R., Cheng, X., *et al.* (2019) Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. Genome Res., **29**, 798–808.

Wang, Y. and Copenhaver, G.P. (2018) Meiotic Recombination: Mixing It Up in Plants. *Annu. Rev. Plant Biol.*, **69**, 577–609.

Weeks, D.E., Tang, X. and Kwon, A.M. (2009) Casares' map function: no need for a "corrected" Haldane's map function. *Genetica*, **135**, 305–7.

Wei, L. and Cao, X. (2016) The effect of transposable elements on phenotypic variation: insights from plants to humans. *Sci. China Life Sci.*, **59**, 24–37.

Wenger, A.M., Peluso, P., Rowell, W.J., *et al.* (2019) Accurate circular consensus long-read sequencing improves variant detection and assembly of a human genome. *Nat. Biotechnol.*, **37**, 1155–1162.

Wetterstrand, K.A. (2020) DNA Sequencing Costs: Data from the NHGRI Genome Sequencing Program (GSP). Available at: www.genome.gov/sequencingcostsdata [Accessed December 12, 2020].

Wicker, T., Sabot, F., Hua-Van, A., *et al.* (2007) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat. Rev. Genet.*, **8**, 973–982.

Wijfjes, R.Y., Smit, S. and Ridder, D. de (2019) Hecaton: reliably detecting copy number variation in plant genomes using short read sequencing data. BMC Genomics, **20**, 818.

Wijnker, E., Velikkakam James, G., Ding, J., *et al.* (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife*, **2**, e01426.

Yelina, N., Diaz, P., Lambing, C. and Henderson, I.R. (2015) Epigenetic control of meiotic recombination in plants. *Sci. China Life Sci.*, **58**, 223–231.

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**, 649.

Zarate, S., Carroll, A., Mahmoud, M., *et al.* (2020) Parliament2: Accurate structural variant calling at scale. Gigascience, **9**.

Zhang, X., Chen, X., Liang, P. and Tang, H. (2018) Cataloging Plant Genome Structural Variations. *Curr. Issues Mol. Biol.*, **27**, 181–194.

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.

Chapter 2

Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between *Solanum lycopersicum* and *Solanum pimpinellifolium*

Published as S. Demirci, A.D.J. van Dijk, G. Sanchez Perez, S.A. Aflitos, D. de Ridder and S. Peters, *The Plant Journal* 89:554-564, 2017. https://doi.org/10.1111/tpj.13406

2.1 Summary

We determined the crossover (CO) distribution, frequency and genomic sequences involved in interspecies meiotic recombination by using parent-assigned variants of 52 F_6 recombinant inbred lines obtained from a cross between tomato, *Solanum lycopersicum*, and its wild relative, *Solanum pimpinellifolium*. The interspecific CO frequency was 80% lower than reported for intraspecific tomato crosses. We detected regions showing a relatively high and low CO frequency, so-called hot and cold regions. Cold regions coincide to a large extent with the heterochromatin, although we found a limited number of smaller cold regions in the euchromatin. The CO frequency was higher at the distal ends of chromosomes than in pericentromeric regions and higher in short arm euchromatin. Hot regions of CO were detected in euchromatin, and COs were more often located in non-coding regions near the 5′ untranslated region of genes than expected by chance. Besides overrepresented CCN repeats, we detected poly-A/T and AT-rich motifs enriched in 1-kb promoter regions flanking the CO sites. The most abundant sequence motifs at CO sites share weak similarity to transcription factor-binding sites, such as for the C2H2 zinc finger factors class and MADS box factors, while InterPro scans detected enrichment for genes possibly involved in the repair of DNA breaks.

Meiotic recombination is a key biological process that generates genetically different haploid gametes in germ cells by reshuffling segments from the parental chromosomes. Understanding the genomic features determining the location and frequency of crossovers (COs) is of both fundamental importance for population genetics and chromosome evolution and practical importance for breeding. Breeders rely on homo(e)ologous meiotic recombination to generate allelic variation in crop species from their wild relatives such as in tomato, in a process known as introgression hybridization breeding. In many cases introgressive hybridization breeding is a time-consuming and costly process. Usually, recombination involves the reciprocal exchange of large chromosome fragments (crossovers) that, besides favourable genes, may also include additional unwanted genes. Breeders aim to maintain the chromosomal regions of interest while removing the unwanted traits from the wild relative by recurrent backcrossing with the crop, followed by trait selection. However, crossing barriers and linkage drag are well-known phenomena that limit the use of wild germplasm. Our objective is to identify key genomic features determining the position and frequency of meiotic CO. These can assist in the selection of compatible breeding parents that can be used for efficient recombination of chromosomal regions with favourable traits.

Previously, recombination for members of the tomato clade has been extensively studied. For *Solanum lycopersicoides* × *Solanum lycopersicum* hybrids the interspecies recombination rate was found to be 10% of the intraspecies recombination frequency (Canady *et al.*, 2006), while recombination within segments derived from *Solanum pennellii* and *Solanum hirsutum* was at 15–30% of intraspecies levels (Van Wordragen *et al.*, 1996; Monteforte and Tanksley, 2000), consistent with the notion that the frequency of interspecies recombination in the tomato clade is correlated with the degree of sequence divergence (Canady *et al.*, 2006). The recombination frequency for *S. lycopersicum* × *S. pennelli* hybrids has been determined using a RFLP analysis (De Vicente and Tanksley, 1991), and in *S. lycopersicum* × *S. lycopersicum* crosses by counting recombination nodules in stained paired homologous chromosomes at the pachytene stage of prophase I (Stack and Anderson, 1986; Sherman and Stack, 1995). According to these studies, recombination events in homologous segments usually begin and accumulate at the distal ends of the chromosomes

where euchromatin is lightly packed and is suppressed or even absent in the distal heterochromatin and the large pericentromere regions. Using chromosome-arm-specific probes Anderson *et al.* (2014) showed synapsis was more often initiated from long arms than from short arms. Early synapsed segments predominantly contained MLH1-positive recombination nodules (RNs) correlating with a higher proportion of class I (interference dependent) COs compared with late synapsed chromosome segments (short arms and pericentromere) in which more MLH1-negative recombination nodules were found. In human, mouse and plants there are recombination 'hotspots' consisting of regions of 1–10 kb (McVean *et al.*, 2004; Drouaud *et al.*, 2013), implying that homologous meiotic recombination generally occurs at non-random locations (Lichten and Goldman, 1995; Kauppi *et al.*, 2004). In contrast, recombination occurs without any hotspots in *Drosophila* species (Heil *et al.*, 2015), suggesting that recombination machineries operate differently between organisms.

In *Arabidopsis thaliana* and *Zea mays* sequence motifs have been found enriched in recombination sites. These include poly-A stretches and a CTT/GAA palindromic repeat in *A. thaliana* and a few diverse motifs in maize (Choi *et al.*, 2013; Wijnker *et al.*, 2013; Rodgers-Melnick *et al.*, 2015). Moreover, homologous meiotic recombination in Arabidopsis and yeast was found around transcription start sites (TSSs) in promoter regions and nucleosome-depleted chromatin-accessible regions (Pan *et al.*, 2011; Choi *et al.*, 2013; Wijnker *et al.*, 2013; Shilo *et al.*, 2015), suggesting that open chromatin provides 'windows of opportunity' for recombination (Heil *et al.*, 2015). In maize, specific hotspots, such as in the Bronze and a1 domains, have been detected in gene-rich regions (Brown and Sundaresan, 1991; Fu *et al.*, 2001). Nonetheless, until now detailed information on sequence motifs and CO position with nucleotide-level resolution for homeologous chromosomes has not been reported for species in the tomato clade. Furthermore, information on proteins that are involved in binding of specific DNA sequences at hotspots of recombination is limited. In humans, apparently, 40% of meiotic recombination hotspots involve a histone lysine *N*-methyltransferase protein PRDM9 (Myers *et al.*, 2010; Parvanov *et al.*, 2010), a member of the C2H2 zinc finger gene family that is associated with a 13-mer sequence motif. No PRDM9 homologue or other factors determining recombination hotspots in homeologous chromosomes, such as specific sequence elements, have been identified in plants as of yet (Bauer *et al.*, 2013).

Here we report on the identification of COs in homeologous chromosomes, their frequency and distribution, local sequence features and chromosome topological context involved in tomato hybrids, by analysing 60 F_6 generation recombinant inbred lines (RILs) obtained from a cross between *S. lycopersicum* cv. Moneymaker (crop tomato) and a wild relative, *Solanum pimpinellifolium*. Mapped CO sites and identified hot regions and their genomic features show: (i) overrepresented sequence motifs and repeats at recombination sites; (ii) a non-random CO distribution with respect to hot and cold regions; and (iii) gene features suggesting a potential relation between recombination and transcription.

2.3 Results

2.3.1 Heterozygosity

We found 3 893 390 discriminative alleles between the *S. lycopersicum* Moneymaker and *S. pimpinellifolium* RIL parents. After applying a strict heterozygosity filtering in the haplotyping step, the average density was 1.2–4.7 kb^{-1}. This is in agreement with the single nucleotide polymorphism (SNP) density of 0.03 and 0.58% that was previously found for the Moneymaker and *S. pimpinellifolium* RIL parents, respectively, when compared with the *S. lycopersicum* cv. Heinz reference genome (Aflitos *et al.*, 2014). We observed 10% of heterozygosity on average in 60 F_6 RILs (Table S1),

which was significantly higher than the expected 3%. Previously, RILs from a cross between *S. lycopersicum* cv. UC204B and *Solanum cheesmaniae* showed an average heterozygosity level of 15% in the F_7 generation, which was significantly higher than the expected 1.5% (Paran *et al.*, 1995). Cross-pollination between the inbred lines could have caused the (re)introduction of heterozygosity in these RILs. Alternatively, a higher level of heterozygosity could have been the result of selection against parental allele combinations during RIL propagation. Taking into account possible effects arising from cross-pollination, eight F_6 RILs exceeding 15% heterozygosity were excluded, leaving 52 RILs for subsequent detection of recombination events.

2.3.2 Crossover distribution and frequency

Since we could not determine the precise CO site down to the nucleotide level, we delineated CO regions based on the observed discriminating alleles and inferred haplotype shifts. In total we detected 1015 such regions with an average interval length of 1577 ± 1660 bp (median 1063 bp) (Data S2). The length distribution of the CO region is shown in Figure 2.1(a). These regions are nearly uniformly distributed over the chromosomes in the euchromatin (Figure 2.1b), although this distribution is not uniform when normalized by euchromatin length. A higher CO frequency is apparent in most of the short arms (Figure 2.1c).

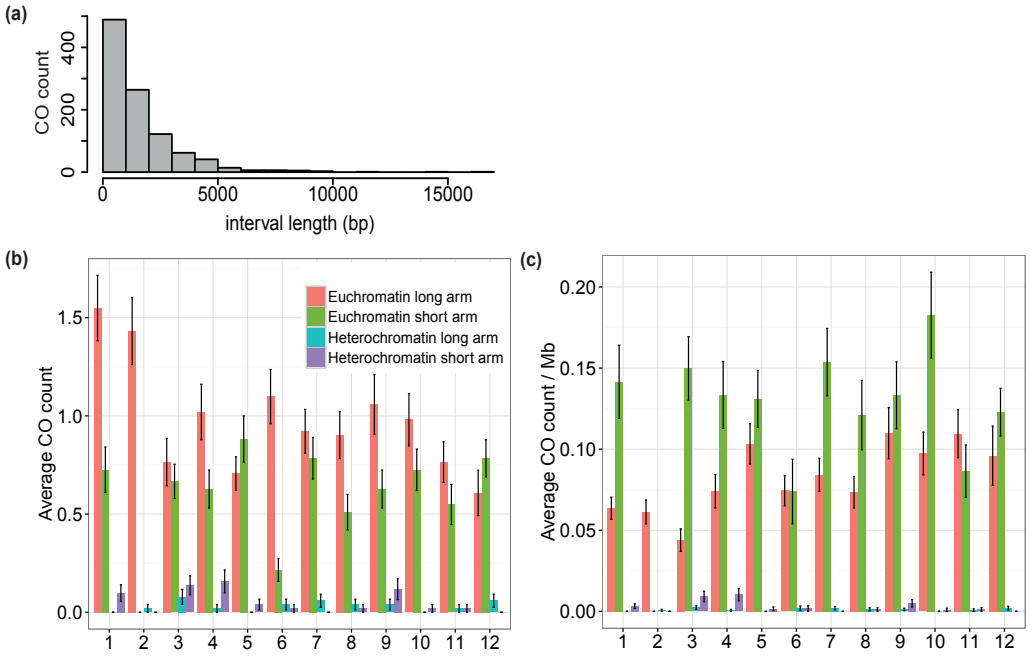

Figure 2.1 The distribution of CO regions. **(a)** Interval length distribution for CO regions in the 52 RILs. **(b)** The average CO count of 52 RILs grouped by euchromatin/heterochromatin of each chromosome. **(c)** The average CO count normalized by the size of euchromatin/heterochromatin of each chromosome.

We next inferred the CO landscape for each of the 12 chromosomes in 52 RILs. Figure 2.2(a) illustrates the consecutive set of parental blocks from which the CO profile for chromosome 6 was inferred. The profiles for the other 11 chromosomes are displayed in Figure S1 In general, COs clearly accumulate distally on both arms of each chromosome, where euchromatin is present, while other so-called 'cold regions' particularly found in heterochromatin are almost devoid of CO. For example, a large heterochromatic cold region spanning 29.7 Mb is indicated by a double-headed red horizontal arrow in the density plot overlapping with the pericentromeric region of chromosome 6 (Figure 2.2b, c), confirming previous reports that heterochromatin is largely devoid of recombination (Sherman and Stack, 1995; The Tomato Genome Consortium, 2012). In the other chromosomes, the large pericentromeric heterochromatic regions were also devoid of COs (Figure S1c). In addition, we found 122 euchromatic cold regions larger than 100 kb (Figure 2.3, Table S2). Nevertheless, 51 COs occur in heterochromatic regions, although most of these appear close to the euchromatin–heterochromatin borders. A few small domains with COs were observed in the pericentromere of chromosomes 2 and 12, far from euchromatin–heterochromatin borders (Figures 2.2 and S1). One possible explanation could be that the CO has taken place within small euchromatic islands of genes in the pericentromere. Indeed, earlier studies indicated substantial numbers of transcribed genes in pericentromere regions (Peters *et al.*, 2009; The Tomato Genome consortium, 2012). Furthermore, we identified seven significantly CO-dense regions, so-called 'hot regions', in tomato chromosome arms 1L, 3S, 3L, 6S, 6L, 9L and 12L (Figures 2.3 and S2) of which, for example, a region of 100 kb in chromosome 6L comprised six CO events. These hot regions could potentially arise from erroneous SNP calling, for example caused by transposable elements, resulting in structural variation (Qi *et al.*, 2013). However, we did not find DNA transposon footprints at the hot regions. To further rule out false positives, regions that contain multiple COs ($n = 49$), including the seven hot regions, were assessed for repetitive domains. The coverage of each region to which multiple COs map was compared with the average genome coverage for both *S. lycopersicum* cv. Moneymaker and *S. pimpinellifolium* RIL parents (Table S3). Results do not point to copy number variation in the corresponding regions, suggesting that the parent-assigned variants (PAVs) originate from unique parental genomic regions that were faithfully mapped for each of the 49 recombination regions.

In contrast, no CO was observed in a small euchromatic region at 40.5–42.5 Mb of chromosome 6L (Figure 2.2b). Previously, this region was identified as an introgression from *S. pimpinellifolium* into the *S. lycopersicum* Heinz and Moneymaker accessions (Aflitos *et al.*, 2015). Also for the chromosome 3L region at 64.6–67.0 Mb we could not detect COs, which perhaps represents another introgression from *S. pimpinellifolium*. The number of SNPs between the two parents in these euchromatic regions (Figures 2.2d and S1) is very low, apparently restricting the power to detect recombination events. Interestingly, a few small cold regions of 1–2 Mb in size with a density of 4 SNP kb^{-1} were observed in chromosomes 5, 7 and 10 (Figure S1, Table S2). Although we currently lack detailed information, structural differences could prevent proper chromosome pairing and recombination in these domains, although the relatively low SNP density would argue against that. Alternatively, these domains might perhaps comprise protected blocks of genes that suppress recombination between homeologous chromosomes (Bedinger *et al.*, 2010).

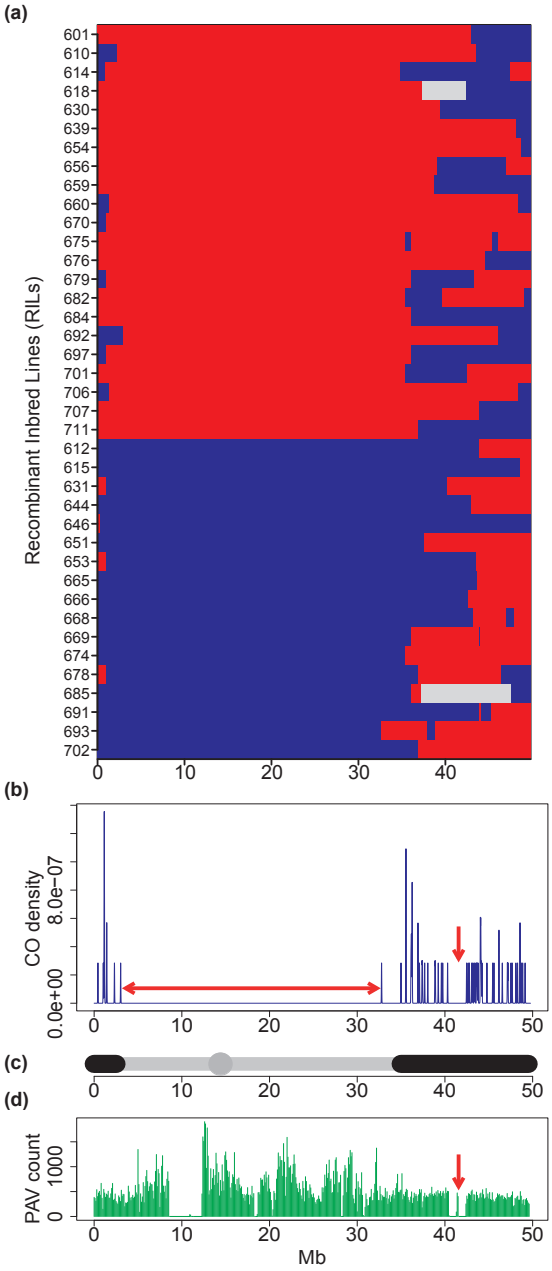

Figure 2.2 Crossover (CO) profile of chromosome 6. **(a)** Schematic overview of parental blocks in the RILs showing at least one CO (*n*=39 RILs). Parental blocks represent the *S. lycopersicum* Moneymaker (blue blocks) and *S. pimpinellifolium* (red blocks) parent. Light grey blocks represent chromosome segments for which the parental source could not be determined. CO regions are inferred at positions where blue blocks change to red blocks or vice versa. **(b)** Kernel density estimate of CO regions in all RILs. Red horizontal and vertical arrows indicate a 29.7 Mb pericentromeric and the position of two euchromatic cold regions respectively. **(c)** Schematic representation of the pachytene chromosome 6, with the centromere depicted by a circle, and the estimated euchromatin and heterochromatin represented by a black and grey line, respectively. **(d)** The distribution of parental alleles (PAVs) over chromosome 6. PAVs are identified with respect to the reference tomato genome (Heinz) and counted for each 100kb region. The region 40.5-42.5 Mb with low allele density, aligning with two euchromatic cold regions shown in B, is indicated by a red vertical arrow.

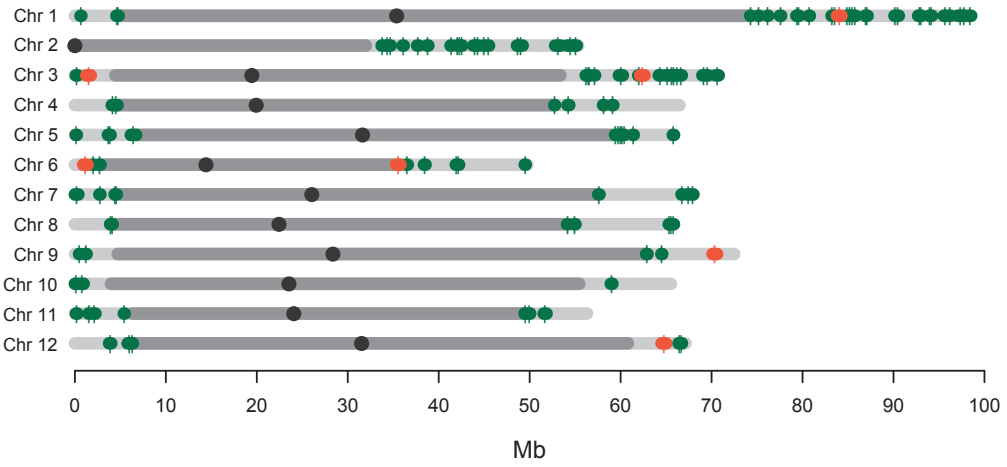

Figure 2.3 Hot and cold regions of crossovers in euchromatin in RILs. The centromere is depicted by a black circle; heterochromatin and euchromatin is shown in dark and light grey respectively; hot and cold regions are shown in red and green respectively, with vertical bars positioned at the midpoint. The left and right positions of the centromere indicate short (S) and long arm (L) respectively. A size bar is indicated at the bottom.

2.3.3 Annotation of CO sites

2.3.3.1 Sequence motifs, protein domains and genomic features

To detect motifs that are enriched in the CO regions, we analysed the sequences of the 453 out of the 1015 total of CO regions harbouring at least one CO with a maximum length of 1 kb, using both discriminative and non-discriminative motif searches. A significant enrichment of 16 motifs was observed. Among these, poly-A, poly-T and poly(AT) stretches were detected at high frequencies (453/453, $E = 3.1 \times 10^{-116}$; 282/453, $E = 3.8 \times 10^{-122}$; 227/453, $E = 2.6 \times 10^{-34}$, respectively) by the non-discriminative MEME algorithm (Figure 2.4a; MEME motifs 1 and 2). In addition, we found a CCN repeat motif at intermediate frequency (143/453, $E = 1.2 \times 10^{-82}$). Using the discriminative DREME algorithm, six AT-rich motifs were found with *P*-values less than 10^{-9} (Figures 2.4a and S3). These were weakly, though significantly, similar to a few different transcription factor (TF)-binding motifs found in the JASPAR motif database (Table S4), in particular C2H2, MADS box and homeodomain motifs.

To assess a possible correlation between recombination and characteristic genome features, the location of CO regions was compared with the annotated tomato reference genome. As shown in Figure 2.4(b), there is a clear preference for CO within 3000 bp upstream of the TSS of genes ($P < 0.05$), with the highest frequency within 1000 bp upstream of the TSS, suggesting that the recombination machinery prefers promoters and intergenic regions in general over gene body domains.

To investigate whether CO regions occur preferentially in the neighbourhood of specific types of genes, we examined genes in these regions for an InterPro domain and Gene Ontology (GO) term enrichment. The GO term enrichment scan revealed a significant GO term, outer membrane-bounded periplasmic space (GO:0030288). However, we failed to see any relationship of this term to meiotic recombination. Furthermore, we found three InterPro domain types (IPR018248, IPR004305 and IPR011598), occurring 20 times in CO regions out of 1344 genes in CO regions that have an InterPro domain. This corresponds to a 50% enrichment compared with the

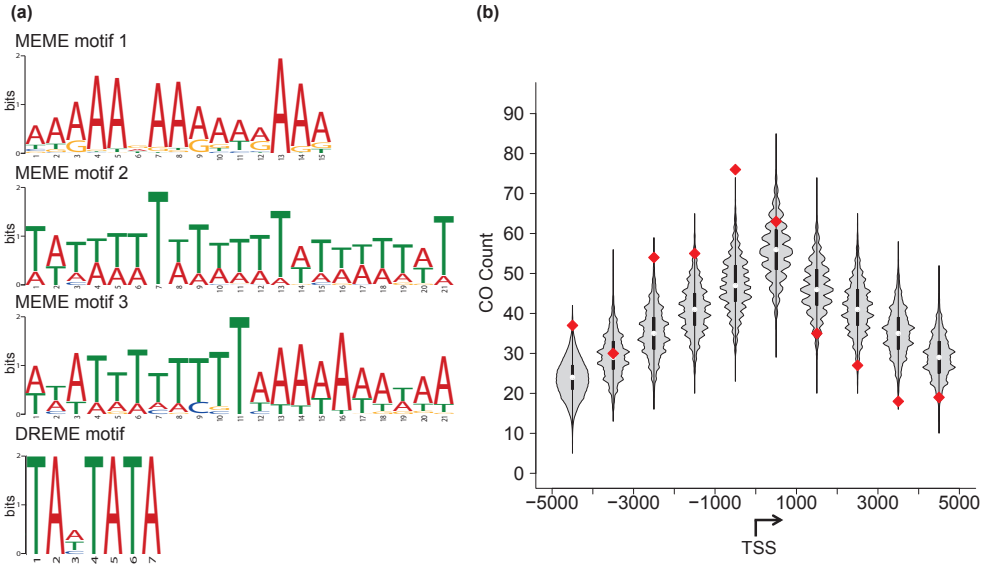

Figure 2.4 Enriched sequence motifs in CO regions and CO distances to transcription start sites (TSS). **(a)** The AARAADAARAWRAAA, WWTWWWTWTWWWTWWWWWTWT, and WWWTWTWTTTTAWAWWWWWAA motifs were discovered with MEME, while the TAHATATA motif ($E=1.5\times10^{-4}$, $p = 1.0\times10^{-11}$) was found over-represented by DREME. **(b)** Distribution of distances between a CO site and its nearest transcription start site. Negative (positive) distances indicate that the CO site is upstream (downstream) of the closest TSS. Red diamonds show the actual CO count between 5 kb upstream and 5 kb downstream the nearest TSS in 1 kb bins. The violin plots indicate the distribution of 10,000 random CO events. White dots indicate the median values of the random samples.

IPR domain containing genes outside CO regions. Interestingly, these IPR domains are associated with DNA damage repair functions and non-transcriptional control of DNA replication. However, although a signal for enrichment was detected, the *P*-value corrected for multiple testing was just below threshold when compared against random recombination sets.

2.4 Discussion

2.4.1 Recombination frequency

Repair of double-stranded breaks (DSBs) usually results in interfering and non-interfering COs, and non-COs. In this study, we mapped regions with CO resulting from interspecific meiotic recombination down to the SNP level in 52 genome sequences of F_6 tomato RILs from *S. lycopersicum* and *S. pimpinellifolium* crosses. We determined both the frequency and the approximate position of CO sites and analysed regions with respect to their annotated positions in the tomato reference genome to identify key genomic features involved in meiotic recombination of homeologous segments. The majority of the initiating events of meiotic recombination in our RILs were resolved into non-COs (Mézard, 2006; Mézard *et al.*, 2007; Lohmiller *et al.*, 2008; Anderson *et al.*, 2014), manifesting either as genetically detectable short gene conversions (25–50 bp) or as silent non-COs (Mézard, 2006; Mézard *et al.*, 2007; Wijnker *et al.*, 2013). The detection of these non-COs, however, was not feasible as our approach depended on genomic regions displaying at least 200 PAVs.

We used strict parameters to limit the number of false positives. As a result, our identification algorithm is likely to have found fewer COs than have actually occurred, although with few false positives. We investigated whether assembly gaps were over-

represented in regions with recombination events, but we did not find a bias when testing against the random background set. We observed approximately 2% of CO events sharing the same upstream or downstream border SNP, respectively. Whether these represent co-occurring COs, or COs in the same region although at different positions, we currently cannot determine.

An important question to address is how representative our results are. The CO frequency obtained for each chromosome indicates that the observed homeologous CO frequency (1.9/5 = 0.38 CO per chromosome per generation) is consistent with the lower homeologous CO rate found in *S. lycopersicoides* (LA2951) × *S. lycopersicum* hybrids (Canady *et al.*, 2006) and far less (80% less, or 20% of intraspecies CO) than the 1.8 homologous CO events per chromosome per tomato generation (Sherman and Stack, 1995). Furthermore, CO in tomato between segments derived from *S. pennellii* and *S. habrochaites* (formerly *Lycopersium hirsutum*) is 15–30% of homologous (or intraspecies) levels (Van Wordragen *et al.*, 1996; Monteforte and Tanksley, 2000). Apparently, such sequence divergence affecting meiotic and mitotic recombination is a more general phenomenon and has also been observed in, for example, maize (Dooner and Martínez-Férez, 1997), yeast (Chen and Jinks-Robertson, 1999) and Arabidopsis (Li *et al.*, 2006). However, while one could expect the frequency of recombination to be positively correlated with the level of homology between parental lines, various studies comparing the genetic map length of intraspecific and interspecific crosses within the tomato clade (Grandillo *et al.*, 2011), which is indicative for the frequency of meiotic recombination, show contrasting results. Genetic map lengths based on intraspecific and interspecific crosses appeared similar to the tomato genetic map length (Tanksley *et al.*, 1992; Paran *et al.*, 1995; Grandillo and Tanksley, 1996; Saliba-Colombani *et al.*, 2000), whereas a map using the same set of genetic markers for the interspecific map *S. lycopersicum* × *S. pennellii* was longer than that of *S. lycopersicum* × *Solanum chmielewskii* (Paterson *et al.*, 1988; Tanksley *et al.*, 1988). The decreased recombination frequency for the latter species appears rather to correlate with more frequent large structural differences (Anderson *et al.*, 2010), notwithstanding a higher sequence divergence for *S. pennellii* (Aflitos *et al.*, 2014).

Strong variation in other species like maize and *Drosophila* has also been observed and seems to be dependent on the sex identity of the gametes. Recombination rates in plants are known to differ between male and female meiosis, but such a genome-wide reduction in CO is not found consistently. De Vicente and Tanksley (1991) used a single F_1 plant for backcrossing to each of the parents, *S. lycopersicum* and *S. pennellii*, and found significantly less recombination for male gametes at all levels. In contrast, Drouaud *et al.* (2007) found the opposite for Arabidopsis chromosome 4, in which male CO rates were higher (up to four times the mean value), whereas female CO rates were equal to or even below the chromosomal average. More recently, Phillips *et al.* (2015) again showed higher overall recombination rates in female meiosis, but also demonstrated a profound effect of temperature on recombination rates in the pericentromeric regions. Thus, while tempting, care should be taken if extrapolating the results for species in the tomato clade and for plants in general, as apparently many factors influence the outcome on recombination frequency and location of COs.

In addition, we found many cold regions larger than 100 kb. Three of them are positioned in the 65.5–66.3 Mb region of chromosome 3L and two of them are positioned inside the 41.8–42.3 Mb region in chromosome 6L which are depleted of PAVs (Table S2). In these two regions, the lack of PAVs points to an introgressed segment in chromosome 6L and a candidate introgressed segment in chromosome 3L. The other cold regions in the euchromatin of several chromosomes suggest that structural rearrangements such as inversions may be present, preventing proper chromosome pair-

ing and blocking recombination, as described above. Indeed, Anderson *et al.* (2010) showed substantial changes in chromosome structures among species of the tomato clade. Among the unusual synaptic configurations for *S. lycopersicum* × *S. pimpinellifolium* F_1 hybrids, mismatched kinetochores and mismatched ends, and foldbacks were found that may restrict meiotic recombination. Besides structural differences, a reduced recombination frequency between homeologous chromosomes could be due to selection against parental allele combinations during RIL propagation, as has been described for tomato × *S. pimpinellifolium* and tomato × *S. pennelli* hybrids (Paran *et al.*, 1995). Interestingly, we observed a lack of COs in euchromatic regions of chromosomes 2L, 10L and 11L for which markers associated with transmission ratio distortion and hybrid incompatibility have been found (Pertuzé *et al.*, 2002; Albrecht and Chetelat, 2009).

The tomato chromosome morphology has been extensively studied, consisting of long stretches of euchromatin in both chromosome arms, flanked by heterochromatin at the telomere ends and the centromere (Ganal *et al.*, 1991; De Jong *et al.*, 2000; Chang *et al.*, 2008). Subsequent studies revealed approximately 220 Mb of euchromatin in the 12 tomato chromosomes (Peterson *et al.*, 1996; Chang *et al.*, 2008) with substantially higher densities of recombination, genes and transcripts (The Tomato Genome Consortium, 2012). In all chromosomes we found COs resulting from recombination mostly at the ends of the chromosomes where euchromatin is present, which is consistent with the homologous recombination distribution found by Sherman and Stack (1995). In yeast, smaller chromosomes undergo more DSBs than larger ones (Pan *et al.*, 2011) and undergo more COs per kilobase (Kaback *et al.*, 1999). It has been speculated that smaller chromosomes may have a longer window of opportunity to make DSBs and that this somehow results in the observed variation in CO density (Lam and Keeney 2015). Therefore we assessed a possible correlation between effective chromosome size and CO by comparing CO frequency and euchromatin size for each of the 12 chromosomes. Our results show that the COs in homeologous segments are not uniformly distributed over the normalized euchromatin length. We find a clear anti-correlation with euchromatin arm size, except for chromosomes 6 and 11 that show similar or higher CO frequency along the long arms, respectively.

2.4.2 Crossover behaviour

Our results show that the most frequently overrepresented sequence motifs at CO sites in homeologous segments are poly-A stretches, poly-T stretches, AT-rich motifs and CCN repeats. Interestingly, the CCN repeat has previously been described to be enriched in CO regions and associated with recombination within Arabidopsis genes but not within promoters, and linked to increased chromatin accessibility (Shilo *et al.*, 2015). Poly-A regions have been reported to be nucleosome-free regions that are targeted by the recombination machinery in yeast (Wu and Lichten, 1994; Pan *et al.*, 2011) and *Arabidopsis* (Wijnker *et al.*, 2013). Also AT-rich motifs, typical of and mainly occurring in non-coding and promoter regions, seem to be CO targets in Arabidopsis (Wijnker *et al.*, 2013). One might assume that the molecular mechanisms underlying recombination between homologous and homeologous sequences are the same and would involve similar sequence motifs. Indeed, the overrepresented sequence motifs involved in CO of homeologous segments from *S. lycopersicum* and *S. pimpinellifolium* substantiate this idea. One intriguing possibility to explain the presence of these motifs could be that transcription is somehow linked to the recombination machinery, promoting recombination. Transcription and recombination involve mechanistic similarities such as DNA unwinding and protein recruitment. Indeed, evidence from lower and higher eukaryotes has accumulated pointing to enhanced recombination at transcriptionally active DNA, a phenomenon known as transcription-associated recombination (TAR) (Gottipati and Helleday, 2009). For example, in

Schizosaccharomyces pombe mitotic and meiotic CO was stimulated at higher transcribed genes and the most highly transcribed allele was the preferred acceptor of genetic information (Grimm *et al.*, 1991). The mechanism, however, remains elusive, and both 'collision' and 'accessibility' theories have been postulated to explain TAR. The observed enhanced recombination in promoter regions might thus be explained by the local structural conformation of the DNA that is accessible for both replication and recombination complex formation and DNA-transcription initiation complex formation, the latter, for example, being needed to transcribe genes promoting recombination such as DNA damage repair proteins. Alternatively, colliding transcription complexes and replication forks at promoter regions stalling the replication could be followed by subsequent recruitment of recombination machinery factors to resolve the collision via recombination. However, this potential relation between recombination and transcription in tomato remains speculative.

It is interesting though that the motifs in tomato and *S. pimpinellifolium* enriched at recombination regions are different from those observed in other organisms, for example maize (Rodgers-Melnick *et al.*, 2015), human or *Drosophila,* pointing at different modes of operation and regulation of recombination. Recombination in tomato and *S. pimpinellifolium* was found to be enriched near TSSs. It has been suggested that the absence of a hotspot-associated protein such as PRDM9, binding to a specific short CCN repeat motif, could provide the recombination machinery with more flexibility to access alternative sites such as promoter regions, indicating that the localization of recombination in tomato and Arabidopsis works differently from, for example, mouse and human. Indeed, PRDM9 knock-outs in mice showed that recombination is initiated at promoters, suggesting that sequence-specific binding of PRDM9 directs the recombination machinery away from gene-promoter regions (Brick *et al.*, 2012). In this respect our observations are in line with the 'windows of opportunity' model proposed for Arabidopsis, dogs and yeast (Lichten, 2008; Auton *et al.*, 2013; Choi *et al.*, 2013; Wijnker *et al.*, 2013; Lam and Keeney, 2015).

The InterPro domains found in genes near recombination sites are enriched for the DNA break repair class. This may be the result of co-evolution of DNA breaks and genes involved in DNA damage repair. Genes responsible for DNA break repair may be connected to promoter regions that are open and accessible for recombination, because these genes need to be transcribed at appropriate levels in case DNA needs to be repaired by hom(e)ologous recombination. In the 'window of opportunity' model mentioned above, this would imply a higher likelihood for these regions to be involved in a recombination event.

In conclusion, we have determined the distribution, position, and genomic characteristics of recombination events for 52 genome sequences of F_6 tomato RILs from a cross between *S. lycopersicum* and *S. pimpinellifolium*. These results provide valuable information for developing a genome-wide predictor for potential recombination sites, aiming at assisting breeders for breeding parent selection and targeted introgression hybridization breeding.

2.5 Experimental procedures

2.5.1 Sequence data

Previously, RILs were constructed from an interspecific cross between *S. lycopersicum* (accession LA2706) and *S. pimpinellifolium* (accession CGN14498), which were used as the male and female parent, respectively (Aflitos *et al.*, 2014). F_1 offspring plants were subsequently selfed and advanced to the F_6 generation (Aflitos *et al.*, 2014). The sequences of 60 F_6 RILs and parental plants were downloaded in BAM file format from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) un-

der project number PRJEB6659. Parental lines were sequenced with an average read depth of 20–45 and the RIL offspring were shallow sequenced with an average read depth of 3–12. Reads from RILs and parental lines were mapped against SL2.40 of *S. lycopersicum* cv. Heinz 1706 LA4345, using bwa (Li and Durbin, 2009). For subsequent analyses, we only considered locations at which a minimum of four reads map uniquely, each at a minimum mapping quality of 60, and for which the base quality of the variant nucleotide was at least 20 in each read. SNPs were called using samtools (Li *et al.*, 2009) and then remapped to the latest SL2.50 assembly using the NCBI Genome Remapping Service (http://www.ncbi.nlm.nih.gov/genome/tools/remap) to retrieve a set of initial SNP locations.

2.5.2 Heterozygosity check

The heterozygosity levels in the RILs were calculated from the variant call format (VCF) data by dividing the number of heterozygous SNPs by the total number of SNPs in each RIL. Taking into account average heterozygosity levels in RILs from interspecific crossing, as observed previously by Paran *et al.* (1995), F_6 RILs showing a genome-wide heterozygosity profile above 15% were discarded.

2.5.3 Haplotyping and identification of CO sites

The composition of a RIL genome can be represented as a set of consecutive haplotypes inherited from either parent. A haplotype change thus coincides with a CO event. To haplotype the RILs, we first determined a set of discriminative allele locations, i.e. genomic positions at which the genotype varies between both parents, *S. lycopersicum* and *S. pimpinellifolium*. Per RIL, discriminative allele locations were then assigned to either parent, yielding a set of parent-assigned variants (PAVs). Haplotype blocks consisting of consecutive variants originating either from the *S. lycopersicum* or *S. pimpinellifolium* parent were then reconstructed.

To detect COs in RILs we searched for regions in the genome containing at least 200 PAVs. We applied a sliding window approach, allowing 95% overlap between two adjacent search windows. A putative CO region was called only when at least 80 of the 100 PAVs on the left of the site were assigned to *S. lycopersicum* and 80 of the 100 PAVs on the right to *S. pimpinellifolium* (or vice versa). We set these parameters to minimize the risk of false CO detection, keeping in mind the 1.8 CO occurrences found on average per chromosome per generation (Sherman and Stack, 1995). Regions overlapping with sequence gaps, as identified by stretches of N base calls in the *S. lycopersicum* reference genome sequence, were discarded to eliminate possible false positives that could arise from assembly errors in the tomato reference genome. To further exclude false positives, such as heterozygous SNPs from local read mapping or sequencing errors, putative CO regions exceeding a local heterozygosity level above 10% were also discarded. This local heterozygosity level threshold was defined according to the observed distribution of SNPs in regions containing at least 200 PAVs, which displays a sharp decrease after 10%. Furthermore, to test whether copy number variation (CNV) in a putative CO region causes false positives, both parents were assessed for possible CNV in the regions containing multiple COs. This was done by the coverage of the raw sequence data from the corresponding CO region to the average genome coverage. The borders for each filtered CO region were then subsequently set according to the corresponding reference genome positions of the two PAVs neighbouring the CO site.

2.5.4 Euchromatic region determination

Tomato chromosome morphology has been well studied by cytogenetic analysis of pachytene chromosomes, which display long continuous stretches of less condensed

euchromatin in chromosome arms flanked by highly condensed heterochromatin at the telomere ends and centromeres (Ganal *et al.*, 1991; De Jong *et al.*, 2000; Chang *et al.*, 2008). The euchromatin–heterochromatin borders for each arm were calculated according to Stack *et al.* (2009). Briefly, the average length and heterochromatin length (in micrometres) of each pachytene chromosome arm based on 100 complete synaptonemal complex (SC) sets was calculated based on table 1 of Sherman and Stack (1992). Then, the euchromatin length of each chromosome arm was calculated simply by subtracting the heterochromatin length from the corresponding arm length. Subsequently, the euchromatin lengths were multiplied with the euchromatin DNA density (1.54 Mb μm^{-1}) to obtain the euchromatin length of each chromosome arm in megabases. The euchromatin–heterochromatin borders were then set based on the length of each euchromatin region (see Data S1).

2.5.5 Sampling of random CO regions

To generate a background CO distribution for testing over- and underrepresented regions, we sampled *n* genomic positions per euchromatic region in each of the chromosome arms per RIL (chromosome 2 only has a long arm, thus comprising 23 arms in total), where *n* is the number of CO events in a particular RIL observed for each euchromatic region. The heterochromatin is not taken into account as recombination is considered to be rare in these regions (Sherman and Stack, 1995), which was confirmed by our observations. We then determined the nearest bordering PAVs for each of the random positions and subsequently filtered these random positions using the same criteria as for observed CO regions. In this way, the distributions of the locations and the interval length of random CO regions faithfully represent those observed in RILs. The whole process was repeated to generate 10 000 random CO sets.

2.5.6 Sequence motif discovery

Sequence motif discovery was done with the MEME suite (Bailey *et al.*, 2009). To construct the set of sequences used for motif search, CO regions larger than 1 kb were not taken into account considering the high uncertainty of the location of the actual CO and to minimize excessive noise arising from using large sequence intervals. CO regions smaller than 1 kb ($n = 490$) were extended to 1 kb (equally on both sides from the midpoint) and used as input for MEME and DREME. MEME found overrepresented motifs occurring zero times or once per region (zoops mode), using the reverse complement activated mode to include both DNA strands which may be accessible to the recombination machinery, with a minimum and maximum motif length of 6 and 50, respectively, and the dinucleotide frequencies of the tomato reference genome as background. For DREME motif discovery, we used an *E*-value cutoff and maximum motif length of 0.1 and 50, respectively, and the random recombination sets as a background. Subsequently, motifs found by DREME were compared with the JASPAR CORE 2016 plant database of TF-binding sites using the Tomtom algorithm with default settings (Gupta *et al.*, 2007).

2.5.7 Gene feature distribution

To assess whether COs between homeologous chromosomes or segments correlated with specific genomic features (e.g. untranslated regions, exons, introns), we determined the nearest gene features to CO regions using the ITAG2.3 annotation for the tomato reference genome (https://solgenomics.net/). The distance between the midpoint of the region and the nearest coordinate of the closest annotated feature was calculated with BEDtools' closest –D option. The same procedure was applied to each random set, resulting in 10 000 random distributions of nearby gene features. To be consistent with these random samples, we only compared CO regions present in euchromatic regions to random CO region sets.

2.5.8 Protein domain and GO enrichment analysis

Genes (partially) within a distance of 7.5 kb from the midpoint of CO regions were analysed for InterPro protein domain enrichment with respect to the 10 000 random recombination site sets. The distance threshold was defined considering an average gene size of 3.3 kb (The Tomato Genome Consortium, 2012). The domain information for unique genes (n = 1843) overlapping with the CO regions was obtained from the ITAG 2.3 functional protein annotation. Significantly overrepresented InterPro domains were selected using a *P*-value adjusted for multiple testing (Bonferroni) with a cutoff of 0.05.

2.5.9 Identification of hot CO and euchromatic cold regions (CO-dense and CO-depleted regions)

To identify hot CO regions, we estimated CO density for each chromosome using the R function 'density' to produce a kernel density estimate. Each CO event is represented by a single Gaussian kernel, with the bandwidth set to 20 000 nucleotides (nt) after manual inspection of the resulting densities. The density was estimated every 1000 nt and an empirical *P*-value was estimated at each local maximum using kernel density estimates in the random recombination region sets, which were calculated in the same way as for the observed recombination set. Significant hot regions were then called by adjusting these *P*-values for multiple testing (Bonferroni) and using a cutoff of 0.05. Cold regions were determined using the same approach, but at every 1000-nt position, employing a *P*-value cutoff of 0.05 without Bonferroni correction to include the regions negatively affected by the large bandwidth (20 000 nt) on the edges (or radius) of kernels. Consecutive cold regions, 10 kb apart at most, were merged. Only cold regions exceeding 100 kb are reported.

2.6 Acknowledgement

The work presented here is supported by the EU FP7 COMREC Marie Curie Initial Training Networks Programme project number 606956.

2.7 Supporting Information

Supplementary files are available at https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13406

2.8 References

Aflitos, S. A., Sanchez-Perez, G., De Ridder, D., Fransz, P., Schranz, M.E., De Jong, H. and Peters, S.A. (2015) Introgression browser: high-throughput whole-genome SNP visualization. *Plant J.* **82,** 174–182.

Aflitos, S., Schijlen, E., de Jong, H. *et al.* (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* **80,** 136–148.

Albrecht, E. and Chetelat, R.T. (2009) Comparative genetic linkage map of *Solanum* sect. *Juglandifolia*: evidence of chromosomal rearrangements and overall synteny with tomatoes and related nightshades. *Theor. Appl. Genet.* **118,** 831–847.

Anderson, L.K., Covey, P.A., Larsen, L.R., Bedinger, P. and Stack, S. (2010). Structural differences in chromosomes distinguish species in the tomato clade. *Cytogenet. Genome Res.* **129,** 24-34.

Anderson, L.K., Lohmiller, L.D. *et al.* (2014) Combined fluorescent and electrom microscopic imaging unveils the specific properties of two classes of meiotic crossovers. *Proc. Natl. Acad. Sci. U.S A.* **111,** 13415-13420.

Auton, A., Rui Li, Y. *et al.* (2013). Genetic Recombination Is Targeted towards Gene Promoter Regions in Dogs. *PLoS Genetics* **9,** e1003984.

Bailey, T.L., Boden, M. *et al.* (2009) Meme suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37,** W202.

Bauer, E., Falque, M. *et al.* (2013) Intraspecific variation of recombination rate in maize. *Genome Biol.* **14**, R103.

Bedinger, P.A., Chetelat, R.T., McClure, B. *et al.* (2010) Interspecific reproductive barriers in the tomato clade: opportunities to decipher mechanisms of reproductive isolation. *Sex. Plant Reprod.* **24,** 171–187.

Brown, J. and Sundaresan, V. (1991) A recombination hotspot in the maize A1 intragenic region. *Theor. Appl. Genet.* **81,** 185–188.

Brick, K., Smagulova, F., Khil, P., Daniel Camerini-Otero R. and Petukhova G.V. (2012) Genetic recombination is directed away from functional genomic elements in mice. *Nature* **485,** 642–645.

Canady, M.A., Ji, Y. and Chetelat, R.T. (2006) Homeologous recombination in *Solanum lycopersicoides* introgression lines of cultivated tomato. *Genetics* **174,** 1775-1788.

Chang, S.-B., Yang, T.-J. *et al.* (2008) FISH mapping and molecular organization of the major repetitive sequences of tomato. *Chromosome Res.* **16,** 919–933.

Chen, W.L. and Jinks-Robertson, S. (1999) The role of the mismatch repair machinery in regulating mitotic and meiotic recombination between diverged sequences in yeast. *Genetics*, **151,** 1299–1313.

Choi, K., Zhao, X. *et al.* (2013) Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet.* **45,** 1327–1336.

De Jong, J.H., Zhong, X.-B. *et al.* (2000) High resolution FISH reveals the molecular and chromosomal organisation of repetitive sequences of individual tomato chromosomes. In Chromosomes Today 13 (Olmo, E. and Rdi, C.A., eds). Switzerland: Birkhäuser Verlag, pp. 267–275.

De Vicente, M.C. and Tanksley, S.D. (1991) Genome wide reduction of backcross progeny derived from male versus female gametes in an interspecific cross of tomato. *Theor. Appl. Genet.* **83,** 173-178.

Dooner, H.K. and Martínez-Férez, I.M. (1997) Recombination occurs uniformly within the bronze gene, a meiotic recombination hotspot in the maize genome. *Plant Cell* **9,** 1633–1646.

Drouaud, J., Mercier, R. *et al.* (2007) Sex-Specific Crossover Distributions and Variations in Interference Level along *Arabidopsis thaliana* chromosome 4. *PLoS Genet.* **3,** e106.

Drouaud, J., Khademian, H. *et al.* (2013) Contrasted patterns of crossover and non-crossover at *Arabidopsis thaliana* meiotic recombination hotspots. *PLoS Genet.* **9,** e1003922.

Fu, H., Park, W., Yan, X., Zheng, Z., Shen, B. and Dooner, H.K. (2001) The highly recombinogenic bz locus lies in an unusually gene-rich region of the maize genome. *Proc. Natl. Acad. Sci. U. S. A.* **98,** 8903–8908.

Ganal, M.W., Lapitan, N.L.V. and Tanksley, S.D. (1991) Macrostructure of the tomato telomeres. *Plant Cell*, **3,** 87–94.

Grimm, C., Schaer, P., Munz, P. and Kohli, J. (1991) The strong ADH1 promoter stimulates mitotic and meiotic recombination at the ADE6 gene of *Schizosaccharomyces pombe*. *Mol. Cell. Biol.* **11,** 289-298.

Gottipati, P. and Helleday, T. (2009) Transcription-associated recombination in eukaryotes: link between transcription, replication and recombination. *Mutagenesis* **24,** 203-2010.

Grandillo, S., Chetelat, R., Knapp, S. *et al.* (2011) *Solanum* sect. Lycopersicum. In Wild Crop Relatives: Genomic and Breeding Resources (Kole, C., ed.). Berlin/Heidelberg: Springer, pp. 129–215.

Grandillo, S. and Tanksley, S.D. (1996) QTL analysis of horticultural traits differentiating the cultivated tomato from the closely related species *Lycopersicum pimpinellifolium*. *Theor. Appl. Genet.* **92,** 935-951.

Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. and Noble, W. S. (2007) Quantifying similarity between motifs. *Genome Biology* **8,** R24.

Heil, C.S.S., Ellison, C., Dubin, M. and Noor, M.A.F. (2015) Recombining without hotspots: A comprehensive evolutionary portrait of recombination in two closely related species of Drosophila. *Gen. Biol. Evol.* **10,** 2829-2842.

Kaback, D.B., Barber, D., Mahon, J., Lamb, J. and You, J. (1999) Chromosome size-dependent control of meiotic reciprocal recombination in Saccharomyces cerevisiae: The role of crossover interference. *Genetics* **152,** 1475-1486.

Kauppi, L., Jeffreys, A.J. and Keeney, S. (2004) Where the crossovers are: Recombination distributions in mammals. *Nature Rev. Gen.* **5,** 413-424.

Lam, I. and Keeney, S. (2015) Nonparadoxical evolutionary stability of the recombination initiation landscape in yeast. *Science* **350,** 932-937.

Li, L., Jean, M. and Belzile, F. (2006) The impact of sequence divergence and DNA mismatch repair on homeologous recombination in *Arabidopsis*. *Plant J.* **45,** 908-916.

Li, H. and Durbin. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25,** 1754-1760.

Li, H., Handsaker, B. *et al.* (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079.

Lichten, M. (2008). Meiotic Chromatin: The Substrate for Recombination Initiation. In *Recombination and Meiosis* (pp. 165–193). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lichten, M. and Goldman, A.S.H. (1995). Meiotic recombination hotspots. *Annu. Rev. Genetics* **29,** 423–444.

Lohmiller, L.D., De Muyt, A. *et al.* (2008) Cytological analysis of MRE11 protein during early meiotic prophase I in *Arabidopsis* and tomato. *Chromosoma* **117,** 277-288.

McVean, G. A.T., Myers, S.R. *et al.* (2004) The fine-scale structure of recombination rate variation in the human genome. *Science* **304,** 581–584.

Mézard, C. (2006). Meiotic recombination hotspots in plants. *Biochemical Society Transactions* **34,** 531-534.

Mézard, C., Vignard, J., Drouaud, J., Mercier, R. (2007) The road to crossovers: plants have their say. *Trends Gen.* **23,** 91-98.

Monteforte, A.J. and Tanksley, S.D. (2000) Fine mapping of quantitative trait locus (QTL) from *Lycopersicum hirsutum* chromosome 1 affecting fruit characteristics and agronomic traits: breaking linkage among QTLs affecting different traits and dissection of heterosis for yield. *Theor. Appl. Genet.* **100,** 471-479.

Myers, S., Bowden, R. *et al.* (2010) Drive against hotspot motifs in primates implicates the PRDM9 gene in meiotic recombination. *Science* **327,** 876–879.

Pan, J., Sasaki, M. *et al.* (2011). A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell* **144,** 719-733.

Paran, I., Goldman, I., Tanksley, S. D. and Zamir, D. (1995) Recombinant inbred lines for genetic mapping in tomato. *Ther. Appl. Genet.* **90,** 542-548.

Parvanov, E. D., Petkov, P. M. and Paigen, K. (2010) Prdm9 controls activation of mammalian recombination hotspots. *Science* **327,** 835.

Paterson, A.H., Lander, E.S. *et al.* (1988) Resolution of quantitative traits into Mendelian factors by using a complete linkage map of restriction fragment length polymorphisms. *Nature*, **335,** 721–726.

Pertuzé , R.A., Ji, Y. and Chetelat, R.T. (2002) Comparative linkage map of the *Solanum lycopersicoides* and *S. sitiens* genomes and their differentiation from tomato. *Genome* **45,** 1003–1012.

Peters, S.A., Datema, E. *et al.* (2009) *Solanum lycopersicum* cv. Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *Plant J.* **58,** 857-869.

Peterson, D.G., Price, H.J., Johnston, J.S. and Stack, S.M. (1996) DNA content of heterochromatin and euchromatin in tomato (Lycopersicon esculentum) pachytene chromosomes. *Genome* **39,** 77-82.

Phillips, D., Jenkins, G. *et al.* (2015) The effect of temperature on the male and female recombination landscape of barley. *New Phyt.* **208,** 421–429

Qi, J., Chen, Y., Copenhaver, G.P. and Ma, H. (2013) Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 10007-10012.

Rodgers-Melnick, E., Bradbury, P.J. *et al.* (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U.S.A.* **112,** 3823–3828.

Saliba-Colombani, A., Causse, M., Gervais and L., Philouze, J. (2000) Efficiency of RFLP, RAPD, and AFLP markers for the construction of an intraspecific map of the tomato genome. *Genome* **43,** 29-40.

Sherman, J.D. and Stack, S.M. (1992) Two-dimensional spreads of synaptonemal complexes from *solanaceous* plants. I. The technique. *Genome* **35,** 354–359.

Sherman, J.D. and Stack, S.M. (1995) Two-Dimensional Spreads of Synaptonemal Complexes from Solanaceous Plants. *Genetics* **141,** 683-708.

Stack, S.M. and Anderson, L.K. (1986) Two-dimensional spreads of synaptonemal complexes from Solanaceous plants. II. Synapsis in *Lycopersicon esculentum* (tomato). *Am. J. Bot.* **73,** 264-281.

Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. and Levy, A.A. (2015) DNA crossover motifs delineate open chromatin regions in Arabidopsis. Plant Cell **27,** 2427-2436.

Tanksley, S.D., Ganal, M.W. *et al.* (1992) High density molecular linkage maps of the tomato and potato genomes. *Genetics* **132,** 1141–1160.

Tanksley, S.D., Miller, J., Paterson, A. and Bernatzky, R. (1988) Molecular mapping of plant chromosomes. *In* Proceedings of the 18th Stadler Genetics Symposium, New York. Plenum Press. pp. 157–173.

The Tomato Genome Consortium. (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **585,** 635-641.

Van Wordragen, M.F., Weide, R.L., Coppoolse, E., Koornneef, M. and Zabel, P. (1996) Tomato chromosome 6: a high resolution map of the long arm and construction of a composite integrated marker-order map. *Theor. Appl. Genet.* **92,** 1065-1072.

Wijnker, E., James, G.V. *et al.* (2013) The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife* **2,** e01426.

Wu, T-Z. and Lichten, M. (1994) Meiosis-induced double strand break sites determined by yeast chromatin structure. *Science* **263,** 515-518.

Chapter 3

DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom

Published as S. Demirci, S.A. Peters, D. de Ridder and A.D.J. van Dijk, *The Plant Journal* 95: 686-699, 2018. https://doi.org/10.1111/tpj.13979

3.1 Summary

A better understanding of genomic features influencing the location of meiotic crossovers (COs) in plant species is both of fundamental importance and of practical relevance for plant breeding. Using CO positions with sufficiently high resolution from four plant species [*Arabidopsis thaliana*, *Solanum lycopersicum* (tomato), *Zea mays* (maize) and *Oryza sativa* (rice)] we have trained machine-learning models to predict the susceptibility to CO formation. Our results show that CO occurrence within various plant genomes can be predicted by DNA sequence and shape features. Several features related to genome content and to genomic accessibility were consistently either positively or negatively related to COs in all four species. Other features were found as predictive only in specific species. Gene annotation-related features were especially predictive for maize, whereas in tomato and Arabidopsis propeller twist and helical twist (DNA shape features) and AT/TA dinucleotides were found to be the most important. In rice, high roll (another DNA shape feature) and low CA dinucleotide frequency in particular were found to be associated with CO occurrence. The accuracy of our models was sufficient for Arabidopsis and rice (area under receiver operating characteristic curve, AUROC > 0.5), and was high for tomato and maize (AUROC ≫ 0.5), demonstrating that DNA sequence and shape are predictive for meiotic COs throughout the plant kingdom.

3.2 Introduction

Meiosis is essential in most reproducing organisms in order to halve the number of chromosomes, which enables the restoration of ploidy levels during fertilization (Villeneuve and Hillers, 2001). At the first meiotic division, homologous chromosomes (homologs) are segregated. In most eukaryotes, accurate homolog segregation is ensured by the formation of at least one recombination event or crossover (CO) between the chromatids of homologs. COs represent a reciprocal exchange of genetic information between homologs (Mercier *et al.*, 2015). In this way, meiotic CO increases genetic diversity in a population of sexually reproducing eukaryotes. Understanding the genomic features influencing the location of COs is of fundamental importance for many areas of biology, ranging from chromosome evolution to population genetics. Knowledge of the location of COs is also key to plant breeding, as breeders are interested in manipulating COs, either to introduce favorable genes from wild relatives to crops or to silence COs in order to generate stable genetic lines of successful crops (Wijnker and de Jong, 2008). There are still numerous gaps of knowledge with respect to meiotic CO and its genetic determinants in plants, however.

The mechanism leading to meiotic COs starts with the formation of double-strand breaks (DSBs) at various chromosomal locations. The DSB distribution deviates from uniform in many species, including mammals, birds and plants (Lichten and Goldman, 1995; Kauppi *et al.*, 2004; Edlinger and Schlögelhofer, 2011; He *et al.*, 2017; Choi *et al.*, 2018). If DSBs are not repaired immediately by DNA repair mechanisms, specific proteins (for example Rad51/Dmc1 in *Arabidopsis thaliana*; Edlinger and Schlögelhofer, 2011) guide one of the loose ends of the DSB to its homologous non-sister chromatid to form a double Holliday junction. Depending on how the junction is resolved, the resulting chromatids can have a non-CO (for example, a gene conversion) or a CO. In Arabidopsis, ~4% of the initial DSBs result in COs (Mercier *et al.*, 2015). COs are formed through two pathways, ZMM-dependent interfering (class-I) and ZMM-independent non-interfering (class-II) pathways. Class-I COs are inhibited from occurring near other class-I COs, whereas class-II COs are unconstrained by the presence of adjacent class-II COs; between class I and class II, weak interference has been reported (Anderson *et al.*, 2014; Mercier *et al.*, 2015). In the current study we focus on the location of any resulting COs without discriminating between class-I or class-II COs.

It is an intriguing question as to how conserved or variable the mechanisms underlying CO formation might be in various plant species. For example, variation exists in the mechanisms underlying DSB formation in different plant species (Lambing *et al.*, 2017). Also, some proteins involved in CO formation have opposing roles in various species. One example is that downregulation of ZYP1/ZEP1 leads to fewer COs in Arabidopsis, yet leads to more COs in *Oryza sativa* (rice; Lambing *et al.*, 2017). A general picture of the conservation of the determinants of CO formation in various plants is lacking, however.

The location of COs is known to be correlated with several genomic features. In many plant species like *Solanum lycopersicum* (tomato), *Zea mays* (maize), Arabidopsis and rice, COs are observed in euchromatic regions, where genes are accumulated and are depleted in pericentromeric regions (Wu *et al.*, 2003; Sato *et al.*, 2012; Choi *et al.*, 2013; Gao *et al.*, 2013; Wijnker *et al.*, 2013; Rodgers-Melnick *et al.*, 2015). More specifically, COs occur preferentially upstream of transcription start sites (TSSs), i.e. in gene promoters in tomato and Arabidopsis (Choi *et al.*, 2013; Wijnker *et al.*, 2013; Demirci *et al.*, 2017; de Haas *et al.*, 2017). In addition to their preferential occurrence in promoters, CO regions are also rich in particular sequence motifs, including poly-A sequence motifs in Arabidopsis and tomato, for example (Choi *et al.*, 2013; Wijnker *et al.*, 2013; Demirci *et al.*, 2017). In maize, GC sequences are over-represented in recombination regions (Rodgers-Melnick *et al.*, 2015). Moreover, *Mu* retrotransposon insertion site frequencies are correlated with recombination in maize (Liu *et al.*, 2009). Finally, DNA methylation was recently shown to be involved in CO silencing in Arabidopsis (Yelina *et al.*, 2015). In this study, we will focus on genomic features rather than epigenetic factors.

To learn about genomic features correlated with CO formation in different plants, we take a predictive machine-learning approach. There have been some previous attempts to predict recombination rates and CO positions. In particular, Rodgers-Melnick *et al.* (2015) used several genomic and epigenetic features to construct a model to predict CO density in maize at the megabase scale. Machine-learning models were successfully used to predict meiotic recombination in yeast based on sequences only (Liu *et al.*, 2012). A consistent, simultaneous analysis of multiple plant species in order to compare the genomic determinants of COs is lacking, however. In this study, we apply machine learning to CO data sets from four different plant species in order to: (i) develop predictive models for the occurrence of COs; and (ii) learn about relevant and important features in these species. This allows us to gain insight into the determinants of CO formation throughout the plant kingdom.

3.3 Results & Discussion

Comparison of the genomic features correlated with the formation of COs requires a consistent analysis of multiple plant species. To this end, we pursued a machine-learning approach, training computational models using available CO data sets obtained in populations derived from crossing parental lines. We specifically focused on high-resolution CO regions (less than 2-kb long). The COs were identified from either recombinant inbred lines, tetrads or double haploid lines. Such data were available for tomato (from a cross between *S. lycopersicum* and *Solanum pimpinellifolium*), *Arabidopsis thaliana* (from a cross between Cvi × L*er* and Col accessions, and between Col and L*er*), maize (from a cross between SK and Zheng58 accessions) and rice (from a cross between PA64s and 93-11). Some general characteristics of the genomes of these four species are presented in Table S1, and a more extensive description of the CO data sets is given in the Experimental procedures. Plots comparing transposable element density, gene density, single-nucleotide polymorphism (SNP) density and CO distribution for the four species are provided in Figure 3.1.

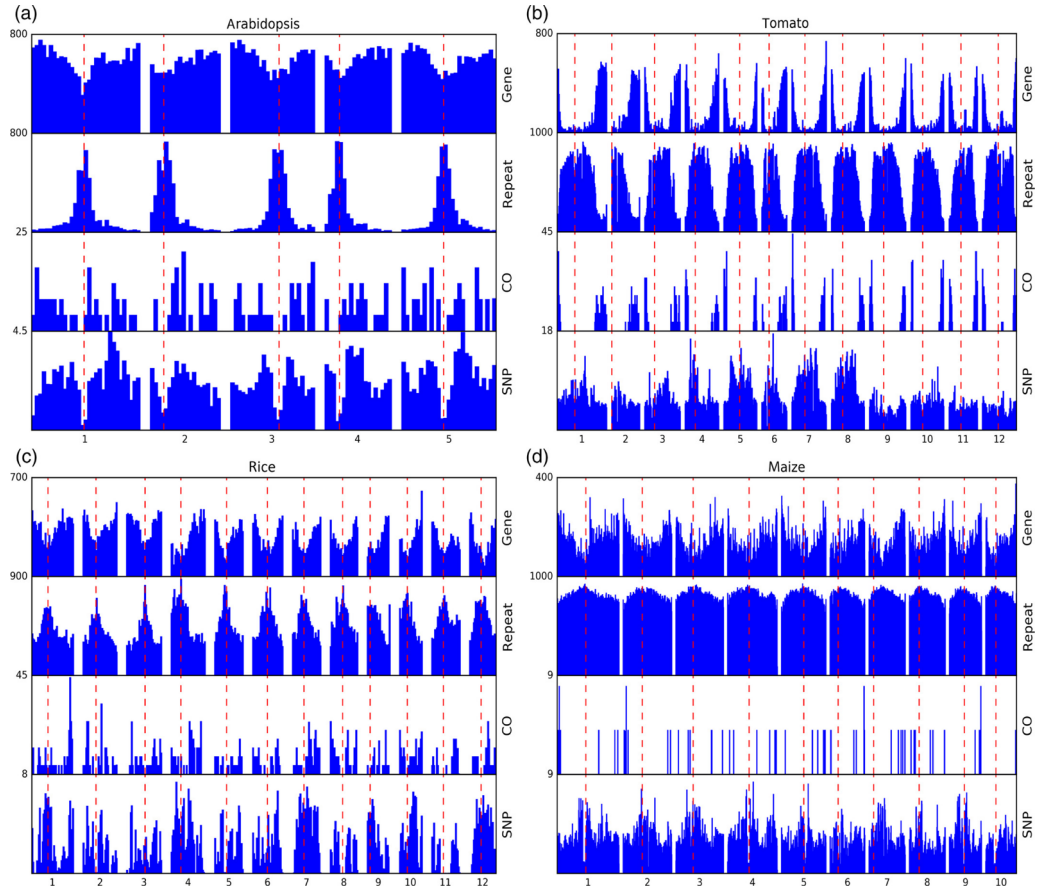

Figure 3.1 Distribution of genomic elements [genes, repeats, crossovers (COs) and single-nucleotide polymorphisms (SNPs)] for **(a)** *Arabidopsis thaliana*, **(b)** *Solanum lycopersicum* (tomato), **(c)** *Oryza sativa* (rice) and **(d)** *Zea mays* (maize). The number of nucleotides covered per kb by the different genomic elements is given in 1-Mb bins for the different chromosomes (indicated by the numbers on the horizontal axes). Dashed lines indicate centromere locations.

We first developed our predictive model on CO data obtained in tomato. Subsequently, we trained similar models for Arabidopsis, maize and rice. We used the CO regions together with their flanking sequences, extending each region to a total length of 4 kb. In these regions, we analysed features based on sequence information, genome annotation and parental genome sequences. We used these features to construct classification (i.e. machine-learning) models that predict the probability of meiotic recombination for a given sequence. After training such a model with a set of known CO regions, the model can be applied to predict likely CO sites throughout the genome. More importantly, we can analyze how the model learned to perform these predictions: i.e. to what extent, and in what direction, is the probability of CO occurrence influenced by the different features, according to the model? In other words, this allows us to learn about genomic features related to CO frequency in different plants.

3.3.1 CO region prediction in the tomato genome

As input for training a machine-learning model, both a positive set (regions containing COs) and a negative set (regions not containing COs) are needed. We prepared a positive set consisting of 4-kb-long CO regions from tomato (n = 664) obtained in our previous study (Demirci *et al.*, 2017). Because the absence of a CO in a given region does not automatically imply that a CO could not occur, generating a negative set is not straightforward. Therefore, we used a random set instead of a negative set. As a first strategy to generate a random set, we simply sampled the same number (n = 664) of 4-kb-long regions randomly from the tomato genome, excluding the 664 CO regions.

Each positive and each random sample was represented by 62 features based on sequence, genome annotation and parental genome sequence variation. Sequence-based features included dinucleotide frequencies and DNA shape features (minor groove width, propeller twist, helical twist and roll). Propeller twist describes how one base in a base pair is rotated about the long axis of the base pair relative to the other base. The helix twist is the angle between two adjacent base pairs as they twist in a DNA helix structure. The roll is the angle between two consecutive base pairs rolling over each other (Chiu *et al.*, 2016). These DNA shape features were predicted using a model trained on experimental DNA structures (Experimental procedures). The values predicted for each nucleotide were averaged over a (positive or random) region to obtain a single value for each region. DNA shape features have recently been shown to be helpful for predicting the binding of proteins to DNA, for example, where using these features showed improved performance compared with using more simple representations of the DNA sequence (Mathelier *et al.*, 2016). Genome annotation features described repeat elements, gene elements and (eu)chromatin state. The latter was defined as described previously (Demirci *et al.*, 2017), based on results from cytogenetic analyses of pachytene chromosomes, which display long continuous stretches of less condensed euchromatin in chromosome arms flanked by highly condensed heterochromatin at the telomere ends and centromeres. Finally, features based on parental genome information included SNPs and INDELs between the parental genomes. Additional information on the exact definition of these features is provided in the Experimental procedures. To assess the discriminative power of the features, we initially applied a Student's *t*-test to compare the means of each individual feature in the positive and in the random set. This indicated that 43 out of 62 features were significantly discriminative (Table S2).

The Student's *t*-test analyzed whether each single feature on its own displayed different values in the CO set compared with the random set. To investigate the discriminative power of features when they are combined, we constructed a classification model that uses all features together. Different types of classifiers were tested to find the best-performing model. In particular, we trained a decision tree, a random forest and a logistic regression classifier. The performances of the prediction models are visualized using receiver operating characteristic (ROC) curves in Figure 3.2(a). The random forest classifier was the best-performing model, in terms of the area under the ROC curve (AUROC = 0.92). Note that performance is calculated using regions not used for training the model, in order to prevent over-optimistic performance estimates.

We subsequently analyzed the importance of each feature according to the random forest model (Figure 3.2b; Table S3). This revealed that whether a region is in euchromatin or not is the most contributing feature (with a positive association, i.e. a region in euchromatin is more likely to be a CO region); euchromatin is defined here as described by Demirci *et al.* (2017). Additional important features included

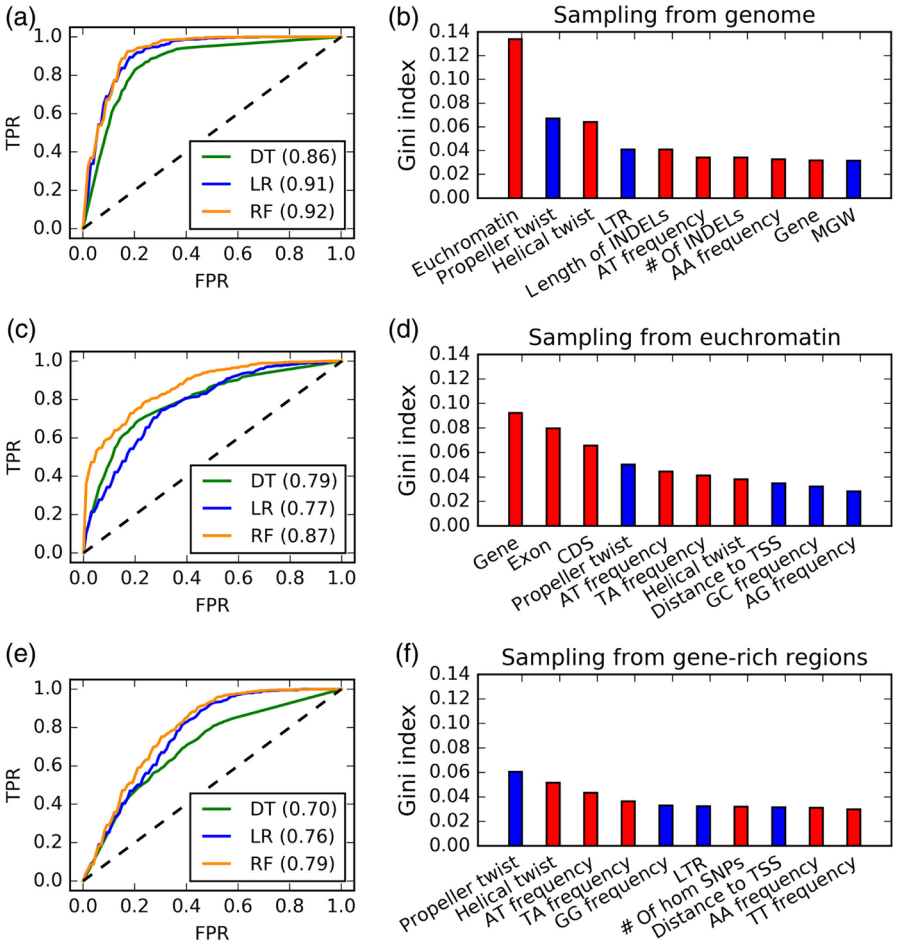

Figure 3.2 Crossover (CO) prediction in *Solanum lycopersicum* (tomato). **(a, c, e)** Assessment of prediction performance with receiver operating characteristic (ROC) curves for the models trained on **(a)** whole genome, **(c)** euchromatin and **(e)** gene-rich regions. Abbreviations: DT, decision tree; FPR, false positive rate; LR, logistic regression; RF, random forest; TPR, true positive rate. Values between brackets indicate area under ROC (AUROC). The dashed lines indicate performance of a random predictor, with AUROC equal to 0.5; the higher the AUROC, the better the predictor. **(b, d, f)** The top-10 most important features (ranked from left to right), according to the random forest classifier using **(b)** whole genome, **(d)** euchromatin and (f) gene-rich regions. The higher the Gini index, the more important the feature, i.e. the bigger its role in determining the CO prediction. The color of the bar for each feature indicates its positive (red) or negative (blue) relationship to the occurrence of CO regions. The DNA shape features given here are mean angular values.

DNA shape features (positive or negative association, depending on the feature), long terminal repeat (LTR) elements (negatively associated, i.e. a region containing LTR repeats is less likely to be a CO region) and the length of insertions and deletions (INDELs) between the two parental genomes (positively associated). The strong contribution of euchromatin presence to CO prediction fits our expectations, as CO regions are known to accumulate in euchromatic regions (Sherman and Stack, 1995; Demirci *et al.*, 2017). The strong contribution of euchromatin in our model may overshadow the effect of other features, however: effectively, the model has learned to discriminate between euchromatin and heterochromatin. In order to find the most relevant features for CO prediction *within* euchromatin, we subsequently followed a second strategy to generate a random set, focusing on the euchromatic regions of the tomato genome.

3.3.2 CO region prediction in euchromatic regions of the tomato genome

To focus on the prediction of CO regions inside tomato euchromatin, we generated an alternative random set: instead of sampling from the whole genome, the regions were sampled randomly from euchromatic regions only. With this new random data set and the same positive data set as above, we again constructed three predictive models using a decision tree, a random forest and logistic regression. Similar to the results obtained with the first random set, the best-performing classifier was the random forest classifier, although the performance slightly decreased (AUROC = 0.86; Figure 3.2c), reflecting an increased difficulty of the prediction problem. As indicated by the AUROC, we could clearly discriminate CO regions from randomly chosen regions in euchromatin. Compared with the results obtained above, the ranking of most contributing features changed drastically (compare Figure 3.2b and d). Top features now are gene density-related features [gene, exon and coding sequence (CDS) coverage], DNA shape, sequence-related features and distance to TSS (Figure 3.2d). This change in the ranking of features, together with the high performance of the model inside euchromatic regions, suggests that not only the (eu)chromatin state but also local sequence properties influence the occurrence of COs. It is particularly revealing that features related to gene density (gene, exon and CDS) constitute the top three (Figure 3.2d). This is in line with existing knowledge on the preference of COs in tomato to be located near genes (Demirci *et al.*, 2017); however, similar to the strong influence of euchromatin found above, we now have features describing high-level annotation that strongly influence the prediction model. In order to further reveal more local sequence properties that influence COs in gene-rich regions, we devised a third and final strategy to generate a random set.

3.3.3 CO region prediction in tomato gene-rich regions

Given the important role of gene annotation-related features in the prediction model found above, we used a third sampling strategy that takes the gene distribution of the tomato genome into account. This new sampling strategy also largely distinguishes euchromatin versus heterochromatin, as euchromatin is more gene rich; moreover, genic regions in heterochromatin where CO potentially could occur are also taken into account. Briefly, this strategy involved the construction of an estimate for the whole-genome gene density, followed by the selection of random regions by sampling from this density. In doing so, the experimental COs were used to find the best value of the bandwidth parameter of the gene density estimation. This procedure ensures that, similar to the positive cases (experimental CO regions), the random cases will preferentially, but not exclusively, occur in gene-rich regions. Further details of this sampling strategy are described in the Experimental procedures.

We constructed three classification models using the same three classifiers with the new random set and the same positive set. Similar to previous trials, the best-performing classifier was the random forest classifier (Figure 3.2e), again with a slightly lower performance than previously (AUROC = 0.79). In this model, the most relevant features are related to DNA shape, sequence, LTRs, distance to TSS and parental sequence differences (Figure 3.2f). In particular, the model revealed local DNA properties as being predictive: the two most important features were the DNA shape features of propeller twist and helical twist. As the second (euchromatin-based) and third (gene density-based) sampling strategies both focus mostly on genic areas, we expect the importance of the features for both strategies to correlate. To test this, we compared the ranking of features obtained by the random forest classifier following the two sampling strategies by Spearman's rank correlation test. The test showed significant positive correlation between the importance scores for the features obtained with these two sampling strategies (Spearman's $\rho = 0.91$, P value $<< 0.001$).

Hence, as expected, out of all features, similar features were selected as important for predicting CO regions in euchromatin (the second sampling strategy) and in gene-rich regions (the final sampling strategy).

We were interested whether this robust behaviour of predictive features was also present between the three different classifiers (decision tree, random forest and logistic regression) trained using the sampling strategy based on gene-rich regions. Such robustness would give credibility to the set of predictive features obtained. To investigate this, we compared the importance of features between the three classifiers by Spearman's rank correlation test; we also included the significance ranking of features obtained from the Student's *t*-test. As summarized in Table 3.1, even the lowest correlation was significant and positive ($\rho = 0.44$; $P < 0.001$). Given that some of the features are related to each other, this correlation between feature importance scores might be an underestimate. It could be strongly influenced by the correlation between features: out of two features that are highly correlated, one may be ranked highly by one classifier and the other ranked highly by another classifier. Note that the correlation between different features describes whether the feature values display similar trends in our data set. Above, we analyzed the correlation between feature importance scores obtained for the same feature with different prediction models. The correlation between feature importance scores could be lowered by correlation between the feature values; to test this, we clustered all features and labeled them with their cluster membership (Figure S1; Table S4). Subsequently, we run the Spearman correlation test for feature importance on cluster ranks (where each cluster was ranked with the rank from its most important feature). As expected, the correlation between the cluster ranks of the features between the different classifiers increased and resulted in a minimum ρ value of 0.56 ($P < 0.001$). The analysis of feature importance thus showed that the ranking of features is robust to the choice of sampling strategy and classifier.

Table 3.1 Spearman correlation coefficients (ρ) between the feature importance values of classifiers in tomato gene-rich regions

	Student's *t*-test	**Decision tree**	**Logistic regression**	**Random forest**
Student's *t*-test	–	0.44***	0.46***	0.70***
Decision tree	*0.57****	–	0.46***	0.75***
Logistic regression	*0.56****	*0.59****	–	0.50***
Random forest	*0.62****	*0.84****	*0.59****	–

Significance: ***$P < 0.001$. Underlined values (above the diagonal) are for the ranking of the importance of individual features; *italic* values (below diagonal) are for the ranking of the importance of feature clusters.

3.3.4 Factors related to crossovers in tomato

As described above, we generated machine-learning models predicting the likelihood of CO formation based on DNA sequence and shape features. In a next step, we aimed to obtain insight into the genomic determinants of CO formation by analyzing how the models make these predictions. This is reflected in the feature importance scores (Figure 3.2); to interpret these, we also made use of the feature values in CO regions and random regions (Figure S2; Table S5). In particular, we observed that the most important features (Figure 3.2) could be grouped into those related to genomic content and those related to genome accessibility.

Two features related to genomic content are euchromatin (Figure 3.2b) and the gene content of a region (Figure 3.2d), which are strongly positively correlated with the

occurrence of CO regions in the first two predictive models. A third feature related to genome content is the presence of LTR repeat regions: according to the final model, the probability of a CO increases with the decreasing occurrence of LTRs (Figure 3.2f). These three genomic features are related to each other, as LTR regions are preferentially positioned in the pericentromeric regions of the chromosomes, where gene density is lower and the DNA is condensed into tightly packed heterochromatin (Sherman and Stack, 1995; Jouffroy *et al.*, 2016).

Among the features important for discriminating CO regions from non-CO regions, there were three features related to the accessibility of genomic regions. First, we found a negative correlation between distance to TSS and the occurrence of CO regions (Figure 3.2d and f). The distribution of TSS distances is shifted towards somewhat more negative values for CO regions, compared with random regions. This implies that, compared with randomly chosen regions, CO regions on average are more often found upstream of the TSS, i.e. in promoter regions. As promoters contain nucleosome-depleted regions (Hartley and Madhani, 2009), and are accessible to transcription factor binding, it is likely that they are also accessible to the recombination machinery during the DSB formation stage, as was found in yeast (Pan *et al.*, 2011) and Arabidopsis (Choi *et al.*, 2018). Moreover, AA, TT, TA and AT dinucleotide frequencies are positively correlated and predictive for CO regions (Figure 3.2f). This finding could be related to the enrichment of TATAT, poly-A and poly-T sequence motifs found in CO regions in tomato (Demirci *et al.*, 2017) and in Arabidopsis (Choi *et al.*, 2013; Wijnker *et al.*, 2013). Similar to the role of promoters, it has been suggested that specific sequence motifs associated with CO occurrence indicate regions of open chromatin (Shilo *et al.*, 2015), which might be explained by the exclusion of nucleosomes, leading to high DSB levels (Choi *et al.*, 2018). Thirdly, we found a relationship between mean propeller twist angle (a DNA structural property) and CO regions (Figure 3.2f): a higher absolute value of propeller twist angle makes a region more likely to be a CO region. Importantly, in yeast a higher absolute propeller twist angle correlates with a lower nucleosome occupancy (Gan *et al.*, 2012). A higher absolute propeller twist angle between particular base pairs could render the DNA more rigid, making the DNA harder to bend around proteins, for example histones (El Hassan and Calladine, 1996). Overall, our results indicate the relevance of genome accessibility for CO formation: nucleosome depletion could render genomic regions more accessible to the recombination machinery.

In addition to features related to genomic content and features related to genome accessibility, the genetic diversity between contributing parental sources is also suggested to be relevant by the model. In particular, the model showed a positive relationship for the number of homozygous SNPs and length of INDELs between parental genomes with CO region presence (Figure 3.2f). Care should be taken when interpreting correlations between SNP rates and CO rates, however: as CO regions are defined by SNPs, it is likely that there is a bias in favor of positive correlation.

3.3.5 CO prediction in Arabidopsis, maize and rice

The results obtained for tomato indicate that it is possible to analyze genomic determinants of CO formation using the set of sequence- and annotation-based features. To investigate the role of these features in other plant species, we constructed prediction models for maize, Arabidopsis and rice. For these three species, we obtained CO regions with sufficient resolution needed for training the models (Wijnker *et al.*, 2013; Li *et al.*, 2015; Si *et al.*, 2015). We prepared positive sets as 4-kb-long regions around CO positions from rice (n = 468), maize (n = 63) and Arabidopsis (n = 159), respectively. We sampled the same number of 4-kb-long regions as in the positive set for each species, using the gene density-based sampling strategy as described above.

We prepared the same features as for tomato, except for the parental sequence-based features. In addition, there are small differences in feature sets between the species as different genomes have different repeat content.

We initially tested the individual discriminative power of features by Student's *t*-test. This yielded 15 significant features among 59 features for Arabidopsis, 13 significant features among 64 features for rice, seven significant features among 55 features for maize and 28 significant features among 56 features for tomato, all with *P* values of <0.05 (Table S6). For tomato, the number of significant features was lower than what was found above when using a random set from the whole genome. This is caused by the fact that it is more difficult to discriminate between CO regions and random regions that are both sampled from gene-rich areas in the genome. Given the smaller number of COs available for Arabidopsis, rice and maize, it is also not surprising that fewer features were found to be significant in these species, compared with tomato. Subsequently, we trained a random forest classifier for each of the three species separately. To compare these three models in a fair way with the tomato model, we also trained a model for tomato without the parental sequence-based features. According to the performance results given in Table 3.2, CO sites are highly predictable for both models of tomato and maize (AUROC $\gg$ 0.5), and are reasonably predictable for Arabidopsis and rice (AUROC > 0.5). The difference in the predictive power is not dependent on the number of COs in our training set: tomato has the most data and maize has the least, whereas in both CO is easier to predict than in Arabidopsis and rice.

Table 3.2 Performance statistics of the random forest model for tomato, rice, maize and Arabidopsis.

	Tomato	Tomato[a]	Arabidopsis	Rice	Maize
AUROC[b]	0.79 (0.04)	0.77 (0.03)	0.63 (0.08)	0.67 (0.05)	0.72 (0.14)
Recall[b]	0.82 (0.03)	0.82 (0.03)	0.64 (0.09)	0.68 (0.08)	0.76 (0.10)
Precision[b]	0.69 (0.04)	0.67 (0.03)	0.58 (0.06)	0.60 (0.05)	0.70 (0.12)
Accuracy[c]	0.95	0.94	0.66	0.76	0.92

[a] Tomato data set without features from the parental genome sequence.

[b] Area under the receiver operating characteristic curve (AUROC), recall and precision are calculated with 10-fold cross-validation using the positive set consisting of experimental crossover (CO) regions and the random set obtained by sampling from gene-rich regions. Values are means obtained with 10-fold cross-validation (with standard deviations given in parentheses).

[c] Accuracy values are calculated on the data set from the pericentromeric regions after training with the positive set and the random set.

To obtain additional validation for the models, we followed two strategies. One was to obtain a set of true-negative cases from pericentromeric regions. Reassuringly, as shown in Table 3.2, the accuracy obtained by applying the models to these regions was again quite decent for Arabidopsis and rice (66–76% correct), and was particularly high for tomato and maize (>90% correct). The second strategy was specific for Arabidopsis, for which we used a genome-wide set of recombination rates (Choi *et al.*, 2013). As expected, the CO regions in our data set showed clearly higher rates compared with random regions (Figure S3a; with a *P* value based on Student`s *t*-test of 10^{-9}). The recombination rate for CO regions correctly predicted by the model was similar to the rate for CO regions not correctly predicted by the model (Figure S3b). Strikingly however, recombination rates for random regions predicted to be CO regions by our model were clearly higher than rates for random regions predicted to be random regions (Figure S3b; with a *P* value based on Student`s *t*-test of 10^{-7}). This provides a clear validation of our model, because it demonstrates that for a set of randomly chosen genome regions the model discriminates between regions with low and regions with high recombination rates.

3.3.6 Factors related to crossovers in Arabidopsis, maize and rice

We further investigated whether similar features are important for CO prediction in the four different species (a complete overview of feature importance values is given in Figure S4). We compared the ranking of the importance of features between species with Spearman's correlation test, as shown in Figure 3.3. On the one hand this revealed that tomato and maize displayed only a modest non-significant correlation, whereas on the other hand all other pairs of species displayed positive significant correlations. The highest correlation was observed between tomato and Arabidopsis, for which very similar features were important to predict CO regions.

To identify common and species-specific features we selected the top ten most contributing features of each species' CO prediction model. Features contributing to the top 10 in at least one species are displayed in Figure 3.4, showing their importance and their influence on the likelihood of COs. Note that the features reported in Figure 3.4 are not necessarily the same as those reported as the result of the Student's *t*-test in Table S6. This is because the Student's *t*-test considers each feature separately, whereas the random forest uses combinations of features, and then ranks the features individually based on their contribution to the model. In addition, for tomato there are small differences between the features shown in Figure 3.2(f) and those shown in Figure 3.4, as the latter includes only features relevant for all four species.

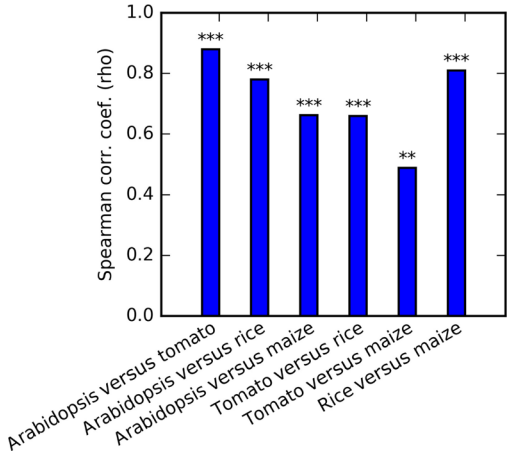

Figure 3.3 Correlation (Spearman's correlation coefficient ρ) between the ranking of the importance of features for crossover prediction in tomato, Arabidopsis, rice and maize. Significance: ***$P < 0.001$; **$P < 0.01$.

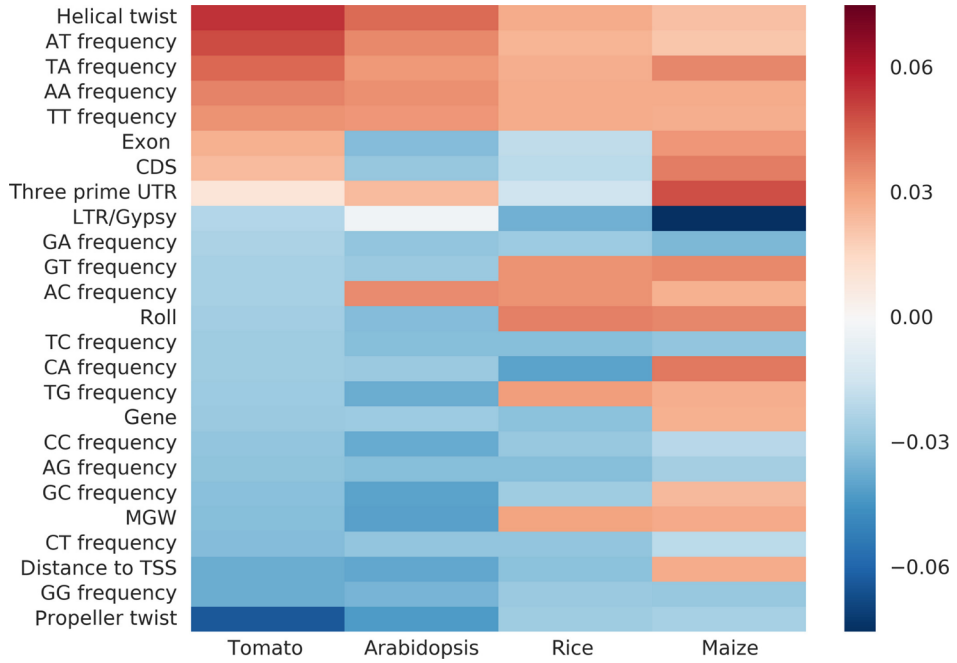

Figure 3.4 Features (vertical axis) contributing to the top-10 most important features in at least one of the species (horizontal axis). Color represents the relationship of features to crossover (CO) prediction (red, positive; blue, negative); intensity represents the importance of the feature. The color-coded full set of features for each species is given in Figure S4.

Interestingly, Figure 3.4 shows that one group of features (consisting of the DNA shape feature helix twist, and AT, TA, AA and TT dinucleotide frequencies) is predicted to have a positive effect in all four studied species, with higher feature values indicating a higher likelihood to be a CO region. Similarly, another group (consisting of the DNA shape feature propeller twist, and GG, GA, TC, CC and AG dinucleotide frequencies) has a negative effect in all four studied species. In addition, the LTR/Gypsy feature has a negative relationship with COs in three of the four species: CO regions are not favored near LTR repeats in maize, tomato and rice. For Arabidopsis the LTR/Gypsy feature is not relevant, as LTR repeats to a large extent are absent from the Arabidopsis genome (The Arabidopsis Genome Initiative, 2000).

These two groups of conserved features, which are consistently positively or negatively related to CO in all four species, can be broadly related to genome content and genome accessibility, as found above for tomato. In particular, the importance of genomic content is reflected in the negative correlation of CO regions with the occurrence of LTR/Gypsy repeats. The negative correlation between recombination and transposon occurrence along chromosome arms has recently been reviewed (Lambing *et al.*, 2017): transposon content increases towards the centromere, whereas the recombination rate decreases towards the centromere. Several other features conserved between species are related to genome accessibility. CO regions are positively correlated with AT, TA, AA and TT dinucleotide frequencies, and with the absolute angles of the propeller twist. As discussed above, the nucleosome occupancy of these regions is expected to be low. This suggests that COs tend to localize in regions of open chromatin that are accessible for the recombination machinery.

In addition to these features that are invariant between species, a more species-specific role was observed for other features. The most important features found in maize were gene annotation-related features like exon, CDS and the 3′ untranslated region (3′-UTR), whereas in tomato and Arabidopsis, propeller twist, helical twist, and AT and TA dinucleotides were the most important features. This difference could partially relate to the observation of CO regions in maize preferentially in 5′-UTRs and 3′-UTRs (Li *et al.*, 2015), and in tomato and Arabidopsis primarily in promoters (Choi *et al.*, 2013; Wijnker *et al.*, 2013; Demirci *et al.*, 2017). Furthermore, in rice, high roll (a DNA shape feature) and low CA dinucleotide frequency in particular favored the occurrence of COs. Two additional features with a species-specific role were minor groove width (MGW) and the distance to TSS. MGW has a negative relationship to COs in Arabidopsis and tomato, and a positive (albeit non-significant) relationship in maize and rice. MGW can strongly influence the binding of proteins to DNA (Rohs *et al.*, 2009). As described in the Introduction, some knowledge exists on the different effects of CO regulators on CO formation in different plant species. The potential influence of MGW on the binding of such CO regulator suggests a possible explanation for why the relationship between MGW and CO formation is positive in some species and negative in others: higher MGWs would have the same effect on binding of the protein in all species, which subsequently would have a differential effect on CO formation. As for distance to TSS, this feature again hints at the importance of genome accessibility. CO regions are localized upstream of the TSS (i.e. in promoter regions) in tomato, rice and Arabidopsis, whereas they are located downstream of TSS (i.e. at the 3′-UTR ends of genes and gene bodies) in maize. Even though CO regions localize at different ends of genes, apparently these positions are associated with nucleosome-depleted regions (Bell *et al.*, 2011), rendering them accessible to the recombination machinery.

3.4 Conclusions

We present a comprehensive application of machine learning to predict CO regions throughout the plant kingdom. CO regions are reasonably predictable in Arabidopsis and rice, and can be predicted with high accuracy in tomato and maize. A few different factors might influence the predictive power. One is that we focus on the prediction of COs in gene-rich regions to be able to find local features, which inevitably means losing predictive power as the difference between random and CO regions gets smaller. The second reason is that there is no proper negative data set to compare against: irrespective of the way we sample, some regions in the random data set may actually be prone to CO formation.

Our results indicate the conservation and the variation of genomic features influencing CO formation throughout the plant kingdom. We found two main groups of conserved features important for predicting CO regions in all four species: genome content and genome accessibility. CO regions are more likely to lie in euchromatic, gene-rich chromosomal regions, be AT-rich, have high absolute propeller twist angles and be depleted of LTR repeats. This could well relate to nucleosome depletion, leading to accessibility by the recombination machinery. In addition to these general rules, we observed that in Arabidopsis, rice and tomato, CO regions are often found in the 5′-UTR ends of genes, whereas in maize the CO regions are more prevalent in the 3′-UTR ends of genes. Yet, in general, in Arabidopsis, rice, tomato and maize, CO regions localize around the UTR ends of genes, which suggests that gene regulatory regions are involved in the CO mechanism.

In addition to these gross similarities between species, our results also indicate the importance of species-specific aspects of CO formation. One example is that MGW is negatively related to CO formation in tomato and Arabidopsis, but is positively relat-

ed in rice and maize. Our findings that both conserved and species-specific genomic features are correlated with COs might be related to the differential effect that proteins have on CO formation. For example, PRDM9 has a specific role in CO formation in human and mouse (Myers *et al.*, 2010; Edlinger and Schlögelhofer, 2011). Similarly, PCH2/CRC1 and ZYP1/ZEP1 seem to have a differential effect on CO formation in Arabidopsis and rice (Lambing *et al.*, 2017). The finding that DNA shape features are important according to our prediction models could be related to the interactions of such proteins with DNA, given that DNA shape is known to be relevant for protein–DNA interactions (Mathelier *et al.*, 2016). The characteristics of the (spatial) interaction between such proteins and their DNA targets is relatively unknown, and in our opinion calls for more detailed studies involving, for example, chromatin immunoprecipitation sequencing (ChIPseq) technology.

Generally speaking, our results indicate the importance of both the conservation and the variation of features influencing COs in various plant species. Our work lays the ground for a comprehensive analysis of features underlying CO formation in plants. Using additional high-resolution data sets, as well as additional relevant features such as epigenetic modifications, will be the next step in order to better understand CO regions. This will be of fundamental biological relevance, and will provide further opportunities for application in plant breeding.

3.5 Experimental procedures

3.5.1 Dataset preparation

Sequences for positive (CO regions) and negative cases were prepared for tomato, rice, thale cress (Arabidopsis thaliana) and maize by using the corresponding genome information.

For tomato, 1015 CO positions were obtained from Demirci *et al.* (2017). CO events were detected in an F6 generation of interspecies recombinant inbred lines (RILs). The parental lines of the RILs were *S. lycopersicum* Moneymaker and *S. pimpinellifolium*. The reference genome *Solanum lycopersicum* Heinz version SL2.50 was used. The genome sequence and gene annotation files (ITAG2.4 gene models and ITAG2.4 repeats aggressive files in gff3 format) were obtained from https://solgenomics.net.

For rice, 1287 CO positions were obtained from Si *et al.* (2015). CO events were detected in F2 lines grown in different environmental conditions; the parental lines were *PA64s* (a hybrid between *O. sativa* indica and javanica) and *93-11* (*O. sativa* indica group). The reference genome *Oryza sativa* Nipponbare version IRGSP-1.0 was used. The genome sequence and gene annotation files were obtained from http://rapdb.dna.affrc.go.jp/download/irgsp1.html.

For Arabidopsis, 191 CO positions in total were obtained from tetrads and double haploids of *Arabidopsis thaliana* (Wijnker *et al.*, 2013). The parental lines of tetrads were Cvi × L*er* and Col accessions of *A. thaliana*. The parental lines of double haploids were Col and L*er* accessions. The reference genome version TAIR 10 genome sequence and gene annotation (gff3) were obtained from https://arabidopsis.org.

For maize, 924 CO positions from tetrads were obtained from Li *et al.* (2015). The parental lines of tetrads were SK and Zheng58 accessions of *Z. mays*. The reference genome B73 RefGen v3 (aka AGPv3) genome sequences and the gene annotation file were downloaded from Ensembl Genomes release 21 (ftp://ftp.ensemblgenomes.org/pub/plants/release-21/fasta/zea_mays/).

Repeats for rice, Arabidopsis and maize genomes were inferred using repeatmasker (Smit *et al.*, 2013), together with its dependencies via the tandem repeat finder

(Benson, 1999) and the National Center for Biotechnology Information (NCBI) blastn programs. The Genetic Information Research Institute Repbase Update database (Bao *et al.*, 2015) was used as the repeat database.

For positive data set preparation, CO sites smaller than 2 kb were selected and extended to 4 kb from their midpoint. After this step, the number of CO regions was 749 for tomato, 485 for rice, 69 for maize and 161 for Arabidopsis. For cases where CO regions overlapped, one of the two overlapping regions was randomly removed when the overlap was more than 25%, i.e. more than 1 kb. Moreover, CO regions were filtered if they overlapped with gaps in the reference genome. After filtering, the number of CO regions was 664 for tomato, 468 for rice, 63 for maize and 159 for Arabidopsis.

3.5.2 Sampling random cases from euchromatin or whole genome in tomato

We randomly selected 664 non-overlapping regions from tomato euchromatin, excluding CO regions and assembly gaps (i.e. N bases). Euchromatic region positions were previously calculated by Demirci *et al.* (2017). To sample these random regions, the bedtools 2.25.0 (Quinlan and Hall, 2010) shuffle function was used with the 'chrom' option, which protects the distribution of sequences among chromosomes. For example, if 10 sequences were present in chromosome 1 in the positive set, 10 sequences will be randomly selected on that chromosome for the random set. The same procedure was used to sample from the whole genome.

3.5.3 Sampling random cases from gene-dense regions

First, we generated a whole-genome gene density estimate using a kernel density procedure [scikit-learn 0.18 (Pedregosa *et al.*, 2011); python 3.5.2 (Python Software Foundation, https://www.python.org)]. We used the center position of every gene from the corresponding species annotation as a representation of the genes. The value of the kernel bandwidth was chosen such that the density would optimize the probability of the experimental CO distribution: the maximum log likelihood of the experimental CO distribution was found using a grid of 1000 different bandwidths, ranging from 1000 to 1 000 000, with increments of 1000. The optimum bandwidths obtained were 36 000, 7000, 171 000 and 54 000 for tomato, maize, rice and Arabidopsis, respectively. Then, to generate the negative set for each chromosome, n regions were randomly sampled, where n is the number of CO regions in that chromosome in the positive set. Then, the candidate regions were filtered for the presence of gaps (Ns), for overlaps between each other and for overlaps with any region in the positive set. If any of the initial candidates failed to pass the filtering, a new candidate was sampled from the distribution and the same filtering was applied. This process was repeated until n candidate negative regions passed all the filtering steps.

3.5.4 Feature preparation

For the positive and negative cases, the following features were calculated.

3.5.4.1 Features derived from sequence information

3.5.4.1.1 Dinucleotide frequencies

For each of the 16 possible dinucleotides, the following calculation was performed:

$$F_{AA} = n_{AA} / (L\text{-}1)$$

where F_{AA} indicates the frequency of dinucleotide AA, n_{AA} is the number of occurrences of AA in the given sequence, and L is the length of the sequence.

3.5.4.1.2 CTT and CCN motifs

As motifs, we used TCTTCTTC (Wijnker *et al.*, 2013) and CCNCCNCCN (Shilo *et al.*, 2015). The absence or presence of a motif in a region was described with a binary feature (motif presence), and the number of times a motif occurred in a region was described in the feature motif occurrence. Finally, motif search scores were obtained with fimo (Grant *et al.*, 2011). In the case of multiple occurrences of a given motif in a region, the following score was used to represent repetitive motifs:

score = ("motif score" / "motif length") x "total length",

where "total length" means the total length of sequences covered by the motif.

3.5.4.1.3 DNA structural features

Helix twist angle, propeller twist angle, MGW and roll were estimated for each nucleotide position in each region using the DNAshapeR algorithm (Chiu *et al.*, 2016). This approach predicts these structural properties for a given sequence using a model trained on experimental DNA structures (Zhou *et al.*, 2013): (i) propeller twist angle is a negative value that measures the perpendicular twist between two paired bases from different strands; (ii) helix twist angle is a positive angle between two adjacent base pairs as they twist in a DNA helix structure; (iii) MGW is the width of the DNA minor groove measured in angstrom (Å); and (iv) roll angle is the angle between two consecutive base pairs rolling over each other, which can be positive or negative. The values predicted for each nucleotide were averaged over a (positive or random) region to obtain a single value for each region. In addition, we calculated the minimum and maximum values estimated for each DNA structural feature for each region.

3.5.4.2 Features derived from genome annotation information

The distance from the centre of sequences to the nearest transcription start site (TSS) was calculated as described by Demirci *et al.* (2017). Briefly, the direct distance from the closest TSS position was calculated with the closest function in bedtools 2.25.0: a negative value means that the midpoint of a sequence lies upstream of the TSS. As the 5′-UTR regions were incomplete in the tomato genome annotation, we used mRNA start positions as TSS. For rice, maize and Arabidopsis, 5′-UTR regions were used.

The coding region fraction was calculated for each region. The gene elements that overlap with the regions were extracted by the intersect function in bedtools 2.25.0 from gene annotation files (ITAG 2.4 gene models file for tomato, IRGSP-1.0 representative locus and transcripts exon files for rice, TAIR 10 genes for Arabidopsis and the AGPv3.21 annotation file for maize). Subsequently, for each region, the total length of exonic regions was divided by the length of the region and reported as the coding region fraction of that region.

For each region, the transposon family fractions were calculated in a similar way as coding region fractions. Repeats that overlap with the regions were extracted by the intersect function in bedtools 2.25.0 from the repeat annotation files (ITAG 2.4 annotation repeat file ITAG2.4_repeats_aggressive.gff3 for tomato and repeat annotation files generated by repeatmasker for other species, see above). Then, for each region, the overlap fractions were calculated for all defined repeat families: the total length of the annotated repeats was divided by the length of the region. Repeat families were excluded as features if they were not present in any region in the data set of each species. For tomato, in addition, eu(chromatin) state was used as a feature in the first model (sampling from the whole genome); it was assigned as described by Demirci *et al.* (2017).

3.5.4.3 Features derived from parental genome information

The sequence divergence between parental genomes for a given region was calculated from VCF files of tomato parental genomes (*S. lycopersicum* Moneymaker and *S. pimpinellifolium*). The fastq files were downloaded from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) for *S. pimpinellifolium* (SAMEA2625653) under project number PRJEB6659 (Aflitos *et al.*, 2015), and for *S. lycopersicum* Moneymaker (SAMEA2340764) under project number PRJEB5235 (Aflitos *et al.*, 2014). These were mapped to the *Solanum lycopersicum* Heinz version SL2.50 reference genome, and variants were called with the same settings as described in Aflitos *et al.* (2014). From the resulting variant VCF files for each parental genome, containing SNPs with respect to the reference genome, SNPs were compared with each other, and homozygous SNPs with the same alternative alleles in the two parents, i.e. identical variants with respect to the reference, were removed. The remaining SNPs from the two genomes were combined to obtain parental SNPs and analysed to calculate the total number of SNPs, heterozygous SNPs and homozygous SNPs present in the regions as three separate features. In a similar way, INDELs with different lengths in the parental genomes were analysed to calculate the number of INDEL positions and the total length of differential INDEL lengths for each region. All five features from SNPs and INDELs were reported as a fraction of each analysed region.

Features were scaled individually by subtracting the mean and dividing by the standard deviation. The scaled features were used in later steps, unless otherwise stated. To cluster features, the absolute value of Pearson correlation between features was converted to a dissimilarity matrix using the equation:

$$D = 1 - \mathrm{abs}(\rho),$$

where D is the distance and ρ is Pearson correlation coefficient.

Based on the dissimilarity matrix, we performed hierarchical clustering with the hclust function in r using complete linkage. After manual inspection, a threshold of 0.4 was applied to define clusters.

To inspect the role of individual features, we performed a Student's *t*-test on non-scaled feature data using scipy 0.17.0 (Jones *et al.*, 2001). *P* values were Benjamini–Hochberg corrected using the multiple test function in statsmodels 0.8.0 (Seabold and Perktold, 2010). To visualize and detect the most significant features, the *P* values were log-transformed.

3.5.5 Comparative genomic analysis

For each species, we used the aforementioned genome annotation files for transposable element (TE) density and gene-density graphs. For CO density, we used the filtered set of CO regions that were used as a positive set to build the models. For SNP density, we used the parental marker set if provided by the original study (for tomato, Arabidopsis and rice); if not provided (in the case of maize), we identified the differential SNPs between parental genomes. To do so, raw sequence data sets of parental genomes Zheng58 (accession nos SRR449340, SRR449342 and SRR449343) and SK (accession no. SRR1585475) were downloaded from the ENA (https://www.ebi.ac.uk/ena). After trimming with trimmomatic 0.36 (Bolger *et al.*, 2014), reads were mapped to the reference genome AGPv3 by bowtie2 2.2.6 (Langmead and Salzberg, 2012) with the fast mapping option, PCR duplicates were removed, and SNPs for each parent were called by samtools 0.1.19 (Li *et al.*, 2009) and bcftools 0.1.19 (Li, 2011). SNPs with a coverage of fewer than four or more than 100 were filtered by bcftools. Finally, we reported the homozygous SNPs between parental genomes. Centromere

information was obtained as follows: for Arabidopsis, we used table S26 from Ziolkowski *et al*. (2017); for maize, we used the 1-Mb flanking region of CRM repeats as identified by repeatmasker; for tomato, we used data S1 from Demirci *et al*., 2017; and for rice, we inferred the approximate locations from a study conducted by Si *et al*. (2015: figure 3). Counts for different elements (COs, TEs, genes and SNPs) were obtained in 1-Mb bins across all chromosomes for a given species.

3.5.6 Classifiers

3.5.6.1 Decision tree classifier

We used the decision tree classifier algorithm implemented in scikit-learn 0.18 with the Gini impurity criterion to split the nodes. To prevent overfitting, the minimum number of samples on each leaf was set to five and the rest of the settings were left as default.

3.5.6.2 Random forest classifier

The random forest algorithm implemented in scikit-learn was used with 1000 trees in the forest. The remaining settings were kept at their defaults, with the number of features used at each split in each tree equal to the square root of the number of features, and the Gini criterion for splitting nodes.

3.5.6.3 Logistic regression

The logistic regression algorithm implemented in scikit-learn was applied. To optimize the regularization factor C, necessary to prevent overfitting, we used cross-validation over 10 different values from 1×10^{-4} to 1×10^{4}. After the prediction model was built, we used the absolute values of the coefficients to determine the feature importance values.

3.5.7 Comparison of feature importance values

Spearman rank correlation was calculated between feature importance values from different classifiers and different species. The resulting ρ value per pair of classifiers or species and the corresponding P value were reported in order to assess the similarity of the ranking of features.

3.5.8 Correlation/relationship of the features with CO prediction

To determine whether the predictive features have a positive or negative relationship on CO prediction, the mean value of a feature in the random set was subtracted from the mean value of a feature in the positive set. A positive sign means that higher values of that feature favor CO regions, and vice versa.

3.5.9 Evaluation of the performance of classifiers

The regions that include COs were defined as positive cases, whereas negative cases were the randomly selected regions. By comparing the prediction for a given case with its real label (CO or random), the following four values can be obtained: FP, the number of false positives (random cases predicted as CO); TP, the number of true positives (CO cases predicted as CO); FN, the number of false negatives (CO cases predicted as random); and TN, the number of true negatives (random cases predicted as random). To evaluate the performance of each predictor, we used the following evaluation metrics based on the values of FP, TP, FN and TN.

I. The AUROC is the area under the receiver operating characteristic (ROC) curve, which visualizes the true positive rate (TPR) versus the false positive rate (FPR). Here, TPR = TP/(TP + FN) is the probability of the detection of COs, and FPR = FP/(TN + FP) is the probability of wrongly predicting a random case as a CO.

II. Precision measures how many of the CO regions were correct among the cases predicted to be CO: Precision = TP/(TP + FP).

III. Recall measures how many of the experimental CO regions were correctly predicted to be CO: Recall = TP/(TP + FN), which is identical to the TPR.

IV. Accuracy measures how many of the instances are correctly predicted.

3.5.10 Validation of prediction models

We used 10-fold cross-validation to validate the prediction model. The data set was randomly split into 10 parts, which in 10 iterations each serve as a test set for a model trained on the remaining nine parts. The performance evaluation metrics are reported as the average and standard deviation over the 10 test sets.

To obtain additional validation on independent data, for the prediction models trained on CO regions and random regions obtained from gene-rich areas in the four species, a negative set was generated by sampling from the pericentromeric regions. The same number of regions as in the positive set (CO regions) was sampled from pericentromeric regions (excluding assembly gaps) with the same method as described above (using the shuffle algorithm in bedtools). The pericentromeric region locations were obtained as follows: for Arabidopsis, we used table S26 from Ziolkowski *et al.* (2017); for maize, we used 20-Mb flanking regions of CRM repeats, as identified by repeatmasker (excluding the CRM repeats); for tomato, we used heterochromatin regions defined in data S1 from Demirci *et al.* (2017); and for rice, we used cold spot regions defined in Si *et al.* (2015: table S4). Features were constructed for these regions in the same way as described above. To estimate the accuracy of the models, we assessed how many of the pericentromeric regions would not be CO regions according to the models.

In addition, for Arabidopsis, we used a genome-wide set of recombination rates (Choi *et al.*, 2013) for validation. For each genome region used in our Arabidopsis model, a single recombination rate was obtained by averaging the values provided by Choi *et al.* (2013). The distributions of these values were obtained separately for CO regions versus random regions, and for both types of regions separately based on whether the model predicted a region to be a CO region or a random region.

3.5.11 Availability

The scripts used for the analyses are available at https://github.com/sdemirci/predCO.

3.6 Acknowledgement

The work presented here is supported by the EU FP7 COMREC Marie Curie Initial Training Networks Programme project number 606956.

3.7 Supporting Information

Supplementary files are available at https://onlinelibrary.wiley.com/doi/full/10.1111/tpj.13979.

3.8 References

Aflitos, S., Schijlen, E., Jong, H. de, *et al.* (2014) Exploring genetic variation in the tomato (Solanum section Lycopersicon) clade by whole-genome sequencing. *Plant J.*, **80**, 136–148.

Aflitos, S.A., Sanchez-Perez, G., Ridder, D. de, Fransz, P., Schranz, M.E., Jong, H. de and Peters, S.A. (2015) Introgression browser: high-throughput whole-genome SNP visualization. *Plant J.*, **82**, 174–182.

Anderson, L.K., Lohmiller, L.D., Tang, X., *et al.* (2014) Combined fluorescent and electron microscopic imaging unveils the specific properties of two classes of meiotic crossovers. *Proc. Natl. Acad. Sci.*, **111**, 13415–13420.

Bao, W., Kojima, K.K. and Kohany, O. (2015) Repbase Update, a database of repetitive elements in eukaryotic genomes. *Mob. DNA*, **6**, 11.

Bell, O., Tiwari, V.K., Thomä, N.H. and Schübeler, D. (2011) Determinants and dynamics of genome accessibility. *Nat. Rev. Genet.*, **12**, 554–564.

Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences. *Nucleic Acids Res.*, **27**, 573–80.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114.

Chiu, T.-P., Comoglio, F., Zhou, T., Yang, L., Paro, R. and Rohs, R. (2016) DNAshapeR: an R/Bioconductor package for DNA shape prediction and feature encoding. *Bioinformatics*, **32**, 1211–3.

Choi, K., Zhao, X., Kelly, K. a, *et al.* (2013) Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat. Genet.*, **45**, 1327–36.

Choi, K., Zhao, X., Tock, A.J. *et al*. (2018) Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis transposons and gene regulatory regions. *Genome Res.* **28**, 532– 546.

Demirci, S., Dijk, A.D.J. van, Sanchez Perez, G., Aflitos, S.A., Ridder, D. de and Peters, S.A. (2017) Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between Solanum lycopersicum and Solanum pimpinellifolium. *Plant J.*, **89**, 554–564.

Edlinger, B. and Schlögelhofer, P. (2011) Have a break: Determinants of meiotic DNA double strand break (DSB) formation and processing in plants. *J. Exp. Bot.*, **62**, 1545–1563.

Gan, Y., Guan, J., Zhou, S. and Zhang, W. (2012) Structural features based genome-wide characterization and prediction of nucleosome organization. *BMC Bioinformatics*, **13**, 49.

Gao, Z.-Y., Zhao, S.-C., He, W.-M., *et al.* (2013) Dissecting yield-associated loci in super hybrid rice by resequencing recombinant inbred lines and improving parental genome sequences. *Proc. Natl. Acad. Sci. U. S. A.*, **110**, 14492–7.

Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics*, **27**, 1017–1018.

de Haas, L.S., Koopmans, R., Lelivelt, C.L.C., Ursem, R., Dirks, R. and Velikkakam James, G. (2017) Low-coverage resequencing detects meiotic recombination pattern and features in tomato RILs. *DNA Res.*, **3**, 1213–6.

Hartley, P.D. and Madhani, H.D. (2009) Mechanisms that Specify Promoter Nucleosome Location and Identity. *Cell*, **137**, 445–458.

El Hassan, M.A. and Calladine, C.R. (1996) Propeller-Twisting of Base-pairs and the Conformational Mobility of Dinucleotide Steps in DNA. *J. Mol. Biol.*, **259**, 95–103.

He, Y., Wang, M., Dukowic-Schulze, S., *et al.* (2017) Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proc. Natl. Acad. Sci.*, **114**, 12231–12236.

Jones, E., Oliphant, T., Peterson, P., *et al.* (2001-) SciPy: Open Source Scientific Tools for Python. Available at http://www.scipy.org.

Jouffroy, O., Saha, S., Mueller, L., *et al.* (2016) Comprehensive repeatome annotation reveals strong potential impact of repetitive elements on tomato ripening. *BMC Genomics*, **17**, 624.

Kauppi, L., Jeffreys, A.J. and Keeney, S. (2004) Where the crossovers are: recombination distributions in mammals. *Nat. Rev. Genet.*, **5**, 413–424.

Lambing, C., Franklin, F.C.H. and Wang, C.-J.R. (2017) Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiol.*, **173**, 1530–1542.

Langmead, B. and Salzberg, S.L. (2012) Fast gapped-read alignment with Bowtie 2. *Nat. Methods*, **9**, 357–359.

Li, H. (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, **27**, 2987.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, X., Li, L. and Yan, J. (2015) Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat. Commun.*, **6**, 6648.

Lichten, M. and Goldman, A.S. (1995) Meiotic recombination hotspots. *Annu. Rev. Genet.*, **29**, 423–44.

Liu, G., Liu, J., Cui, X. and Cai, L. (2012) Sequence-dependent prediction of recombination hotspots in Saccharomyces cerevisiae. *J. Theor. Biol.*, **293**, 49–54.

Liu, S., Yeh, C.-T., Ji, T., Ying, K., Wu, H., Tang, H.M., Fu, Y., Nettleton, D. and Schnable, P.S. (2009) Mu transposon insertion sites and meiotic recombination events co-localize with epigenetic marks for open chromatin across the maize genome. *PLoS Genet.*, **5**, e1000733.

Mathelier, A., Xin, B., Chiu, T.P., Yang, L., Rohs, R. and Wasserman, W.W. (2016) DNA shape features improve transcription factor binding site predictions *in vivo*. *Cell Syst.* **3**, 278– 286.

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. and Grelon, M. (2015) The Molecular Biology of Meiosis in Plants. *Annu. Rev. Plant Biol*, **66**, 297–327.

Myers, S., Bowden, R., Tumian, A., Bontrop, R.E., Freeman, C., MacFie, T.S., McVean, G. and Donnelly, P. (2010) Drive Against Hotspot Motifs in Primates Implicates the PRDM9 Gene in Meiotic Recombination. *Science*, **327**, 876–879.

Pan, J., Sasaki, M., Kniewel, R., *et al.* (2011) A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, **144**, 719–731.

Pedregosa, F., Varoquaux, G., Gramfort, A., *et al.* (2011) Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.*, **12**, 2825–2830.

Python Software Foundation. The Python Language Reference, version 3.5.2. Available at https://docs.python.org/3.5/reference/index.html.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–2.

Rodgers-Melnick, E., Bradbury, P.J., Elshire, R.J., Glaubitz, J.C., Acharya, C.B., Mitchell, S.E., Li, C., Li, Y. and Buckler, E.S. (2015) Recombination in diverse maize is stable, predictable, and associated with genetic load. *Proc. Natl. Acad. Sci. U. S. A.*, **112**, 3823–8.

Rohs, R., West, S.M., Sosinsky, A., Liu, P., Mann, R.S. and Honig, B. (2009) The role of DNA shape in protein–DNA recognition. *Nature*, **461**, 1248–1253.

Sato, S., Tabata, S., Hirakawa, H., *et al.* (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Seabold, S., and Perktold, J. (2010) Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*. 57-61.

Sherman, J.D. and Stack, S.M. (1995) Two-Dimensional Spreads of Synaptonemal Complexes from Solanaceous Plants. VI. High-Resolution Recombination Nodule Map for Tomato (Lycopersicon esculentum). *Genetics*, **141**, 683–708.

Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. and Levy, A.A. (2015) DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell*, **27**, 2427–36.

Si, W., Yuan, Y., Huang, J., *et al.* (2015) Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F 2 plants. *New Phytol.*, **206**, 1491–1502.

Smit, A.F.A., Hubley, R and Green, P. (2013-2015) *RepeatMasker Open-4.0*. http://www.repeatmasker.org/.

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

Villeneuve, A.M. and Hillers, K.J. (2001) Whence meiosis? *Cell,* **106**, 647-650.

Wijnker, E. and de Jong, H. (2008) Managing meiotic recombination in plant breeding. *Trends Plant Sci.*, **13**, 640–6.

Wijnker, E., Velikkakam James, G., Ding, J., *et al.* (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife*, **2**, e01426.

Wu, J., Mizuno, H., Hayashi-Tsugane, M., *et al.* (2003) Physical maps and recombination frequency of six rice chromosomes. *Plant J.*, **36**, 720–730.

Yelina, N.E., Lambing, C., Hardcastle, T.J., Zhao, X., Santos, B. and Henderson, I.R. (2015) DNA methylation epigenetically silences crossover hot spots and controls chromosomal domains of meiotic recombination in Arabidopsis. *Genes Dev.*, **29**, 2183–202.

Zhou, T., Yang, L., Lu, Y., Dror, I., Dantas Machado, A.C., Ghane, T., Felice, R. Di and Rohs, R. (2013) DNAshape: a method for the high-throughput prediction of DNA structural features on a genomic scale. *Nucleic Acids Res.*, **41**, W56–W62.

Ziolkowski, P.A., Underwood, C.J., Lambing, C., *et al.* (2017) Natural variation and dosage of the HEI10 meiotic E3 ligase control Arabidopsis crossover recombination. *Genes Dev.*, **31**, 306–317.

Chapter 4

Chasing breeding footprints through structural variations in *Cucumis melo* and wild relatives

Published as S. Demirci, R.R. Fuentes, W. van Dooijeweert, S. Aflitos, E. Schijlen, T. Hesselink, D. de Ridder, A.D.J. van Dijk and S. Peters, *G3* 11: 1–12, 2021.
https://doi.org/10.1093/g3journal/jkaa038

4.1 Summary

Cucumis melo (melon or muskmelon) is an important crop in the family of the *Cucurbitaceae*. Melon is cross pollinated and domesticated at several locations throughout the breeding history, resulting in highly diverse genetic structure in the germplasm. Yet, the relations among the groups and cultivars are still incomplete. We shed light on the melon breeding history, analyzing structural variations ranging from 50 bp up to 100 kb, identified from whole genome sequences of 100 selected melon accessions and wild relatives. Phylogenetic trees based on SV types completely resolve cultivars and wild accessions into two monophyletic groups and clustering of cultivars largely correlates with their geographic origin. Taking into account morphology, we found six mis-categorized cultivars. Unique inversions are more often shared between cultivars, carrying advantageous genes and do not directly originate from wild species. Approximately 60% of the inversion breaks carry a long poly A/T motif, and following observations in other plant species, suggest that inversions in melon likely resulted from meiotic recombination events. We show that resistance genes in the linkage V region are expanded in the cultivar genomes compared to wild relatives. Furthermore, particular agronomic traits such as fruit ripening, fragrance, and stress response are specifically selected for in the melon subspecies. These results represent distinctive footprints of selective breeding that shaped today's melon. The sequences and genomic relations between land races, wild relatives, and cultivars will serve the community to identify genetic diversity, optimize experimental designs, and enhance crop development.

4.2 Introduction

The family of the *Cucurbitaceae* consists of around 975 species across 98 genera, some of which have economic importance like the genus *Cucumis* of which species like *melo* (melon) and *sativus* (cucumber) are well known for their nutritious value and taste, while other species have a characteristic fragrance like *dudaim*, or fruit shape (*e.g.*, "horned melon" and "armenian cucumber"). *Cucumis melo* is grown worldwide and has been cultivated into many varieties and include for example netted cultivars (*e.g.*, cantaloupe) or smooth skinned varieties (*e.g.*, honeydew, casaba). Over the past decades melon breeding focused on higher productivity and adaptation to different growing systems. According to the FAO, worldwide melon production in 2018 (www.fao.org/faostat) amounted to 27.4 million tons, reflecting the success of modern melon breeding. However, this massive production is threatened by lack of disease resistance and (a)biotic stress tolerance traits in melon. A few disease resistance genes have been identified in a few melon accessions such as Fom-2, Fom-1 locus for Fusarium wilt disease (Joobeur *et al.,* 2004; Tezuka *et al.,* 2009), QTL for Cucurbit yellow stunting disorder virus (Palomares-Ríus *et al.,* 2016), *cmv1* gene and two other QTLs for Cucumber mosaic virus (Guiu-Aragonés *et al.,* 2014), and a QTL for powdery mildew (Li *et al.,* 2017). Further research for discovering resistance genes and QTL as well as generating new commercial resistant lines are ongoing (https://cuccap.org/breeding/melon/). Furthermore, changing environmental conditions, the demand for improved food quality (nutritional properties, increased shelf life), and new consumer preferences become increasingly important. This motivates the development of adapted and advanced crops and efficient agricultural production systems that are more in line with environmental, economic, and social needs. To achieve this goal, it is important to disclose and use genetic diversity underlying important target crop traits by breeding. The wild species that are closely related to cultivated melon represent a rich genetic diversity that can be used for advanced introgression hybridization breeding to introduce economically important traits, either by direct crossing or via bridging species.

Selection of compatible breeding parents for inter-specific hybridization breeding requires insight into the phylogenetic relationships and the genomic context of target genes underlying key economical traits. However, phylogenetic relationships for the *Cucumis* clade are incomplete and are not undisputed. Cucumber and melon diverged only 10 million years ago (Sebastian *et al.,* 2010). Although the chromosome number of cucumber ($2n=14$) and melon ($2n=24$) is different, there is high synteny between two genomes (Huang *et al.,* 2009; Li *et al.,* 2011). Ancestral fusion of five melon chromosome pairs in cucumber was suggested besides the observation of several intra- and interchromosomal rearrangements (Garcia-Mas *et al.,* 2012). Phylogenetic trees based on internal transcribed spacer sequences of nuclear ribosomal DNA (Garcia-Mas *et al.,* 2004) and partial mitochondrial DNA (Renner *et al.,* 2007) revealed that the genus *Cucumis* is monophyletic and contains many diverse melon and cucumber like species. The two aforementioned studies, however, disagreed on the branching of species within *Cucumis*, most likely due to usage of nongenomic data. *Cucumis melo* is suggested to be domesticated several times independently in Africa, Australia, Asia, and two times in India (South Asia) (Endl *et al.,* 2018; Gonzalo *et al.,* 2019; Zhao *et al.,* 2019). Currently, *C. melo* is divided into two subspecies, *C. melo* ssp. *melo* and *C. melo* ssp. a*grestis* based on morphology, yet the genetic data (SNP) also supports the distinction (Zhao *et al.,* 2019). These subspecies have been further divided into groups even though some of these groups are heterogenous and phenotypically complex (mixed). The ssp. *melo* has been divided into eleven groups (*cantalupensis*, *reticulatus*, *adana*, *chandalak*, *ameri*, *inodorus*, *chate*, *flexuosus*, *dudaim*, *chito*, and *tibish*), whereas ssp. *agrestis* has been divided into five (*momordica*, *conomon*, *chinensis*, *makuwa*, and *acidulus*) (Pitrat 2008), referred to as the 16-group classification. A later study suggested a 19-group classification with changes as (1) to merge the *reticulatus* and *cantalupensis*; (2) to split the large *inodorus* into three sub groups; (3) and the addition of two new groups *kachri* and *indicus* (Pitrat 2016). It is also suggested to remove the subspecifies classification as these 19 groups sufficiently distinguish melon. Since the new classification has not been applied in the databases yet, we have followed the 16-group classification in this study. The population diversity of melon was further studied with genetic marker assays such as AFLP, SNP assays, and WGS (Esteras *et al.,* 2013; Zhao *et al.,* 2019). One study focused on the group level and found a mixed (diverse) population structure in the ssp. *agrestis* groups, more than the commercially important ssp. *melo* groups *inodorus* and *cantalupensis* (Esteras *et al.,* 2013). Another study focused on the differences between cultivated and wild accessions within subspecies, revealing that cultivated *melo* has a much higher nucleotide diversity than cultivated *agrestis* (Zhao *et al.,* 2019).

In addition to SNPs, structural variants (SVs) can hold valuable and distinctive phylogenetic information (Saxena *et al.,* 2014). SVs provide information on variations in genic content such as copy number variation (CNV), inversions, translocations and deletions. SVs in melon were previously analyzed at a small scale, including seven accessions (González *et al.,* 2013; Sanseverino *et al.,* 2015). It has been found that the rate of occurrence of SVs is similar to that found in other species (Saxena *et al.,* 2014; Sanseverino *et al.,* 2015). In contrast to melon, an extensive SV study was conducted by Zhang *et al.* (2015) on 115 cucumber accessions. The study revealed that SVs derived mostly from nonhomologous rearrangement followed by transposable element (TE) movement, and very few by nonallelic homologous recombination (Zhang *et al.,* 2015). Regions with increased recombination frequency were correlated (co-located) with high CNV in soybean and barley, but not in maize (Saxena *et al.,* 2014). On the other hand, in maize, presence/absence variation (PAV) of SV was not influenced by recombination but by short direct repeats (Woodhouse *et al.,* 2010). For genetic diversity assessments and *Cucurbitaceae* genomic studies, the production of an annotated melon reference genome represented a valuable resource (Garcia-Mas *et al.,*

2012). This reference was constructed from the accession DHL92, which is a cross between accession Songwhan Charmi (*C. melo* ssp. *agrestis)* and Piel de Sapo (*C. melo* ssp. *melo*). The reference DHL92 genome is improved by optical mapping, and transposons were reannotated in the improved version (Ruggieri *et al.,* 2018). The annotated reference revealed a considerable amount of the genome (44%) composed of TEs as a major source of SV. Three quarters of these TEs are retrotransposons, whereas the rest consists of DNA transposons and unclassified TEs (Ruggieri *et al.,* 2018).

To benefit from the allelic richness occurring in wild species and to advance (introgression) breeding for melon crop improvement, assessment of genetic diversity and insight in chromosome topology is required, which at the moment is lacking. Disclosure of the genetic diversity and chromosome synteny and collinearity will not only be of use for precision breeding, but is also required to achieve a thorough understanding of the fundamentals of melon genome evolution and the genetic basis of complex traits. To achieve this goal, we have sequenced 94 melon (*Cucumis melo*) accessions and 6 wild species related to melon. We show that this provides a valuable resource for genome wide SV analysis across a large panel of accessions that is representative of the major phylogenetic groups. We have assessed these SVs in 100 melon and melon related wild accessions and use these data to shed light on the evolutionary history of melon breeding from a structural genomics point of view.

4.3 Experimental procedures

4.3.1 Melon accessions

The melon accessions selected in this study consist of 25 ssp. *agrestis*, 69 ssp. *melo*, and 6 wild species *C. zeyheri, C. prophetarum, C. anguria, C. dipsaceus, C. myriocarpus*, *and C. ficifolius*. The details of accessions including country of origin, seed provider, accession number, common name and groups, and subspecies they belong to are given in Supplementary Table S1.

4.3.2 DNA isolation and sequencing

Leaf material from 100 melon accessions were grinded and subsequently used for DNA isolation using the Qiagen DNeasy Plant Mini Kit. DNA concentration was measured using Qubit fluorometric quantitation (Lifetechnologies, www. lifetechnologies.com/qubit.html). For each sample ~2 μg DNA was used as input for TruSeq DNA PCR-Free Library Preparation. DNA was sheared using a Covaris M220 to an average insert size of ~550 bp according to the manufacturer protocol. Samples were individually barcoded and combined in two pools of each 33 samples and one pool of 34 samples. Prior to sequencing on the HiSeq2500 a qPCR and a MiSeq Nano run were performed to balance the samples and optimize cluster density. Aired-end 126 bp sequences at a mean coverage of 31-fold (assuming a genome size of 450 Mbp) were generated.

4.3.3 Alignment of reads

The reads were trimmed by Trimmomatic version 0.36 (Bolger *et al.,* 2014) with settings of minimum base quality of 20 and minimum read length of 50. Then the first 3 bp from the 5' end was trimmed by cutadapt v1.16 (Martin 2011). We mapped the Illumina short read by bwa mem v0.7.7 (Li and Durbin 2009) with default settings to melon reference genome CM3.6.1 which can be found in melonomics website (https://www.melonomics.net).

4.3.4 SV detection

We followed the method described in Fuentes *et al.* (2019) to identify the SV events with some modifications. Briefly, inversions, deletions, and duplication events were identified with four SV calling methods namely Pindel (Ye *et al.,* 2009), LUMPY (Layer *et al.,* 2014), GROM (Smith *et al.,* 2017), and DELLY 2 (Rausch *et al.,* 2012) based on the aligned reads. The results from different callers were merged if an SV event was called by at least two callers and the locations had 80% reciprocal overlap. When merging the SV events, the borders of the SV events were used from a single caller with the following priority: Pindel, DELLY, GROM, and LUMPY, as explained by benchmarking in Fuentes *et al.* (2019). The SV is filtered out if the ratio of the breakpoint error, or the span of inaccurate boundaries, to the intersection of the SV calls from different callers is larger than 1. We also discarded SV events in chromosome 0 (combination of unassigned contigs) and SV events shorter than 50 bp.

4.3.5 SV validation criteria

To validate each SV, we looked at the discordant read information and read coverage in the read alignment file via the Integrative genomics viewer, IGV version 2.5.2 (Robinson *et al.,* 2011). A read pair is called proper if the first read is aligned in forward direction (F) and the second read is aligned in reverse direction (R)—so-called FR pair—and the insert size is approximately the library size; otherwise, the read pair is called discordant. We called a tandem duplication event if sufficient (at least 4) RF-oriented pairs are present at both ends of the duplication event. For the inversion events, we searched for either one of RR or FF-oriented pairs or presence of both of RR and FF-oriented pairs at the ends of inversion events. Finally, for large deletion events, we searched for FR pairs with insert size larger or smaller than expected insert size; for smaller deletion events, that is, smaller than the read size, we checked split reads. Alternatively, for deletions and duplications, when there is no supporting paired-read information, we looked at differences of read coverage at the inspected SV event compared to nearby regions (or its flanking regions).

4.3.6 Functional annotation (GO term enrichment)

The SVs were grouped based on their type (duplication, inversion, deletion) and based on the subspecies label of accession they were observed either *C. melo* ssp. *melo* or *C. melo* ssp. *agrestis*. The SVs observed in wild related species were grouped together under the name wilds. In total, nine groups were prepared. For each group, the genes overlapping with SV events are identified by using bedtools intersect tool version 2.25 (Quinlan and Hall 2010). Overlap is defined when either the full length of an SV is covered by a gene or 50% of a gene is covered by an SV. We used the genome annotation version 4.0: CM4.0.gff3 (Ruggieri *et al.,* 2018) downloaded from the melonomics website (https://www.melonomics.net).

The genes overlapping with the SVs in a group were tested against a population of genes (present in all nine SV groups) on enrichment of a certain Gene Ontology (GO) term (molecular function, biological process and/or cellular component) in that group compared to the rest of the SVs. For GO term enrichment analysis, we used Ontologizer 2.1 (Bauer *et al.,* 2008) with Parent-Child-Union approach (Grossmann *et al.,* 2007) and Benjamini–Hochberg (BH) multiple testing correction method (Benjamini and Hochberg 1995). As a summary file for GO terms, the go-basic.obo file (http://purl.obolibrary.org/obo/go/go-basic.obo) is used. The gene association file (required by Ontologizer) is prepared with custom scripts from the melon genome annotation v4.0. The alternative GO term IDs given in CM4.0 annotation file were replaced with the main GO term IDs (as given in go-basics.obo file) as the current version (2.1) of ontologizer does not take alternative IDs into account.

Any GO term in any gene set was counted as significant if BH-adjusted *P*-values were less than 0.05. Subsequently, these GO terms were summarized for each SV type with REVIGO (Supek *et al.,* 2011). A value is given to a GO term based on how many melon groups it was significantly overrepresented in [one for single, two for two groups, and three for all groups (ssp. *agrestis*, ssp. *melo*, and wilds)]. We did not take into account the GO terms which were significantly overrepresented in all groups.

4.3.7 Dendrograms

We constructed a cluster map as a combination of a heat map and a tree based on presence/absence of SV events. We used Euclidean distance to cluster the SV events via the clustermap function in the seaborn module version 0.9.0 (Waskom *et al.,* 2017) with Python 3. The presence/absence of SV events in 100 genomes were used to construct a maximum-likelihood (ML) tree dendrogram via RAxML version 8.2.12 (Stamatakis 2014) with settings BINGAMMA and 100 bootstrap. The best tree was visualized in MEGA version 10.0.5 (Kumar *et al.,* 2018) and Iroki (Moore *et al.,* 2020).

4.3.8 Correlation between SV and genomic features

The gene models and transposon data of melon genome annotation version 4.0 were retrieved from the melonomics website. The genome was binned with 100 kb bin size. For each bin, the coverage of duplication, inversion, deletion, gene models, and transposons were calculated. Pearson's correlation was applied to the coverage values of each pair of dataset.

4.3.9 Motif discovery

Sequence motif discovery was done with the MEME suite 5 (Bailey *et al.,* 2009). The 1000 bp flanking regions of inversion breakpoints were extracted from the reference genome where the inversion size is more than 1 kb to eliminate the redundancy (n=855). In total, 1710 sequences were used to discover sequence motifs around the breakpoints. We ran MEME in 'any number of repetitions' (anr) mode, allowing it to find motifs which may occur in any number as long as they are nonoverlapping; a minimum and maximum motif length of 6 and 50, respectively; and including reverse complement settings. As a background we used first-order Markov frequencies, in other words dinucleotide frequencies in the reference genome.

4.3.10 Data availability

The sequences generated during the current study are available in the European Nucleotide Archive repository, https://www.ebi.ac.uk under study PRJEB37978. Supplementary Figure S1 contains allele frequency counts for duplications, inversions, and deletions. Supplementary Figure S2 contains clustering of the duplications and melon accessions based on presence/absence of duplications. Supplementary Figures S3–S5 contain ML trees based on duplications, inversions, and deletions, respectively. Supplementary Figure S6 contains a ML tree with bootstrap values based on combined SV events. Supplementary Table S1 contains the details of accessions used in this study including country of origin, accession number, common name and groups, and subspecies they belong to. Supplementary Table S2 contains the PAV of 104 genes in linkage group V in 100 melon accessions. Supplementary Table S3 contains the overrepresented GO terms in genes overlapping with SV.

4.4 Results & Discussion

We have identified SVs in 100 melon genomes by combining multiple SV detection tools which use different types of information, as this proved better than using a single type of information or a single tool (Kosugi *et al.* 2019). Briefly, the SVs are identified from whole genome pair-end sequencing data, using three types of information including read depth split-read and paired-end information detected by four SV detection algorithms (GROM, LUMPY, Pindel, and DELLY). SV events were categorized into three types: deletions (*i.e.*, absence of sequence compared to the reference genome), inversions, and duplications. The observed length of SV events ranged from 50 bp (the minimum length threshold) to 500 kb (observed in deletions). In total, we find 1,805,000 SV events in 100 genomes, which were combined into 50,271 distinct events that are present in one or more accessions. The number and coverage of the SV events for each type is given in Table 4.1. A large number of deletions is observed compared to inversions and duplications. Similar large differences in number of observed SV between types are also reported in a population of cucumber and a few accessions of melon (Sanseverino *et al.* 2015; Zhang *et al.* 2015). Despite their large number, deletions in this melon population only cover twice the size of the genomic region covered by duplications and almost three times that of inversions. Most deletions are small in size. More than 65% of the deletions are shorter than 1 kb where only ~35% and ~45% of the events are shorter than 1 kb in inversions and duplications, respectively. The size compositions of events per SV type is given in Figure 4.1.

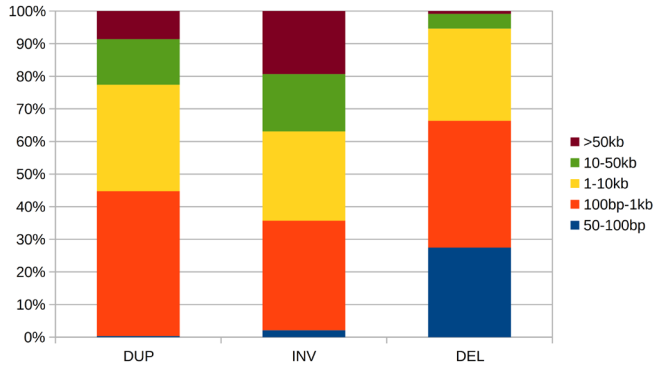

Figure 4.1 Size distribution of SV events per SV type. DUP: duplications; INV: inversions; DEL: deletions.

4.4.1 Distribution of structural variations

To see the patterns of structural variation (SV) along the genome, we performed a correlation analysis of SV types with respect to TEs and genes and to each other (Figure 4.2). A high proportion of all the SV types (47%–69%) were found to be co-located with TEs (Table 4.1). In line with this, all SVs showed positive correlation with TEs. While inversions and duplications show a weak correlation (r=0.1), deletions show a higher correlation (r=0.4) with TEs. The correlation between deletions and TEs is similar but opposite to the correlation between genes and TEs (r = −0.47). Furthermore, Fuentes *et al.* (2019) reported that 51.2% of SV's in the rice population overlap with TEs. A plausible explanation for the correlation between TEs and deletions is that TE insertions in the reference genome were identified as deletions in the studied genome. Additionally, a weak positive correlation of TEs with duplications might be due to the 'copy and insert' mechanism of TEs in new positions of the genome.

Table 4.1 Basic statistics of SVs found in all 100 melon samples (nonredundant SVs).

	Duplication	Inversion	Deletion
Number of SVs[a]	3,375	1,330	45,566
Coverage (total length of SV in bp)	82,650,295	66,786,711	166,990,885
Average[b] number of SV of 100 genomes	781	439	16,830
Average[b] coverage of 100 genomes (in bp)	38,218,127	47,630,536	81,786,636
Average[b] coverage of 94 cultivar genomes (in bp)	39,088,254	49,017,849	83,569,717
Percentage of SV in genic regions (at least 50% of the gene is covered)	28.39	37.59	5.77
Percentage of SV in genic regions (at least 100 bp of SV overlaps)	49.57	52.86	28.77
Percentage of SV in Transposons (at least 100 bp of SV overlaps)	67.38	69.32	47.77

[a] Nonredundant SV list. An SV can be observed in more than one genome.

[b] Average number of SVs per genome, over 100 genomes.

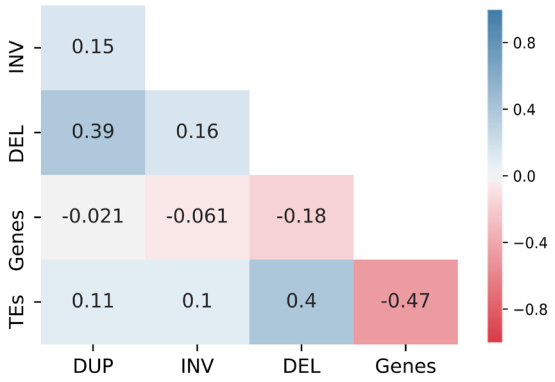

Figure 4.2 Correlation among SV events, genes and transposons. The Pearson correlation coefficients between pairs are given in cells. DUP, duplications; INV, inversions; DEL, deletions; TEs, transposable elements.

Half of the duplications and inversions partially overlap with genes, whereas only one quarter of deletions partially overlap with genes (Table 4.1). Despite this high rate of overlap, there is no correlation between duplications or inversions and genes, while there is a weak, negative correlation between deletions and genes ($r = -0.18$, Figure 4.2). In other words, inversions and duplications occur more in genic regions than deletions. Even though deletions have the lowest percentage of overlap with genes, since the number of deletions is much higher than other SV types, they affect more genes than inversions or duplications. Deletions affect 8869 genes of which 7821 were completely deleted, while inversions and duplications affected 4841 and 6198 genes, respectively.

We also observed a positive correlation between SV types. In particular, deletions and duplications ($r = 0.39$) showed a positive correlation. The overlapping sequences between duplications and deletions were 59 Mb (58,952,903 bp), comprising 15.7% of the melon genome. Highly variable regions of the genome are susceptible to more than one SV type. Specifically, we find that 62% of duplicated genes were deleted in either the same or different accessions. When a duplication occurs in one accession and a deletion of the same gene in another accession, this suggests that these genes were favored in some accessions, while in the other accessions the genes are lost or the gene function is compensated for by other genes. A possible explanation of the deletion and the duplication of the same gene on the same accession could be that these regions are hemizygous.

We further visualized the distribution of SV to capture the specific SV patterns in the chromosomes with respect to gene and TE distribution over the melon genome (Figure 4.3). The genome-wide correlation between deletions and TEs (r=0.4) can be seen in Figure 4.3 as they follow a similar pattern. Both deletions and TEs frequently occur in gene depleted regions, a pattern which is most striking on chromosome 6. Although inversions and duplications are sparse and general patterns were hard to deduce by visual inspection, there are some noteworthy regions. Since consistently empty bins, shown as noncolored areas in the *x*-axis of Figure 4.3, coincide with assembly gaps (NNs), finding their origin is nontrivial. Nevertheless, there are some large (Mb scale) contiguous empty areas in the SV plots. In chromosome 3 inversions are almost absent from the 9 to 14 Mb region. This region has 162 genes, many TEs and a substantial number of duplications and deletions. Similar to chromosome 3, a low number of inversions were observed on chromosome 10 compared to other chromosomes. Interestingly, chromosome 10 has the lowest recombination rate among the 12 chromosomes when measured in hybrids of *agrestis* and *melo* accessions (Argyris *et al.,* 2015; Chang *et al.,* 2017). While genome structure and recombination influence each other, cause and effect are difficult to distinguish. The low number of inversions in low recombination regions seems in contrast with the assumption of inversions blocking recombination; however, it is in line with recombination as a source of SV (see section "Inversions are the result of meiotic recombinations").

4.4.2 Allele frequency of SV among melon genomes

An SV event is regarded as unique if observed in only one accession. On the contrary, a general SV is observed in all genomes (such an event can also be regarded as unique to the reference genome). The unique SVs build up a large proportion of the total SVs, in particular for inversions (68%) and to a lesser extent also for duplications (30%) and deletions (22%) (Supplementary Figure S1). These SVs are specific to a single genome, conferring a high degree of diversity. Remarkably, the unique SVs originating from six wild species correspond to 30% of unique deletions, 25% of unique duplications, and 15% of inversions. An interesting observation is that a substantial number of unique inversions do not directly originate from wild species. It appears that inversions rather than duplications and deletions are distinctive in cultivars. Duplication and deletion events are more often shared between cultivars, perhaps reflecting breeding efforts carrying advantageous genes (or traits) into other lines.

4.4.3 Diversity in melon accessions based on SV

As SVs are a measure of genetic diversity (like SNPs), they are useful to investigate the evolutionary relationships within wild accessions, melon subspecies, and even melon groups. To analyze these relationships, two sets of phylogenetic trees (or dendrograms) based on SV events in 100 genomes were constructed by using two approaches: clustering and ML. Both approaches were based on the combination of common and unique SVs present in the accessions. The first approach revealed clustering of SV events separating melon accessions into four main subsets (Figure 4.4 for inversions; Supplementary Figure S2 for duplications) for which SVs form distinct patterns for wild species, melo, agrestis I, and II subsets. Following the clustering of accessions based on SVs, we used a second approach for differentiating the evolutionary (phylogenetic) relations between accessions. A ML tree was constructed not only on each SV type but also on the combined SVs. In the combined tree, the higher number of deletions likely influenced the branching more than the relatively lower number of duplications and inversions. To understand the effect on the tree topology, multiple ML dendrograms based on different SV types were analyzed (Supplementary Figures S3–S5). The observations were based on the tree constructed from combined SVs (Figure 4.5) and reported when consistent with at least two of the

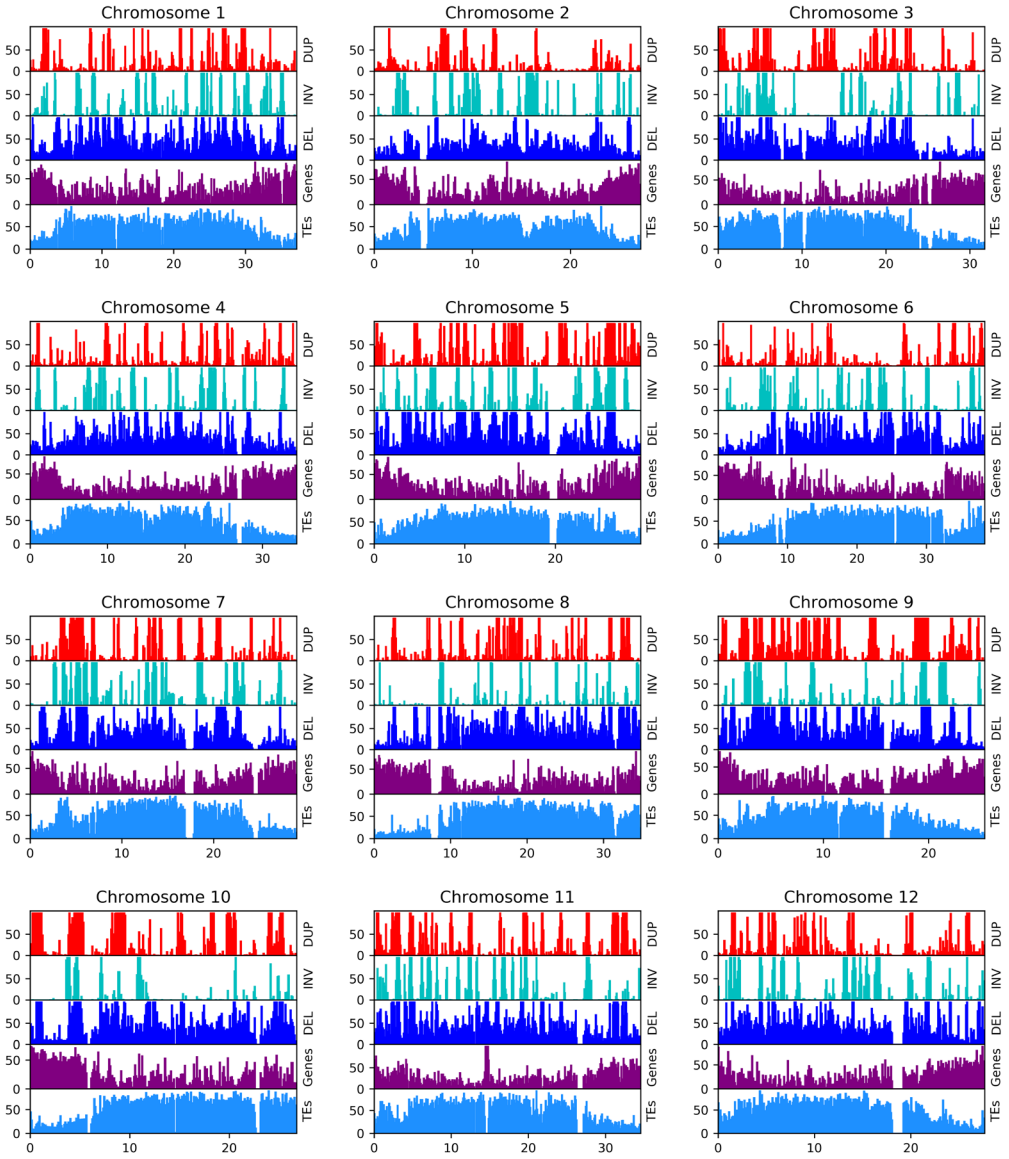

Figure 4.3 The distribution of SV elements over melon chromosomes. DUP, duplication (red); INV, inversion (cyan); DEL, deletion (blue). TEs, transposable elements (light blue) and genes (purple) of melon reference genome were added for comparison reasons. The *x*-axis is Mb scale; the *y*-axis is the percentage of coverage over bins of 100 kb. On the *y*-axis, 50 means that 50 kb of the 100 kb in that bin is covered by a given element.

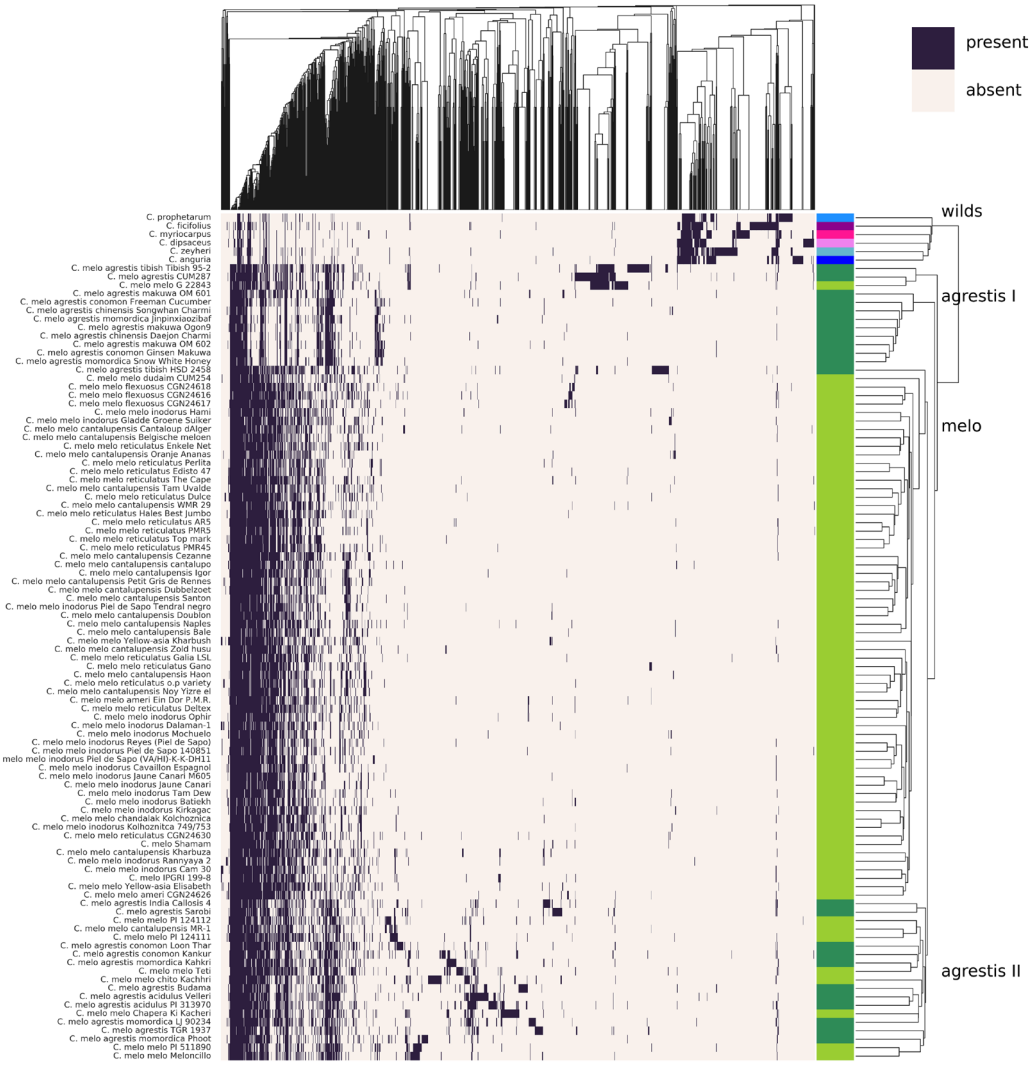

Figure 4.4 Clustering of the inversions (top) and melon accessions (right) based on PAV of inversions. The leaves of the cladogram of accessions are color coded based on their species and subspecies; light green: *C. melo* ssp. *melo*, dark green: *C. melo* ssp. *agrestis*, light blue: *C. zeyheri*, blue: *C. prophetarum*, dark blue: *C. anguria*, rose: *C. dipsaceus*, pink: *C. myriocarpus*, and purple: *C. ficifolius*.

three SV types. As can be seen in the tree, a subset of accessions belonging to ssp. *agrestis* (*agrestis* I in Figure 4.5) is clustered together independent from phenotypic grouping. Interestingly, this branch contains all *agrestis* accessions found in eastern Asia (China, Japan, and Korea). The remaining *agrestis* accessions originate from middle-western Asia (India, Pakistan, and Afghanistan) and Africa in line with the suggested independent domestication events (Endl *et al.,* 2018; Gonzalo *et al.,* 2019; Zhao *et al.,* 2019). These are polyphyletic, as they show high diversity within ssp. *agrestis*. This is in agreement with earlier phylogenetic studies, showing that ssp. *agrestis* is polyphyletic with accessions clustering by geographic origin (Stepansky *et al.,* 1999; Esteras *et al.,* 2013).

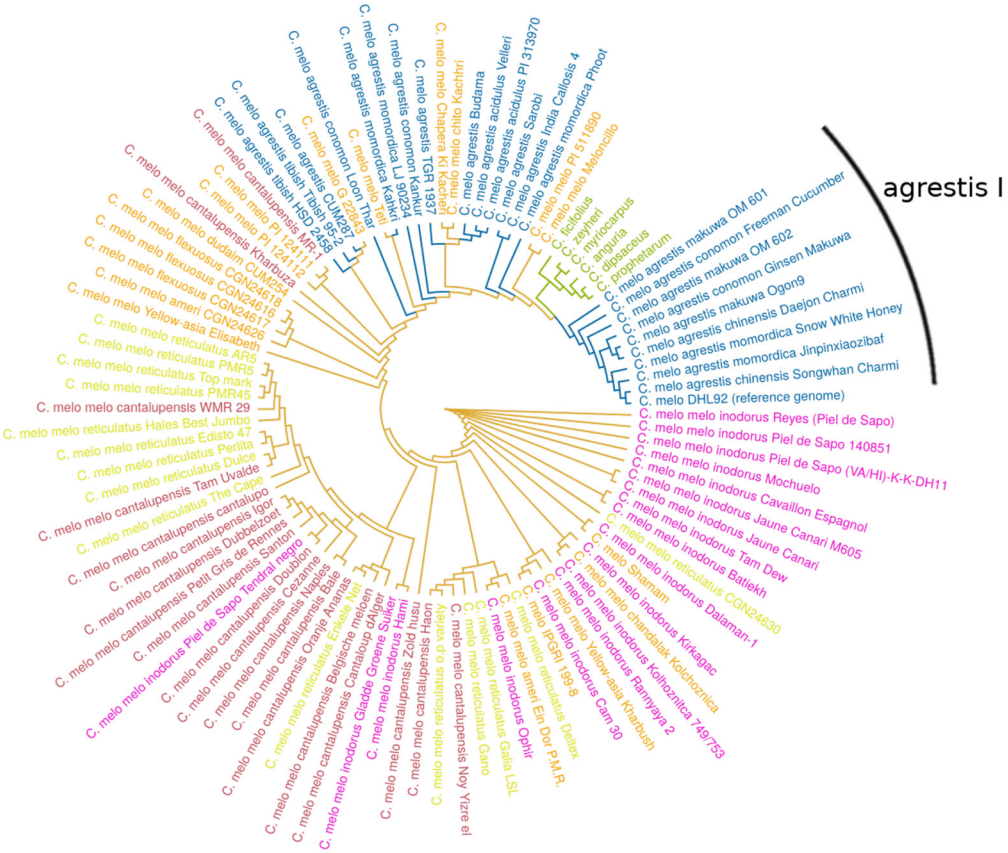

Figure 4.5 ML tree based on combined SV events (duplications, inversions, and deletions). Color codes refer to species identity. Blue: ssp. *agrestis*; green: wild Cucumis species; pink: *inodorus*; brown: *cantalupensis*; light green: *reticulatus*; orange: the rest of ssp. *melo* (yellow Asia, *chandalak*, *flexuosus*, *dudaim*, *ameri*, *chito*, ungrouped ssp. *melo*).

Recently, Zhao *et al.* (2019) analyzed more than thousand melon accessions, including cultivars and landraces. In their SNP-based phylogenetic tree wild African *agrestis* accessions cluster (*C. melo* var. *agrestis*) and are sister to the cultivated African accesions (*C. melo* ssp. *agrestis* var. *tibish*). As can be seen in the SV-based tree (Figure 4.5), indeed the branch with *tibish* accessions, corresponding to the African group (as these accessions originate from Sudan, Senegal, and Nigeria), cluster together. The *tibish* subclade is sister to the agrestis II group represented by *agrestis* accessions predominantly originating from India (*e.g.*, accs. Phoot, Kahkri, Velleri, Budama, Kacheri, Kahrbuza, and Teti). This is in line with Zhao *et al.*_(2019), reporting on the unexpected close relationship between African cultivated *agrestis* accessions with *agrestis* accessions from India. Furthermore, the *momordica* accessions from India grouped closely together with cultivated melon accessions in the SNP tree (Zhao *et al.,* 2019). Indeed, we observed some cultivated melo accessions grouping with *momordica* accessions from India in the SV tree. Interestingly, the *agrestis* group with Indian accessions, is relatively closely related to six *Cucumis* species (*C. ficipholus*, *C. zeyhri*, *C. myriocarpus*, *C. anguria*, *C. dipsaceus*, and *C. prophetarum*) all originating from Africa, clearly separating the African/Indian agrestis group from the agrestis I group, containing cultivated *agrestis* accessions, predominantly originating from

China, Japan, and Korea (*e.g.*, accessions OM601, OM602, Ginsen Makuwa, Daejon Charmi, Songwhan Charmi, and Jinpinxiaozibaf). Consistent for the observation of wild and cultivated *melo* accessions by Zhao *et al.* (2019), the landraces of ssp. *melo* such as the *flexuosus* accessions, *ameri* CGN24626, and *dudaim* CUM254, which are wild *melo* accessions are branching of from the *reticulatus*, *cantalupensis*, and *inodorus* accessions, which are cultivated melo accessions. Both the SV-based tree and the SNP tree from Zhao *et al.* (2019) show strong geographic separation and the overall topology of distinct groups in both trees is comparable.

The reference genome (DHL92) is grouped with one of its parents, Songwhan Charmi (SC; ssp. *agrestis*) on one end of the tree. Although the other parental accession (T111, Piel de Sapo ssp. melo) is not included in the SV-based tree (Figure 4.5), accessions related to T111, cluster with the *inodorus* clade on the other end of the tree. Previous findings on the comparative analysis of parental lines to the DHL92 genome showed that SC has slightly less SNPs than T111 (Garcia-Mas *et al.,* 2012; Sanseverino *et al.,* 2015). On the other hand, PAV analyses of genes showed that DHL92 is closer to T111 (González *et al.,* 2013). In conclusion, due to lack of T111 in the SV tree, and conflicting evidence from SNP and PAV analysis, we cannot confirm nor deny that DHL92 is closer to one or the other parent.

The wild accessions consistently grouped together; however, the relationship among the accessions could not be deducted as branching was not consistent among the trees (Figures 4 and 5, Supplementary Figures S2–S5). Previously, two studies have also found deviating branching (Garcia-Mas *et al.,* 2004; Renner *et al.,* 2007), probably due to the use of different sets of SV data and different algorithms. Therefore, the topological context of wild species should be considered cautiously. More consistent genomic relationships between these wild species could be obtained using whole genome assemblies.

The *C. melo* ssp. *melo* group together except for six accessions, five of which previously have not been assigned to any group. These six accessions currently group with ssp. *agrestis* [Meloncillo, PI 511890, Kachhri (*chito* group), Chapera Ki Kacheri, Teti, G22843]. It is likely that the level of mixture in some of these accessions makes it hard to categorize phenotypically. On the other hand, the morphology of G22843 and Kachhri accessions are similar to ssp. a*grestis*, they have a very small fruit size and green exocarp, suggesting that these accessions are possibly mis-categorized. Moreover, we observed phenotypic similarities between the rest of these ssp. *melo* accessions and their neighboring ssp. *agrestis* accessions. This raises suspicion for further mis-categorization of these ssp. *melo* accessions.

In the study by Sanseverino *et al.* (2015), the sample PI 124112 groups with ssp. *melo*, although it is assigned to the *momordica* group belonging to ssp. *agrestis*. In our study, accessions PI 124111 and PI 124112 also cluster with ssp. *melo* together with accession MR-1, a cultivar claimed from the *cantalupensis* group (bootstrap value 100, Figure 4.5, Supplementary Figure S6). Since the breeding line MR-1 was derived from PI 124111 (Thomas 1986; Li *et al.,* 2017), it is logical that these lines cluster together. Moreover, PI 124111 and PI 124112 lines are landraces which could explain the location of the branch next to other landraces from ssp. *melo* and ssp. *agrestis*. Another accession showing discrepancies with previous studies is accession CUM254 from the *dudaim* group. This accession clusters with *C. melo* ssp. *melo*, in agreement with Sanseverino *et al.* (2015), but unlike previous studies where an unknown *dudaim* accession was grouped with ssp. *agrestis* (Stepansky *et al.,* 1999; Esteras *et al.,* 2013). Although only seven samples were studied in Sanseverino *et al.* (2015) and care should be taken to make conclusions about the phylogenetic tree position of *dudaim*, we suggest that the *dudaim* accession CUM254 is more similar to ssp. *melo*.

In summary, both phylogenetic trees based on SV types completely resolve cultivar and wild accessions into two monophyletic groups. The two *C. melo* ssp. *agrestis* and *melo* are separately grouped as shown previously by SNP (Esteras *et al.,* 2013) and Inter-SSR (microsatellite) data (Stepansky *et al.,* 1999). Only 6 cultivars out of 94 were misplaced in these two subspecies branches when taking into account their morphology. However, the groups within these branches have not been resolved completely (Figure 4.5), implying a lack of specific SV events. Alternatively, this could be partially due to outdated classification of groups. As stated in the introduction, it has been proposed to merge the *reticulatus* and *cantalupensis* and split the large *inodorus* into three subgroups (Pitrat 2016). Yet, even if the groups were reclassified based on morphology, we would still expect polyphyletic groups. As melon cultivars have been transported between continents/countries and frequently crossed, a mixed diversity can be expected. This is consistent with the findings of Esteras *et al.* (2013), showing the mixed population structure of *momordica, inodorus landraces, ameri, flexuosus, indian agrestis*, and *dudaim.* Yet, the clustering of commercial groups like *cantalupensis* and *inodorus* (Figure 4.5) suggests a close but distinct relationship.

4.4.4 Validation of SV

We manually verified the paired-read alignment file of each sample to validate the SVs. We randomly selected 50 SVs per SV type (in total 150 SVs) to determine whether there is evidence of SV in the paired-read alignment file, either by looking at discordant read information or read coverage. We find that 93% (139 out of 150) of the SV events were detected correctly. We also observed that 8 of the 11 incorrect events constituted a more complex event, such as overlapping or nested SVs. This is in line with earlier reports on complex SV occurrences in plant genomes (Saxena *et al.,* 2014). Although we acknowledge the existence of complex events, for simplicity we reported all SVs as single SVs even though they might contribute to a complex event.

We observed that 48% (24 out of 50) of the deletion events overlap with assembly gaps (NN regions) of at least 80% of their size. This high rate of overlap between deletions and gaps is not observed between inversions or duplication and gaps. In fact, less than 1% of these events were found to overlap with gaps. Among all deletions observed (n=1,154,949), only 36% overlap at least 80% of their size with a gap, less than in the 50 randomly selected deletions. These 36% show a similar pattern of allele distribution compared to all deletion events, which suggests that gaps do not bias deletion detection. Since deletion events are correlated with transposons (see section "Distribution of SV"), we argue that these gaps could be unassembled transposons in the reference genome. Due to the difficulty of the assembly of repeat regions, the unassembled repeat regions were filled with NNs to the same size as in the reference genome. Indeed, a recent study revealed young transposons in the melon genome assembly v4.0 in the regions that previously were unassembled (Castanera *et al.,* 2020). Although probable, currently it is not yet clear whether these new young LTR transposons in the genome v4.0 can be assigned to the gaps of the v3.6.1 genome used in this study.

4.4.5 SV diversity in linkage group V

Previous studies showed that linkage group V, which contains resistance genes of melon including the NBS LRR TIR region and the CC NBS LRR protein coding *Vat (Virus aphid transmission)* locus, exhibits PAV (González *et al.,* 2013; Sanseverino *et al.,* 2015). In this study, we have not only confirmed the PAV genes in the same sample (Songwhan Charmi), we also find PAV in the genes annotated in the recent genome assembly version (Supplementary Table S2). The *Vat* locus shows either partial or full deletion in most of the samples (94%). The eight samples, having a complete *Vat* locus, belong to the *inodorus* (n=7) and *cantaloup* (n=1) group. This region might have

originated from a related wild species. Indeed, wide-cross attempts between cultivated and wild *Cucumis* species, such as *C. melo* × *C. metuliferus* and *C. metuliferus* x *C. dipsaceus*, resulted in viable offspring (Van Raamsdonk *et al.,* 1989; Chen and Zhou 2011). Also *C. melo* × *C. dipsaceus* resulted in fruit induction, although the fruits did not contain seeds. The closest wild species having nearly complete *Vat* locus genes is *C. dipsaceus* containing 10 out of 13 genes. *C. dipsaceus* could possibly be one of the ancestors of this region in melon perhaps via another "bridging species." Alternatively, there could be other ancestors of this locus that we have not analyzed in this study.

For the NBS LRR TIR region, we observed that the last third of this region is deleted in four of the wild species and partially deleted in *C. zeyheri* and *C. ficifolius*. Noting that the beginning of the region is absent in *C. zeyheri*, we suspect that these two species contributed together to shape the current NBS LRR TIR region. A summary of PAV of the whole linkage region, spanning 1.1 Mb and 104 genes, is given in Supplementary Table S2.

4.4.6 Genes affected by SV events are linked to melon breeding

In our analysis, 17,300 genes with a GO annotation were found to be affected by SV events in 100 melons. To investigate whether genes with the same function are specific for a particular group of melon accessions, we applied GO enrichment analysis for genes affected by SV in three sets of melon accessions belonging to ssp. *agrestis*, ssp. *melo* and wild relatives of *C. melo*. We found 26 overrepresented GO terms related to molecular and biological processes (Supplementary Table S3) linked to genes that are likely targeted for breeding. These GO terms are related to fragrance, fruit ripening, and stress response. In particular, for the fragrance-related GO term, the l-phenylalanine metabolic process is found to be overrepresented in inversions of wild species, compared to cultivar SVs. l-phenylalanine is an aromatic amino acid, and secondary metabolites derived from l-phenylalanine are fragrance and plant defence related. Another GO term, fruit ripening-related pectinesterase inhibitor activity, likewise differentiates wild species from cultivars. It has been found to be overrepresented in deletions and duplications of ssp. *melo* and ssp. *agrestis*, but not in wild species. Pectinesterase inhibitors are regulators of pectin esterases which are involved in fruit ripening, thickness of morphology of peel (skin), as well as defense against pathogens (Bethke *et al.,* 2016). Most likely, genes involved in pectin biosynthesis underlying fruit development have been selected for multiple times through different lines of breeding in different locations for particular cultivars, although it seems not specific for all melon subspecies. Beside general functions, we observed different stress response functions specific to melon subspecies. In agrestis duplications we observed enrichment of the wounding response process as well as oxylipin biosynthetic and metabolic processes. Oxylipins are an important class of signaling molecules in plants related to plant stress responses and innate immunity, and we speculate that ssp. *agrestis* gained resistance genes by duplication events during the breeding process. Additionally, another stress response-related function, calcium-dependent phospholipid binding, is found enriched in melo duplications. This function is associated with the annexin class of genes in duplications, which is differentially regulated by calcium changes induced by abiotic stress (Cantero *et al.,* 2006). Given that the melon genome has relatively few resistance genes compared to other plant species (Garcia-Mas *et al.,* 2012), it is not surprising that melon breeding resulted in an increase of stress response or resistance related genes via duplication events. We also see the effect of different domestication paths during melon breeding on the expansion of different sets of stress response genes as in two subspecies of melon.

Similar findings were reported for genes affected by SV in ssp. *agrestis* and *melo* by Sanseverino *et al.* (2015). Their study also finds genes associated with agronomically relevant pathways including disease resistance, aroma volatiles metabolism, sugar metabolism, and more. Together with results reported here, these findings shed more light on the selective breeding history that shaped today's melon.

4.4.7 Inversions as a result of meiotic crossovers

Meiotic recombination can result in various SV types, including inversions (Gaut *et al.,* 2007). It should therefore be possible to find the origin of SV from the breakpoint features. To this end, we searched for sequence motifs in the regions flanking the breakpoints of inversions whose size is more than 1 kb (n=855). We observed a 29-bp long poly A/T motif present in 1000 out of 1710 sequences, with an E-value of 8.4e−10. The observation of similar poly A/T stretches (Figure 4.6), at meiotic recombination breakpoints in relatively gene-rich euchromatic regions of tomato and Arabidopsis (Wijnker *et al.,* 2013; Demirci *et al.,* 2017) suggests that inversions are mostly the result of meiotic recombination events in melon as well. Although heterochromatic regions that are known to have a low recombination rate also can show high rates of rearrangements such as observed in Arabidopsis and wheat (Hall 2006; Lysak *et al.,* 2006; See *et al.,* 2006), the rearrangement dynamics in heterochromatic regions is quite different, as in general natural selection is inefficient and less deleterious (Gaut *et al.,* 2007), allowing for more uncompromised accumulation of rearrangements over time, while rearrangements in euchromatin swept through generations. Thus, our results for melon are consistent with the view that recombination is the major driving force for inversions in plants and possibly in all eukaryotes (Demirci *et al.,* 2017).

4.5 Conclusions

In this study, we investigated the evolutionary history of melon breeding from a SV point of view. We presented a broad range of SVs in 100 melon genomes, ranging from 50 bp up to 100 kb, which gives an idea of how collinear these genomes are for a region of interest as well as over the whole genome. By analyzing the SVs among the cultivated melon and wild melon relatives, we shed light on the phylogenetic relation between melon accessions. We showed that the resistance genes in the linkage V region are expanded in the cultivar genomes compared to wild relatives. We also found that particular agronomic traits are specifically selected in the melon subspecies such as fruit ripening, fragrance, and stress response. The findings reported here can help us to understand the selective breeding history through backward deductive analysis of events that shaped today's melon. Furthermore, we provide an inventory of SVs which can be used to shape future melon breeding strategies.

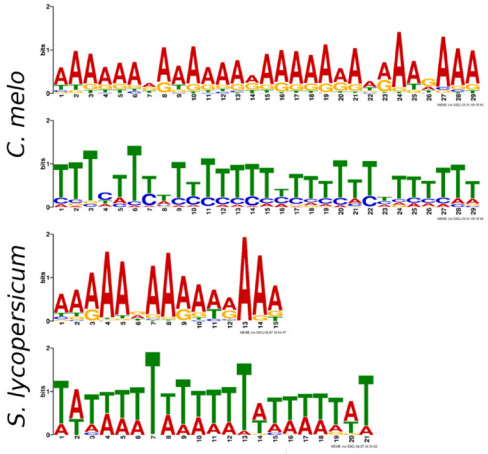

Figure 4.6 Sequence motifs around the inversion breakpoints (*Cucumis melo*) and crossover breakpoints (*Solanum lycopersicum*). Note: the two *C. melo* motifs are each other's reverse complement.

4.6 Acknowledgements

We thank Hortigenetics Research (S.E. Asia) Limited, Rijk Zwaan Zaadteelt en Zaadhandel B.V., Nunhems Netherlands B.V., Vilmorin & Cie for providing materials, Keygene N.V. for sequencing, Bas te Lintel Hekkert for sequencing, and Henri van de Geest for QC and bioinformatics support.

S.D. was supported by the EU FP7 COMREC Marie Curie Initial Training Networks program project number 606956. This project was financed by the Topsector Horticulture and Propagation Materials project 100 Melon Genome Project (https://topsectortu.nl/nl/100-meloen-genoom-project) project number 1310-034.

4.7 Supporting Information

Supplementary files are available at figshare, https://doi.org/10.25387/g3.13312088.

4.8 References

Argyris, J.M., Ruiz-Herrera, A., Madriz-Masis, P., Sanseverino, W., Morata, J., Pujol, M., Ramos-Onsins, S.E. and Garcia-Mas, J. (2015) Use of targeted SNP selection for an improved anchoring of the melon (Cucumis melo L.) scaffold genome assembly. *BMC Genomics*, **16**, 4.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W. and Noble, W.S. (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, **37**, W202-8.

Bauer, S., Grossmann, S., Vingron, M. and Robinson, P.N. (2008) Ontologizer 2.0--a multifunctional tool for GO term enrichment analysis and data exploration. *Bioinformatics*, **24**, 1650–1651.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J. R. Stat. Soc. Ser. B*, **57**, 289–300.

Bethke, G., Thao, A., Xiong, G., *et al.* (2016) Pectin Biosynthesis Is Critical for Cell Wall Integrity and Immunity in Arabidopsis thaliana. *Plant Cell*, **28**, 537–556.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–20.

Cantero, A., Barthakur, S., Bushart, T.J., Chou, S., Morgan, R.O., Fernandez, M.P., Clark, G.B. and Roux, S.J. (2006) Expression profiling of the Arabidopsis annexin gene family during germination, de-etiolation and abiotic stress. *Plant Physiol. Biochem.*, **44**, 13–24.

Castanera, R., Ruggieri, V., Pujol, M., Garcia-Mas, J. and Casacuberta, J.M. (2020) An Improved Melon Reference Genome With Single-Molecule Sequencing Uncovers a Recent Burst of Transposable Elements With Potential Impact on Genes. *Front. Plant Sci.*, **10**, 1–10.

Chang, C.-W., Wang, Y.-H. and Tung, C.-W. (2017) Genome-Wide Single Nucleotide Polymorphism Discovery and the Construction of a High-Density Genetic Map for Melon (Cucumis melo L.) Using Genotyping-by-Sequencing. *Front. Plant Sci.*, **8**.

Chen, J. and Zhou, X. (2011) Cucumis. In C. Kole, ed. *Wild Crop Relatives: Genomic and Breeding Resources*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 67–90.

Demirci, S., Dijk, A.D.J. van, Sanchez Perez, G., Aflitos, S.A., Ridder, D. de and Peters, S.A. (2017) Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between Solanum lycopersicum and Solanum pimpinellifolium. *Plant J.*, **89**, 554–564.

Endl, J., Achigan-Dako, E.G., Pandey, A.K., Monforte, A.J., Pico, B. and Schaefer, H. (2018) Repeated domestication of melon (Cucumis melo) in Africa and Asia and a new close relative from India. *Am. J. Bot.*, **105**, 1662–1671.

Esteras, C., Formisano, G., Roig, C., *et al.* (2013) SNP genotyping in melons: Genetic variation, population structure, and linkage disequilibrium. *Theor. Appl. Genet.*, **126**, 1285–1303.

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

Garcia-Mas, J., Benjak, A., Sanseverino, W., *et al.* (2012) The genome of melon (Cucumis melo L.). *Proc. Natl. Acad. Sci.*, **109**, 11872–11877.

Garcia-Mas, J., Monforte, A.J. and Arús, P. (2004) Phylogenetic relationships among Cucumis species based on the ribosomal internal transcribed spacer sequence and microsatellite markers. *Plant Syst. Evol.*, **248**, 191–203.

Gaut, B.S., Wright, S.I., Rizzon, C., Dvorak, J. and Anderson, L.K. (2007) Recombination: an underappreciated factor in the evolution of plant genomes. *Nat. Rev. Genet.*, **8**, 77–84.

González, V.M., Aventín, N., Centeno, E. and Puigdomènech, P. (2013) High presence/absence gene variability in defense-related gene clusters of Cucumis melo. *BMC Genomics*, **14**.

Gonzalo, M.J., Díaz, A., Dhillon, N.P.S., Reddy, U.K., Picó, B. and Monforte, A.J. (2019) Re-evaluation of the role of Indian germplasm as center of melon diversification based on genotyping-by-sequencing analysis. *BMC Genomics*, **20**, 1–13.

Grossmann, S., Bauer, S., Robinson, P.N. and Vingron, M. (2007) Improved detection of overrepresentation of Gene-Ontology annotations with parent child analysis. *Bioinformatics*, **23**, 3024–3031.

Guiu-Aragonés, C., Monforte, A.J., Saladié, M., Corrêa, R.X., Garcia-Mas, J. and Martín-Hernández, A.M. (2014) The complex resistance to cucumber mosaic cucumovirus (CMV) in the melon accession PI161375 is governed by one gene and at least two quantitative trait loci. *Mol. Breed.*, **34**, 351–362.

Hall, A.E. (2006) Dynamic evolution at pericentromeres. *Genome Res.*, **16**, 355–364.

Huang, S., Li, R., Zhang, Z., *et al.* (2009) The genome of the cucumber, Cucumis sativus L. *Nat. Genet.*, **41**, 1275–1281.

Joobeur, T., King, J.J., Nolin, S.J., Thomas, C.E. and Dean, R.A. (2004) The Fusarium wilt resistance locus Fom-2 of melon contains a single resistance gene with complex features. *Plant J.*, **39**, 283–97.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. and Kamatani, Y. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, **20**, 8–11.

Kumar, S., Stecher, G., Li, M., Knyaz, C. and Tamura, K. (2018) MEGA X: Molecular Evolutionary Genetics Analysis across Computing Platforms F. U. Battistuzzi, ed. *Mol. Biol. Evol.*, **35**, 1547–1549.

Layer, R.M., Chiang, C., Quinlan, A.R. and Hall, I.M. (2014) LUMPY: a probabilistic framework for structural variant discovery. *Genome Biol.*, **15**, R84.

Li, B., Zhao, Y., Zhu, Q., Zhang, Z., Fan, C., Amanullah, S., Gao, P. and Luan, F. (2017) Mapping of powdery mildew resistance genes in melon (Cucumis melo L.) by bulked segregant analysis. *Sci. Hortic. (Amsterdam).*, **220**, 160–167.

Li, D., Cuevas, H.E., Yang, L., *et al.* (2011) Syntenic relationships between cucumber (Cucumis sativus L.) and melon (C. melo L.) chromosomes as revealed by comparative genetic mapping. *BMC Genomics*, **12**, 396.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Lysak, M.A., Berr, A., Pecinka, A., Schmidt, R., McBreen, K. and Schubert, I. (2006) Mechanisms of chromosome number reduction in Arabidopsis thaliana and related Brassicaceae species. *Proc. Natl. Acad. Sci.*, **103**, 5224–5229.

Martin, M. (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal*, **17**, 10.

Moore, R.M., Harrison, A.O., McAllister, S.M., Polson, S.W. and Wommack, K.E. (2020) Iroki: automatic customization and visualization of phylogenetic trees. *PeerJ*, **8**, e8584.

Palomares-Ríus, F.J. and Garcés-Claver, A. (2016) Detection of two QTLS associated to Cucurbit yellow stunting disorder virus in the TGR 1551 melon line. In: E.U. Kozik, H.S. Paris and M.L. Gómez-Guillamón, eds. *Cucurbitaceae 2016, XIth Eucarpia Meeting on Genetics and Breeding of Cucurbitaceae*, July 24–28, 2016. Warsaw, Poland, pp. 334–337.

Pitrat, M. (2008) Melon. In J. Prohens and F. Nuez, eds. *Vegetables I. Handbook of Plant Breeding, vol 1*. New York, NY: Springer New York, pp. 283–315.

Pitrat, M. (2016) Melon Genetic Resources: Phenotypic Diversity and Horticultural Taxonomy. In R. Grumet, N. Katzir, and J. Garcia-Mas, eds. *Genetics and Genomics of Cucurbitaceae*. Cham: Springer, pp. 25–60.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rausch, T., Zichner, T., Schlattl, A., Stutz, A.M., Benes, V. and Korbel, J.O. (2012) DELLY: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, **28**, i333–i339.

Renner, S.S., Schaefer, H. and Kocyan, A. (2007) Phylogenetics of Cucumis (Cucurbitaceae): Cucumber (C. sativus) belongs in an Asian/Australian clade far from melon (C. melo). *BMC Evol. Biol.*, **7**, 58.

Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G. and Mesirov, J.P. (2011) Integrative genomics viewer. *Nat. Biotechnol.*, **29**, 24–26.

Ruggieri, V., Alexiou, K.G., Morata, J., *et al.* (2018) An improved assembly and annotation of the melon (Cucumis melo L.) reference genome. *Sci. Rep.*, **8**, 1–9.

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 Insertions, 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**.

Sebastian, P., Schaefer, H., Telford, I.R.H. and Renner, S.S. (2010) Cucumber (Cucumis sativus) and melon (C. melo) have numerous wild relatives in Asia and Australia, and the sister species of melon is from Australia. *Proc. Natl. Acad. Sci. U. S. A.*, **107**, 14269–73.

See, D.R., Brooks, S., Nelson, J.C., Brown-Guedira, G., Friebe, B. and Gill, B.S. (2006) Gene evolution at the ends of wheat chromosomes. *Proc. Natl. Acad. Sci.*, **103**, 4162–4167.

Smith, S.D., Kawash, J.K. and Grigoriev, A. (2017) Lightning-fast genome variant detection with GROM. *Gigascience*, **6**.

Stamatakis, A. (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–3.

Stepansky, A., Kovalski, I. and Perl-Treves, R. (1999) Plant Systematics and Evolution Intraspecific classifcation of melons (Cucumis melo L.) in view of their phenotypic and molecular variation. *Plant Syst. Evol*, **217**, 313–332.

Supek, F., Bošnjak, M., Škunca, N. and Šmuc, T. (2011) REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms C. Gibas, ed. *PLoS One*, **6**, e21800.

Tezuka, T., Waki, K., Yashiro, K., Kuzuya, M., Ishikawa, T., Takatsu, Y. and Miyagi, M. (2009) Construction of a linkage map and identification of DNA markers linked to Fom-1, a gene conferring resistance to Fusarium oxysporum f.sp. melonis race 2 in melon. *Euphytica*, **168**, 177–188.

Thomas, C.E. (1986) Downy and powdery mildew resistant muskmelon breeding line MR-1. *Hortic. Sci.*, **21**, 329.

Van Raamsdonk, L.W.D., Nijs, A.P.M. den and Jongerius, M.C. (1989) Meiotic analyses of Cucumis hybrids and an evolutionary evaluation of the genus Cucumis (Cucurbitaceae). *Plant Syst. Evol.*, **163**, 133–146.

Waskom, M., Botvinnik, O., O'Kane, D., *et al.* (2017) mwaskom/seaborn: v0.8.1 (September 2017).

Wijnker, E., Velikkakam James, G., Ding, J., *et al.* (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *Elife*, **2**, e01426.

Woodhouse, M.R., Schnable, J.C., Pedersen, B.S., Lyons, E., Lisch, D., Subramaniam, S. and Freeling, M. (2010) Following Tetraploidy in Maize, a Short Deletion Mechanism Removed Genes Preferentially from One of the Two Homeologs K. H. Wolfe, ed. *PLoS Biol.*, **8**, e1000409.

Ye, K., Schulz, M.H., Long, Q., Apweiler, R. and Ning, Z. (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics*, **25**, 2865–2871.

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.

Zhao, G., Lian, Q., Zhang, Z., *et al.* (2019) A comprehensive genome variation map of melon identifies multiple domestication events and loci influencing agronomic traits. *Nat. Genet.*, **51**, 1607–1615.

Chapter 5

Transposon insertion polymorphisms in tomato cultivars

Sevgin Demirci, Xuenan Pi, Dick de Ridder, Sandra Smit and Ruud A. de Maagd

5.1 Summary

Transposable element mobility is an important mechanism of genome evolution in plants and contributes to the formation of many important traits in crops. Here we survey the activity of Long Terminal Repeat Retrotransposons (LTR-RTs) in a set of tomato (*Solanum lycopersicum*) accessions by cataloguing their shared and rare insertions. We first benchmarked our computational pipeline and found that in spite of a sometimes low recall, we are able to detect LTR-RT insertions with high precision. Inspection of the insertions of 21 (18 unique) tomato LTR-RTs in 6 diverse accessions resulted in a selection of four putatively active RTs, based on the number of insertions in genes and unique versus shared insertion ratios. For these active elements, insertions in 60 accessions including two *S. pimpinellifolium*, one *S. pennellii* and one *S. peruvianum* were surveyed. We found all four RTs to produce insertion polymorphisms, with *Rider* producing most (2248) insertions and the highest proportion in euchromatin and in genes. Annotation of the genes with insertions gave evidence of the presence of alleles known to be produced by *Rider* insertion or translocation, such as *sun*, *potato leaf* (*c*), *jointless-2* (*j-2*), and *yellow flesh* (*r*). Moreover, a large number of newly identified insertions from all four LTR-RTs may result in phenotypes with potential for tomato breeding. While some of the insertion polymorphisms in cultivated tomato can be traced back to identifiable introgressions from wild species, many unique insertions are testimony of likely LTR-RT mobility within the timeframe of tomato domestication and breeding. Induction of mobility of these transposons can in the future be used for the production of more genetic diversity and improved traits in tomato.

5.2 Introduction

Transposable elements (TEs) or transposons are relatively autonomously acting DNA sequences able to move through genomes, excising - or copying in case of retrotransposons - from their location and inserting into a new genome location. Plant transposable elements can be divided into Class I elements or retrotransposons (RTs), which are the most common type, and Class II elements or DNA transposons. Retrotransposons transpose via a "copy-and-paste" mechanism in which an mRNA molecule is produced from an internal promoter, which is reverse transcribed into cDNA and then integrated at a new position in the genome by an integrase. They can be further divided into LTR (Long Terminal Repeat) retroelements (dominant in plants) and non-LTR elements, and may be autonomous or non-autonomous (the latter requiring enzymes encoded by related autonomous elements) (Lisch, 2013).

Depending on the order of the genes they contain (Group-specific Antigen, Protease, Integrase, Reverse transcriptase, Ribonuclease H), the LTR-RTs are further divided into the superfamilies Ty1/*Copia* and Ty3/*Gypsy* (Galindo-González *et al.*, 2017). Further division within the superfamilies is based on the total sequence identity: if the RT sequences are more than 80% identical, they are assumed to belong to the same family (Wicker *et al.*, 2007). If an RT retained all the above genes, it is referred to as a "full" copy. With time passing after insertion, and as a result of silencing of its activity, RTs degenerate, causing sequence divergence in their LTRs and losing part or all of the coding sequence and thereby their activity. Alternatively, recombination between LTRs may eliminate the internal sequence leaving solo LTRs or truncated elements.

The transposition of TEs in genomes may cause mutations in genes, by insertion, imprecise excision, or gene duplication and rearrangements during excision (Huang *et al.*, 2012; Lisch, 2013). 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 specific varieties, such as in maize where they were first discovered, 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) when comparing two individuals, in which case the insertion is assumed to have occurred after divergence from a common ancestor.

Many examples of transposon insertions leading to phenotypes (including important traits) in crops exist. 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). Variation in epigenetic modification of transposons may lead to phenotypes, as demonstrated by the tissue culture-induced hypomethylation of a *Karma* retrotransposon in oil palm, leading to alternative splicing of a floral homeotic gene with the mantled flower phenotype as a result (Ong-Abdullah *et al.*, 2015). It is therefore clear that transposons in crops have had a role in the formation of both advantageous as well as detrimental traits.

High throughput sequencing techniques have revolutionized the discovery of transposable elements, their diversity, the resulting polymorphisms and their role in evolution, particularly for the human genome as exemplified by the results of the 1000 genomes project (http://www.1000genomes.org/, (Xing *et al.*, 2013)). The detection of mobile elements at the whole-genome level comprises two primary types of methods: i) targeted methods, which involve enrichment of DNA fragments related to TE's before sequencing or genotyping; and ii) post-sequencing bioinformatics methods using available whole-genome sequencing data, as reviewed in (Xing *et al.*, 2013; Ewing, 2015). With the advent of whole-genome (re-)sequencing, large-scale detection of structural variants, including RT insertion polymorphisms can be accomplished computationally using whole genome data, without much prior knowledge of TE characteristics. Several published methods exist for detecting RT insertions by anchored discordant read–pair mapping, in which one of the paired-end reads maps uniquely to the genome ("anchored") and the other maps to a consensus TE sequence. Additionally, with longer read pairs, a single read may span the genome/TE-junction, allowing it to locate the junction precisely ("split read-mapping"). Some algorithms integrate both methods (Rishishwar *et al.*, 2017). However, benchmarking these tools on simulated and real data for different types of retrotransposons (i.e. LINEs and SINEs) showed that each tool has unique biases and can report many false positives (Rishishwar *et al.*, 2017; Makałowski *et al.*, 2019).

Identification of insertion polymorphisms in tomato accessions will further improve our insight in their contribution to phenotypic diversity in cultivated tomato, including the part of that diversity that has been selected for during domestication or more recent breeding efforts. It will also enhance our understanding of their contribution to general genome plasticity and genetic diversity of tomato, as it has done for several crops already (Oliver *et al.*, 2013; Grzebelus, 2018). In several species, gene function is analysed by studying phenotypic effects of TE insertions. Moreover, induced transposition may increase genetic diversity in crops like tomato, supplementing more traditional mutagenesis methods. However, in-depth knowledge on the activity of transposons in this important crop is still lacking. Here, we present an inventory of putatively active retrotransposons, which we used to investigate insertion patterns. To this end, we screened whole genome resequencing data of selected tomato cultivars and a number of wild species for variation in RT insertions, annotated the resulting polymorphisms and linked these to known or possible phenotypes. This

provides a resource for further investigating polymorphisms in other tomato accessions as well as for studying the phenotypes of insertion events near or in genes of interest.

5.3 Results

5.3.1 Identification of LTR-retrotransposon families with insertion polymorphisms in tomato cultivars

To detect common RT insertions as well as polymorphic ones, we used ITIS (Identification of Transposon Insertion Sites), a method which uses split read and paired-end read information in NGS data that was previously applied to *Medicago truncatula* (Jiang *et al.*, 2015). We selected 21 LTR-RTs potentially active in tomato as they were identified as having recent and intact insertions (with identical LTRs and non-degenerate open reading frames) in the reference genome (Xu and Du, 2014; Paz *et al.*, 2017) for an initial screen. These 21 candidate RTs have an intact size range of 4294 to 14054 bp and an LTR size range of 275 to 3602 bp. They belong to the superfamilies Ty1/*Copia* (*n*=17) and Ty3/*Gypsy* (*n*=4), which are common in the tomato genome. The sizes, clades, and sources of selected RTs are listed in Table 5.1 (locations for sequence extraction are listed in Table S1).

ITIS is limited by the insert size of the reads, in this study 500 bp, to detect insertions. To check whether ITIS can distinguish the different RTs by their first 500 bp, we constructed a similarity matrix based on the first and last 500 bp (mostly LTR sequences) of each RT, which is shown in Table S2. As a control we used 2 pairs of RTs already known to be similar (Paz *et al.*, 2017): SL_RT_F17-SL_RT_F142 (F17/F142) and SL_RT_F66- SL_RT_F25 (F66/F25). We found that the 500 bp end regions of RTs were on average 66% identical to each other, while those of the known similar pairs were at least 94% identical. Beside the two known pairs, one other pair of selected RTs, SL_RT_F50-SL_RT_F108 (F50/F108), showed high (98.4%) similarity between the RT ends. Due to these similarities, the ITIS results of each pair were merged, resulting in a set of 18 unique LTR-RT families.

5.3.1.1 Benchmarking ITIS

To understand how well ITIS performs at detecting TE insertions based on similarity to given TE sequences, we benchmarked the method by comparing the insertions reported by ITIS with those actually present in the Heinz genome. Briefly, ITIS was applied to read data generated from the reference genome *S. lycopersicum* cv. Heinz (with the actual TEs masked, see Experimental Procedures) to find the TE insertions. As a ground truth, the output of RepeatMasker (RM) for these TE types was taken. We focused only on calls which were labelled ‘copy’ or ‘copy,gap’ by ITIS, not on novel insertions (labelled ‘new’, see section 5.5.3 for definitions of these labels). Based on a comparison of these reported insertions to the ground truth, we calculated a false discovery rate (FDR) and false negative rate (FNR). As an illustration, examples of positive and negative cases are shown in Figure S1.

We find that the FDR is often 0 and always less than 10% for all 18 TE families (Table S3, column set A). Since the content of assembly gaps in ‘copy,gap’ cases obviously cannot be fully confirmed, we calculated the same statistics on only ‘copy’ cases. Similar to before, the FDR was under 10% (Table S3, column set B). The FNR was high, reflecting the conservative settings of ITIS as well as the difference in the approaches: ITIS is read-based and focuses on similarity of short, outermost regions of LTRs only, whereas RM is reference-based and can also report longer similar sequences including TE gene content. Manual inspection confirmed that a number of false negatives were due to an insufficient number of supporting reads or the presence of

Table 5.1 Tomato LTR-RTs for which we determined characteristics in 6 tomato accessions.

Name[1]	Name[2]	Name	Length	LTR	Super family	Inser-tions	Unique /All	% in heteroch-romatin	% in genes
SL_RT_F322	GypsySL_01	*Jinling*[3]	8834	1824	Gypsy	1724	0.68	84.8	0.9
SL_RT_F160	CopiaSL_37	*Rider*[4]	4877	400	Copia	1164	1.78	38.9	22.2
SL_RT_F108	CopiaSL_25		5538	275	Copia	306	0.35	67.3	5.2
SL_RT_F120	CopiaSL_29		6112	859	Copia	1268	0.36	80.3	5.0
SL_RT_F132	CopiaSL_19	*ToRTL1*[5]	9651	812	Copia	353	0.24	76.5	8.5
SL_RT_F142	GypsySL_05		12281	3602	Gypsy	1108	0.82	80.4	6.4
SL_RT_F155	CopiaSL_mt		4822	141	Copia	13	0	76.9	0.0
SL_RT_F159	GypsySL_01		1134	424	Gypsy	950	0.08	64.0	7.9
SL_RT_F164	CopiaSL_33	*T135*[6]	5504	735	Copia	313	0.14	71.6	10.5
SL_RT_F170	CopiaSL_32	*Tnt1/ TLC1/ Retrolyc*[7]	5367	630	Copia	98	0.15	67.4	12.2
SL_RT_F17	GypsySL_05		11634	3279	Gypsy	merged with F142			
SL_RT_F204	CopiaSL_12		4977	285	Copia	24	1.5	41.7	12.5
SL_RT_F233	CopiaSL_14		5502	322	Copia	20	2.5	25.0	20.0
SL_RT_F251	CopiaSL_08		5070	240	Copia	20	0.17	45.0	20.0
SL_RT_F272	CopiaSL_11		4966	283	Copia	22	0.18	31.8	13.6
SL_RT_F274	CopiaSL_16		5417	266	Copia	11	NA	18.2	27.3
SL_RT_F50	CopiaSL_38		4294	275	Copia	merged with F108			
SL_RT_F66	CopiaSL_08		5043	237	Copia	merged with F251			
SL_RT_F94	GypsySL_04		14054	1554	Gypsy	3709	0.26	96.2	1.3
	Copia_mt	*TARE1*[8]	5375	199	Copia	697	0.04	74.5	6.2
SL_RT_F2	CopiaSL_01	*TGRE1*[9]	4625	129	Copia	14	0.29	50.0	14.3

[1]Name according to (Xu and Du, 2014); [2]Name according to (Paz *et al.*, 2017); [3](Wang *et al.*, 2006); [4](Cheng *et al.*, 2009) [5](Daraselia *et al.*, 1996);[6,7](Tam *et al.*, 2005); [8,9](Yin *et al.*, 2013)

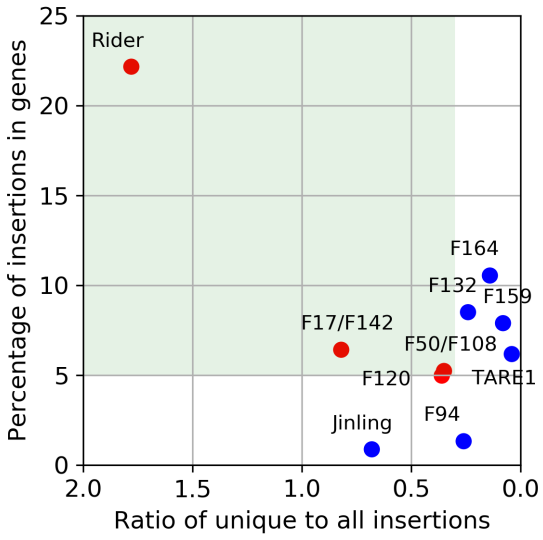

Figure 5.1 Tomato LTR-RT families with over 100 insertions in the 6 accessions characterized by the ratio of unique (present in only one accession) to all (present in all 6 accessions) insertions and percentage of insertions in genes. The four RTs shown in red were subsequently selected for further analysis of insertions and their polymorphisms in a larger set of tomato accessions. The green shaded box indicates the cut-off for selection by 5% in genes and by 0.3 unique/ all ratios.

conflicting read information, mostly for truncated TEs or solo LTRs. Therefore we subsequently focused on nearly-full length (80% length of the reference LTR-RT sequence) insertions, both in the RM truth set and reported ITIS cases. We calculated FDR and FNR for all copy cases ('copy' and 'copy,gap') as well as non-gap cases, since assembly gap size is unreliable as an indicator for TE length (Table S4). The FNR in non-gap cases of nearly-full length LTR-RTs (Table S4, column set B) was much lower than the FNR in non-gap cases of mixed size LTR-RT families (Table S3, column set B). Some of the RT families' FDRs were slightly increased while other RT families' FDRs dropped to 0 when short cases and gaps were not considered. We conclude that ITIS can detect LTR-RT insertions based on read information, yielding a low number of false discoveries (<10%) at a reasonably high number of false negatives (mostly 10-40%, with some extreme cases of >90%) in a situation which can be verified by reference-based similarity search (RM). Moreover, even though we cannot benchmark 'copy,gap' cases, we assume ITIS here has similar strengths and weaknesses as for 'copy' cases. Therefore, we included 'copy,gap' cases in our further analyses.

5.3.1.2 Identification of most active TE families

To identify the most active LTR-RT families from the initial set of 18 families, we explored the frequency of TIPs in few tomato accessions. We expected TIPs to be more frequent between genotypically different accessions, as they are likely to have diverged from a common parent longer ago, allowing for more lineage-specific transposition events. Therefore, we selected 6 representative *S. lycopersicum* accessions with increasing sequence divergence from the reference as shown by their SNP frequencies relative to the reference (Aflitos et al., 2014): Heinz 1706 (reference genome), Moneymaker (RF001/LA2706), Gardener's Delight (RF003/PI406760), Large Pink (R019/EA01049), and *S. lycopersicum* var. cerasiforme accessions RF037/LA1324 and RF105/LA1479. We obtained the LTR-RT insertions for 18 families in these 6 accessions. The observation of a unique insertion in a cultivar suggests that it is not inherited from a common ancestor but results from a transposition event after divergence from the nearest relative or from an introgression from another cultivar or wild species.

We defined an RT as recently active if it had a relatively high number of unique insertions compared to shared insertions among the 6 selected tomato accessions. We also took into account whether the insertions were found to overlap with gene bodies (UTRs, CDS and introns) to further narrow down our RT selection for analysis in all resequenced tomato genomes (for details, see Experimental Procedures). A summary of basic statistics of the 18 RT families is given in Table 5.1. This table lists, for each RT, the total number of insertions, the ratio of unique to shared with all insertions, the percentage of insertions in genes and in heterochromatin. Moreover, a plot of the ratio of unique to shared insertions versus the percentage of insertions in genes for RTs with more than 100 insertions is shown in Figure 5.1. Since calling insertions on a population of 60 accessions is computationally intensive, we focused on the most active RTs that have the most unique polymorphims and potentially the most effect on gene expression (i.e. on phenotype). In order to simultaneously optimize both the number of found polymorphisms relative to used computing time as well as the number of insertions found in genes, we selected four putatively active RT with at least 5% of insertions in genes and a higher than 3:10 ratio of unique to shared insertions for further analysis: SL-RT_F160 (*Rider*), F17/F142, and F50/F108, and F120.

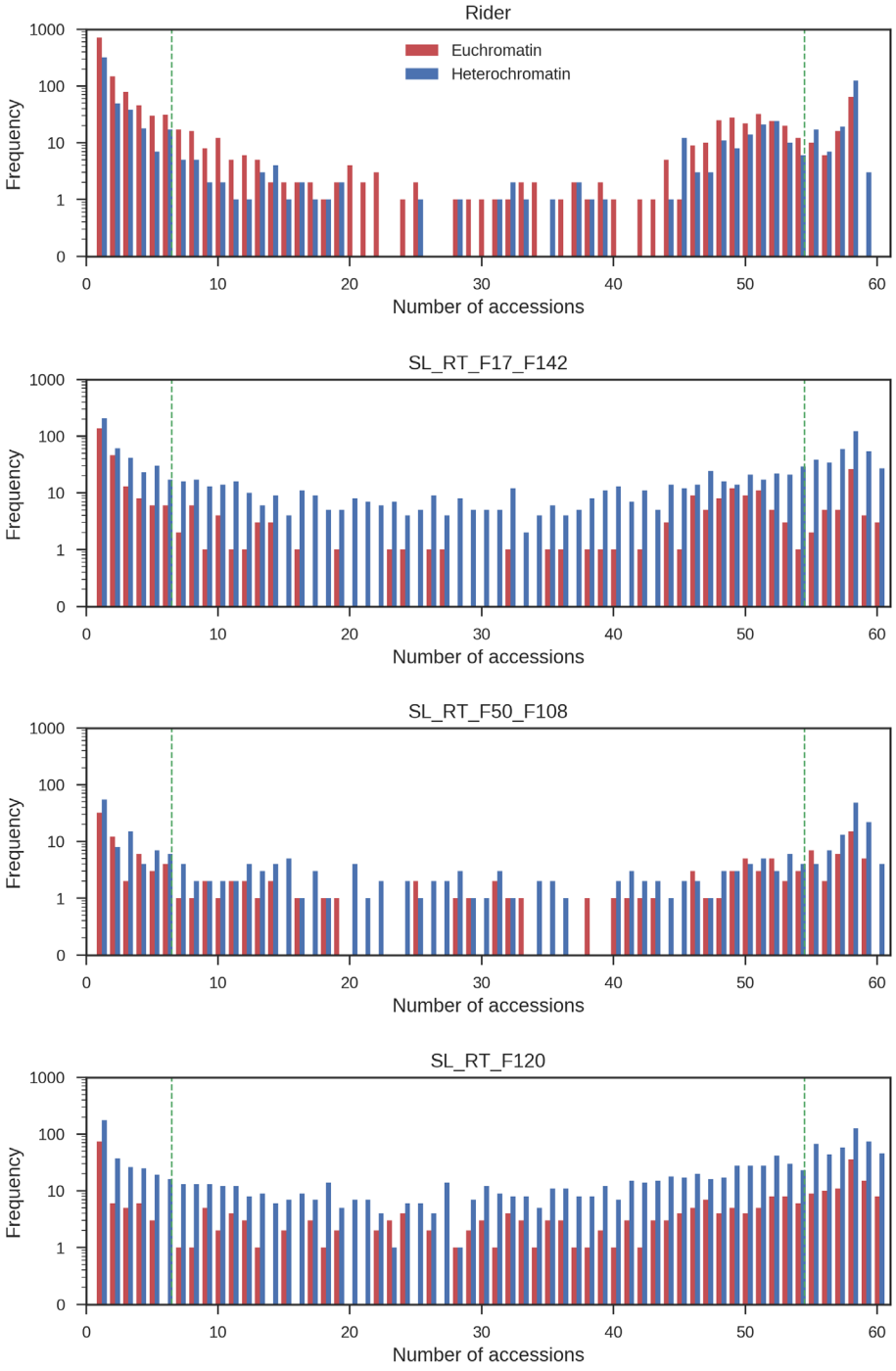

Figure 5.2 Frequency of insertions occurrence versus the number of accessions sharing them, found for four families of LTR-RTs over the 60 samples, and in euchromatin (red bars) or heterochromatin (blue bars). The thresholds for considering alleles rare (at most 10% allele frequency) and common (at least 90% allele frequency) are indicated by green dashed lines.

5.3.2 LTR-RT insertion polymorphisms in 60 tomato accessions

In order to identify RT insertions within *Solanum lycopersicum* (including. var. *cerasiforme*), we screened 60 accessions of the species comprising tomato cultivars and landraces as well as two *S. pimpinellifolium*, one *S. peruvianum* and one *S. pennellii* accessions (listed in Table S5) resequenced in the 150 Tomato Genome project at a coverage of ~35x (Aflitos *et al.*, 2014).

Insertions for all four RTs are present in all chromosomes, but not uniformly distributed as insertions of F17/F142, F50/F108 and F120 families were found mostly in the heterochromatin while *Rider* insertions are present at the ends of chromosomes in the gene-rich euchromatin (Table 5.2 and Figure S2).

Table 5.2 Distribution of active RT insertions over genomic regions in 60 accessions.

LTR-RT	number of insertions	ratio rare/ common	% in eu-chromatin	% in hetero-chromatin	% in genes	% in coding regions
Rider	2248	5.57 (1510/271)	64.0	34.3	28.1	12.7
SL_RT_F17/F142	1587	1.55 (595/383)	22.7	76.2	9.4	3.3
SL_RT_F50/F108	451	1.18 (157/133)	32.8	65.9	5.5	2.0
SL_RT_F120	1655	0.77 (399/518)	18.8	79.1	6.1	1.6

Note: % in euchromatin and % in heterochromatin do not add up 100% as some of insertions are in chromosome 0 of the assembly, which does not correspond to a physical chromosome.

In order to detect correlations between euchromatin/heterochromatin distribution and the frequency of sharing of insertions in the 60 accessions, we plotted the distribution of shared insertions over the chromatin as shown in Figure 5.2. For *Rider*, out of a total of 2248 insertions, 271 are shared by at least 90% of the accessions (common insertions) and 1510 occur in at most 10% of the accessions (rare insertions). Clearly *Rider* has the most insertions of the 4 RTs and, confirming the earlier conclusion with 6 accessions (Figure 5.1), the highest ratio of unique to shared insertions as well as the highest ratio of rare to common insertions. All four RTs show a peak of insertions shared by 58 accessions (Figure 5.2), which represent all *S. lycopersicum* and *S. pimpinellifolium* accessions excluding the two other wild accessions. The latter have fewer insertions in common with the former. Curiously, this peak is much more pronounced for *Rider* than for the other three: only one *S. pennellii Rider* insertion is shared with the 58 *S. lycopersicum* accessions (and not with *S. peruvianum)*, and 2 *S. peruvianum* insertions are shared with the 58 (and not with *S. pennellii)*. The other three RTs have considerably more insertions that are shared among 59 or even all 60 accessions. As shown in Table 5.2, 64% of all *Rider* insertions were found in the euchromatin. This percentage stays approximately equal throughout the graph in Figure 5.2, suggesting that unique insertions follow the general pattern of *Rider* insertions and are not further biased towards euchromatin or heterochromatin (Figure S3). In contrast to *Rider* the other three RT families, F50/F108, F17/F142 and F120, have fewer insertions in euchromatin, 33%, 23% and 19%, respectively (Table 5.2). However, similar to *Rider*, these percentages are independent of the frequency of occurrence or sharing (Figure S3).

5.3.3 Retrotransposons contribute to genetic diversity

We next investigated to which extent RT activity contributes to the differentiation of tomato genomes. To understand the variation between genomes, we generated a clustermap of insertions for each of the four RT families. In Figure 5.3, a heatmap based on the presence of each *Rider* insertion in 60 tomato genomes is shown, together with a tree based on the clustering of the insertion patterns. An outgroup containing *S. pennellii* and *S. peruvianum* is clearly visible in the clustermap (marked "1") as particularly *S. pennellii* has multiple unique *Rider* insertions and only a single insertion in common with the other genomes, on chromosome 10 between positions 45509901 and 45510299. The latter insertion is only 398 bp long and is found in all accessions except *S. peruvianum*. Second in the outgroup is *S. peruvianum*, in which we found fewer insertions (n=54) than in *S. pennellii* (n=134) overall, but 3 possible RT insertions were shared with cultivars. Among them the largest insertion is close to the full size of Rider (4877 bp) used as reference. Other outer branches (marked "2" and "3") are characterized by increasing numbers of insertions in common with cultivated tomatoes, and groups of insertions that are hardly or not shared with other accessions. The two *S. pimpinellifolium* accessions LA1578 and LYC2798 (RF47 and RF44, respectively) are in these branches and each share similarity with two *S. lycopersicum* accessions, LYC2962 (RF42)/LA1479 (RF105) (branch 2) and LYC3340 (RF17)/TR00024 (RF54) (branch 3), which may point to introgressions from wild species. These four accessions share only approximately half of the insertions common in most cultivars and have a relatively high number of unique insertions (n=50-158), comparable to the two outgroups. The remaining *S. lycopersicum* accessions are grouped together, as they mostly share Rider insertions. There is one more distinct group, composed of the accessions RF003/S.lyc PI406760 (cv. Gardeners Delight), RF007/S.lyc EA00375 (cv. Katinka Cherry), RF005/S.lyc EA00325 (cv. Galina) and S.lyc RF045/LYC2740. Investigation of an insertion map where Rider insertions are ordered according to their chromosome position (Figure S4) shows that much of the difference between these four accessions and the other accessions can be attributed to an apparent introgression spanning a large part of chromosome 4. This is confirmed by the similarity of SNP positions and density between these accessions and *S. pimpinellifolium* generated by iBrowser (Aflitos et al., 2015) (Figure S5a). When the insertion-based cluster tree of the accessions is compared to a phylogenetic tree of the 60 accessions based on the SNP distribution that was published earlier (Aflitos et al., 2014), there is 84% difference between the topologies of two trees according to the Robinson-Foulds metric (Figure 5.3 and Figure S6). This is most likely due to different topologies in the group of cultivars as outgroups of wild species are similar in both trees. As Rider has the highest frequency of insertions in genes and euchromatin, covering roughly 30% of the whole genome (TheTomatoGenomeConsortium, 2012), it is likely that if Rider insertions influence evolution, it will be noticeable in these regions rather than in the whole genome.

Clustermaps based on absence-presence variation of insertions of the other three active RTs are given in Figures S7-S12. These also show high dissimilarity between the RT insertion pattern-based phylogenetic tree and the SNP-based phylogenetic tree (normalized Robinson-Foulds distances 90%-95%). This suggests that RT activity independently contributes to genetic diversity, as it does not follow the general pattern of whole genome SNP variation, assuming the false negatives are not biased along the genome.

Overall, the impression from Figure 5.3 and Figures S7-S12 is that both *S. pennellii* as well as *S. peruvianum* carry far fewer insertions of all four investigated RTs than *S. lycopersicum* and *S. pimpinellifolium*. This is also reflected in a region on chromosome 9 that is particularly devoid of insertions of all four RTs in three tomato accessions

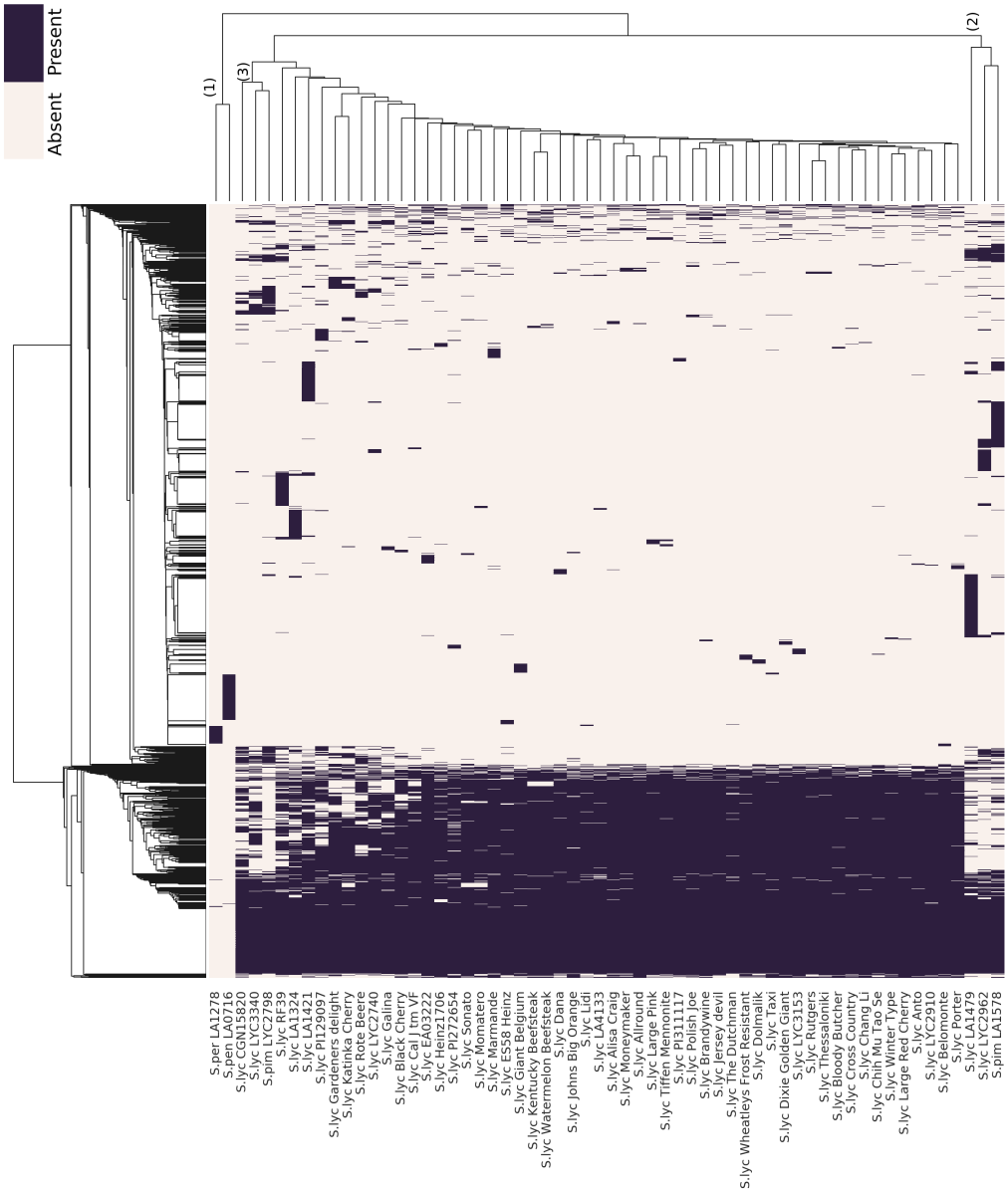

Figure 5.3 Clustermap of *Rider* insertions in 60 tomato genomes. A heatmap based on absence (light) and presence (dark) of specific *Rider* insertions, clustered by Euclidean distance (left). The 60 accessions (bottom) are ordered according to a phylogenetic tree based on insertion patterns (top). The clustermap clearly distinguishes a set of *Rider* presences unique to one or a few genomes (top) and a set of *Rider* absences unique to one or a few genomes (bottom). Numbers in parentheses represent branches discussed in the main text.

(from top to bottom in Figure S4): cv Katinka Cherry (EA00375), cv Sonato (LYC1969), and cv Momatero (TR00003), which are known to have large introgressions from *S. peruvianum* chromosome 9 carrying the Tomato Mosaic Virus resistance gene *Tm2-2* (Lanfermeijer *et al.*, 2005). The presence of these introgressions and their origin in *S. peruvianum* was confirmed by the similarity in SNP patterns from the relevant accessions in iBrowser (Figure S5b). Although we cannot independently establish the RT content of S. *peruvianum,* as no complete reference genome sequence is available for that species at this time, and thus the lack of detectable insertions in that species may have various causes, for *S. pennellii* many more insertions of at least three of the RTs are expected (see below).

5.3.4 Presence of the LTR-RTs in other *Solanaceous* species

To investigate the possible origin and distribution of the four retrotransposons in the Solanaceae, we searched the available genomes of the Solanaceae family species for the presence of full or partial copies of the RTs. Apart from *S. lycopersicum*, we detected full-sized copies of *Rider* in *S. pimpinellifolium,* as reported earlier (Xu and Du, 2014; Paz *et al.*, 2017), as well in *S. pennellii.* We found truncated copies of *Rider* (containing one of the LTRs) in *S. arcanum*, *S. galapagense* and *S. habrochaites* (Figure 5.4). In previous studies, *Rider* was not detected in potato, pepper or tobacco (Jiang *et al.*, 2009; Cheng *et al.*, 2009; Jiang *et al.*, 2012, Benoit et al., 2019). Like for *Rider*, intact copies of F50 and F120 were found in *S. pimpinellifolium* and *S. pennellii* and for the latter, also in *S. habrochaites*. Only truncated copies were found in the intermittent wild tomato species, but unlike *Rider* also in potato. Outside tomato, F17 full-length copies were only found in *S. pimpinellifolium*.

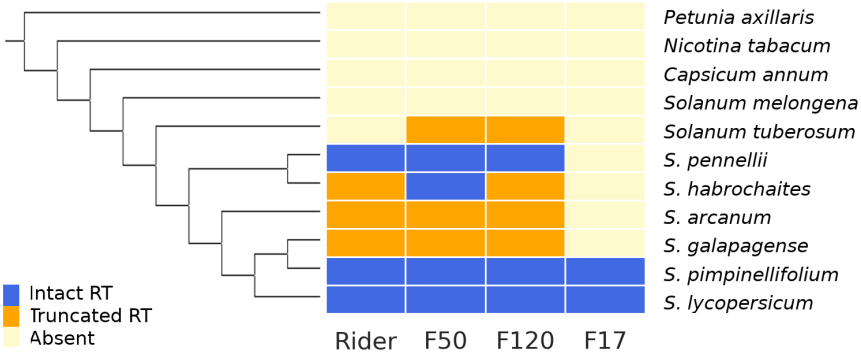

Figure 5.4 Absence (yellow) and presence of full length (intact, blue) or truncated (orange) copies of RT insertions in Solanaceae species together with an inferred phylogenetic tree of the Solanaceae family from (Aflitos *et al.*, 2014) and (Olmstead *et al.*, 2008). For all species we used the assembled draft genomes or reference genomes (Table S6) to infer the RT insertions.

5.3.5 LTR-RT insertions in genes

We compiled the results for all analysed tomato accessions into tables listing all shared and unique RT insertions. This includes the results of the analysis of an additional 14 accessions (RF201 and up, Table S5) that were resequenced at lower coverage but yielded sufficient insertions. The resulting tables are included in Data S1. When linking all detected insertion positions of the LTR-RTs to gene positions and annotations (Data S1), it became clear that five previously described *Rider* insertions in the tomato genome could be identified in some or most of the cultivated tomato accessions of this study (Table 5.3). These are the *sun* allele leading to elongated fruit types, the *c* mutation leading to potato-like leaves, the *r* mutant allele leading to yellow fruit colour, the *j-2* allele leading to jointless fruit pedicels, and an insertion in *ALMT9*, a gene involved in malate transport (referenced in Table 5.3). Some of these

Table 5.3 A selection of earlier described *Rider*-associated mutations with a known phenotype and of novel insertions of all four studied LTR-RTs in gene bodies (UTR, CDS, intron) of tomato genes with known, putative (), and unknown (-) functions. In the first group, the reference is to the description of the insertion, in the second group the reference is to the (putative) function of the targeted gene. Accessions printed in bold have been previously identified as having the insertion.

Previously identified LTR-RT insertions found in this study

Gene/mutation	Gene id	Trait	TE insertion	Accessions	Reference
DEFL1/sun	Solyc07g007760	fruit shape	*Rider* (cds)	RF19,22,**24**,26,33, 34,35,38,43,78	(Xiao *et al.*, 2008)
Bli-2/c	Solyc06g074910	leaf shape	*Rider* (cds)	RF19,34,89,90,**206**	(Busch *et al.*, 2011)
PSY1/r	Solyc03g031860	fruit colour	*Rider* (cds)	RF203	((Fray and Grierson, 1993)
ALMT9	Solyc06g072920	malate transport	*Rider* (2nd intron)	>50	(Ye *et al.*, 2017)
MBP21/j-2TE	Solyc12g038510	fruit abscission	*Rider* (1st intron)	RF**27**,**232**,233,234	(Soyk *et al.*, 2017; Roldan *et al.*, 2017)

Selection of newly identified LTR-RT insertions identified in this study

Gene/mutation	Gene id	(putative) trait	TE insertion	Accessions	Reference
ANL2b/HDG1	Solyc06g035940	(cuticle biosynthesis)	*Rider* (intron) *Rider* (3' UTR)	>50 RF42	(Lashbrooke *et al.*, 2015)
ABI3	Solyc06g083590	ABA signalling	*Rider* (cds)	RF37	(Gao *et al.*, 2013)
SP5G3	Solyc11g008650	(flowering time)	*Rider* (cds)	RF103,105	(Cao *et al.*, 2015)
MBP13	Solyc08g080100	-	*Rider* (1st and 5th intron)	RF16,44,45	(Hileman *et al.*, 2006)
MBP6	Solyc01g093960	parthenocarpy	*Rider* (3'UTR)	RF103	(Takisawa *et al.*, 2018)
MYB11	Solyc12g049300	-	*Rider* (cds)	RF91	(Fernandez-Moreno *et al.*, 2016)
TOMLOXC	Solyc01g006540	volatile biosynthesis	*Rider* (2nd intron)	RF20	(Chen *et al.*, 2004)
TPS39	Solyc10g005390	volatile biosynthesis	*Rider* (6th intron)	RF17,44	(Cao *et al.*, 2014)
HMGRCoAR	Solyc02g082260	mevalonate pathway	F17	RF103	(Narita and Gruissem, 1989)
BCAT2	Solyc07g021630	amino acid biosynthesis	F50	>40	(Maloney *et al.*, 2010)
CCoAOMT	Solyc02g093270	phenylpropanoid pathway	F108	RF54	(Lashbrooke *et al.*, 2016)

DEFL1: DEFENSIN-LIKE; PSY1/r: PHYTOENE SYNTHASE 1; mutant allele of R(Red); ALMT9: Al-ACTIVATED MALATE TRANSPORTER 9; MBP21/J-2: MADS-BOX PROTEIN 21/jointless-2; ANL2b/HDG1: ANTHOCYANINLESS 2B; HOMEODOMAIN GLABROUS 1; ABI3: ABSCISIC ACID-INSENSITIVE 3; SP5G3: SELF-PRUNING 5G3; MBP13; MADS-BOX PROTEIN 13; MBP6: MADS-BOX PROTEIN 6; MYB11: MYELOBLASTOMA 11; TOMLOXC: TOMATO LIPOXYGENASE C; TPS39: TERPENE SYNTHASE 39; HMGRCoAR: 3-HYDROXY-3-METHYLGLUTARYL CoA REDUCTASE; BCAT2: BRANCHED-CHAIN AMINO ACID AMINOTRANSFERASE 2; CCoAOMT: CAFFEYOYL-CoA O-METHYLTRANSFERASE.

accessions had been described earlier in literature as having these particular insertions (bold print in the "accessions" column). Moreover, both *Rider* (and, to a lesser extent, the other three LTR-RTs) showed as yet undescribed insertions into gene bodies with known or putative functions. A selection of the latter is listed in Table 5.3.

5.4 Discussion

We have surveyed tomato LTR-RT insertions and their polymorphisms in order to estimate the contribution of these polymorphisms to genetic diversity and to discover active RTs among cultivated tomato accessions and landraces, with comparisons to a small number of wild accessions.

We used ITIS as a tool to detect insertions based on read information. A low precision and recall were observed in ITIS results on human non-LTR retrotransposon data (Rishishwar *et al.*, 2017), but in a benchmark on Heinz we found it to be conservative for LTR-RTs using our settings, yielding a low number of false discoveries at a reasonably high number of false negatives. The false negative rate in more distant genomes is expected to be higher due to mismatches and structural variations between reference genome and the genome from which the reads originate. Due to the nature of LTR-RTs, it is challenging to detect insertions from short read data. Different tools have been developed for different problem settings (genomes, TE types) and will vary in performance when applied in other settings. In this study, we attempted to overcome some of the biases of ITIS by modifying it for performance on tomato LTR-RT detection. Even so, as ITIS still has a high false negative rate it cannot specifically distinguish insertions present in the reference genome but absent in the resequenced accessions. As it is nearly impossible to perform de novo assembly for the newly detected RT insertions of multiple copy RT families from short reads, we lack information on their length, but insertion polymorphisms are expected to be the result of relatively recent transpositions and therefore to have maintained full length.

Our first selection of 21 LTR-RTs for characterization was based on the relatively young age of insertions in the reference genome as reported earlier in other studies (Xu and Du, 2014; Paz *et al.*, 2017). This selection contains most (7) of the previously reported LTR-RTs studied before in tomato like *Rider, Jinling, ToRTL1, T135, Tnt1/ Retrolyc, TARE1* and *TGRE1* (Table 5.1). Additionally, many of these were shown to have insertions that were transcribed, as evidenced by their appearance in data from RNAseq experiments published earlier, another clue to their putative present activity (Paz *et al.*, 2017)

RT insertions are not uniformly distributed (Wang *et al.*, 2006). Most RTs accumulate in pericentromeric heterochromatin (Morata *et al.*, 2018; Contreras *et al.*, 2015) but some are found in gene-rich euchromatic regions. In tomato, the *Copia* superfamily preferentially targets genes, while the *Gypsy* superfamily is mostly found in heterochromatic regions. In total, RTs cover 62% of the tomato genome (Paz *et al.*, 2017; TheTomatoGenomeConsortium, 2012). Several examples of RT-related phenotypic variation in tomato were identified earlier, mostly related to fruit shape, colour and taste. Cultivated tomato and closely related red- and orange-fruited species accumulate hexoses in mature fruit, while green-fruited wild species accumulate sucrose. This difference appeared to be determined by the level of transcription of the soluble acid invertase gene *TIV*, due to a *Copia*-like RT insertion in the promoter of red-fruited species (Moy *et al.*, 2007). More recently it was shown that the relatively low volatile ester production of red-fruited *Solanum* species is caused by the activation of an esterase gene in these species, mediated by the insertion of a *Copia*-like retrotransposon in its promoter (Goulet *et al.*, 2012). Thus, TE mobility has clearly occurred during evolution of the tomato clade and related *Solanum* clades, and has contributed to changes related to fruit shape and quality, properties that to some extent

may have been subject to selection during domestication. *Rider*, a Ty1/*Copia*-family retrotransposon that appears to be tomato-specific (i.e. found in wild tomato relatives but not in tobacco and potato), showed most evidence for recent transposition, including in our present study. *Rider* is constitutively transcribed and several copies have identical LTR's and intact ORFs, all of which are hallmarks of active transposons (Cheng *et al.*, 2009).

Both ITIS and a Blast search against the assembled *S. pennellii* genome revealed a small number of *Rider* insertions, however at different locations that those found in the *S. lycopersicum* accessions. For the other three RTs there are considerable numbers of insertions shared between the two wild accessions and the cultivated tomato accessions (Figure 5.2). This may indicate considerable differences in the evolutionary history of *Rider* as opposed to the other three. The latter may have their origin in the last common ancestor of *S. lycopersicum/pimpinellifolium* and *S. pennellii,* and with a low transposition rate since their split could retain a large number of common insertions. In contrast *Rider* may have been introduced independently in the two lineages or, if present in the last common ancestor may have experienced independent transposition bursts, explaining the lack of common insertions.

5.4.1 Perspectives for breeding and functional genomics

Of the four LTR-RTs for which insertions were determined in all accessions, *Rider* had the largest rate of insertions in gene bodies (UTR, exon, intron, CDS, Figure 5.1) and many more near to gene bodies, where they may affect gene expression. Five previously reported insertions (references in Table 5.3) in tomato genes (and additionally, the ancestor of the *sun* locus, on chromosome 10) were found with varying frequency in our collection, as listed at the top of Table 5.3. Also in four out of five cases we found insertions in accessions in which these had been reported before, in addition to many new accessions. This clearly validates our approach for finding LTR-RT insertions in tomato accessions using resequencing data and for linking these to potentially new interesting phenotypes. The *Rider*-mediated translocation event originating in chromosome 10, giving rise to the insertion of *Rider* together with an IQD (IQ67-Domain) protein encoding gene (called *SUN1* on chromosome 10, (Huang *et al.*, 2013) in the *DEFL*1 gene on chromosome 7, was found in ten accessions. Curiously, only two of these (RF024, Jersey Devil, and RF026, Polish Joe) have retained the ancestral *Rider* copy near *SUN1* on chromosome 10, while another 15, including the reference genome, contained this ancestral copy without the translocation event on chromosome 7. Possibly the former has lost the ancestral copy during breeding, with the *Rider*-containing allele being replaced with one without the *Rider* insertion after the translocation to chromosome 10 had already occurred. Other previously characterized *Rider* insertions are *cut leaf* (*c*), and *yellow flesh* (*r,* mutant form of red); which was found in the pale yellow/ivory-fruited variety "Snowstorm" (RF203). *Al-ACTIVATED MALATE TRANSPORTER9* (*ALMT9*) was recently shown to underlie a Quantitative Trait Locus (QTL) for fruit malate content (*TFM6*) and was shown to contain a *Rider* (there called: CopiaSL_37) insertion in the second intron. The most recent example of an agronomic trait influenced by RT insertion is *jointless-2* in tomato. In this mutant, insertion of a *Rider* copy in the first intron of the MADS-box protein encoding gene *MBP21* led to decreased expression - presumably by *Rider*-induced epigenetic modifications - and hence to a disruption of abscission zone development in the fruit.

Insertions not reported before include a single homeobox-leucine zipper protein-encoding gene, ANL2b/HDG1, which was found to be targeted in two distinct events. Similarly, *MBP13*, a MADS-box transcription factor encoding gene with yet unknown function, has two *Rider* insertions in different introns, albeit all occurring in the same three accessions. A selection of insertions inside genes with a known function

and/or insertion phenotype (see above) or with a putative function based on homology is shown in Table 5.3. Obviously, the phenotypic effects of these new insertions will need to be further experimentally assessed. A complete list of all insertions is given in Data S1.

5.4.2 Conclusions and future perspectives

With our identification of sufficiently abundant and diverse, actively transposing (*Rider*) RTs, the way is also open for a more targeted approach to identifying insertion polymorphisms. This could be a combination of dedicated amplification of particular RT families with their adjacent genome sequence (Transposon Display; (Vandenbussche *et al.*, 2013)) and high-throughput sequencing. It would allow a much larger number of accessions to be screened. ITIS and similar protocols could also be applied to the much less-characterized, but ubiquitous tomato non-LTR retrotransposons and class II (DNA) transposons or MITES (Kuang *et al.*, 2009). This requires a more detailed characterization of the latter, as is now available only for LTR-RTs. On the other hand, much more detailed information about transposons present in tomato and their varying insertion positions and architecture will likely become available as high-quality long-read sequencing technologies become increasingly more available (Schmidt *et al.*, 2017). With average read lengths much larger than most TEs and an error rate equal to that of short reads (circa. 0.1 %), PacBio Hi-Fi offers a good candidate technology for detecting TE insertions and TIPs. Moreover, Oxford Nanopore Technologies (ONT) long-reads getting more accurate with new base callers, making it another candidate for detecting kb long insertions with high accuracy.

Activating transposable elements in crops is an attractive strategy for creating genotypic diversity and new traits for plant breeding (Paszkowski, 2015). Recognized early, various biotic and abiotic stresses activate mobility of transposons (McClintock, 1984). Alternatively, normally silent transposons may be activated in mutants that are compromised in their epigenetic modification of transposable elements (Tsukahara *et al.*, 2009). Transfer of an active transposon from another species is so far the only successful use of transposons in tomato (Meissner *et al.*, 2000). Curiously, activation of endogenous transposons and characterization of their (semi-) random insertions, such as those studied in more detail here can provide an attractive, non-genetic modification type of mutagenesis in tomato.

5.5 Experimental procedures

5.5.1 Sequence data

We downloaded sequencing data of all tomato cultivars and land accessions; two *S. pimpinellifolium* samples (known to be the closest wild relative to tomato); *S. peruvianum* (known to have introgressions into tomato cultivars) and *S. pennellii* as an outgroup. The paired-end sequences were downloaded from the European Nucleotide Archive (ENA, http://www.ebi.ac.uk/ena) under project number PRJEB5235 (Aflitos *et al.*, 2014). The sequence data for *S. lycopersicum* Heinz 1706 were obtained from Sandra Smit (Wageningen University). The sequence data for 14 additional accessions (RF201 and up) were produced as described before (Aflitos *et al.*, 2014) but with approximately 10x coverage and these were uploaded to the European Nucleotide Archive (project number PRJEB29506) The full list of tomato cultivars used in this study can be found in Table S5.

Briefly, the sequences were obtained by Illumina Hiseq 2000 sequencing (Aflitos *et al.*, 2014) and filtered with minimum base quality 20 and minimum sequence length 50 by Trimmomatic version 0.36 (Bolger *et al.*, 2014). The insert size of the paired-end reads was 500 bp. The read length was 100 bp.

5.5.2 Retrotransposon sequences

Most candidate RTs were selected from two studies, Xu and Du (2014) and Paz et al. (2017) where the authors listed RT insertions in the tomato genome together with an estimated time of insertion. We selected those RT insertions with an estimated age of less than 10000 years and found to be intact in the aforementioned studies. For each RT family, we used the sequence of the first copy listed in the genome as representative for that family. We thus obtained 17 RT families belonging to either the *Copia* or *Gypsy* superfamily. We also investigated previously studied RTs, like *Rider* (Cheng et al., 2009) and *Jinling* (Wang et al., 2006). The sequences were downloaded from NCBI under accession numbers EU195798.2 and DQ445619.1 for *Rider* and *Jinling*, respectively. Moreover, we included *TGRE1* (SL_RT_F2 in Xu and Du, 2014; Yin et al 2013) and *TARE1* as they are described as young in Yin et al (2013). The location of the sequences based on the tomato reference genome version 2.40 is given in Table S1. We only used reference genome version 2.40 to obtain the RT sequences, for all other analyses we used reference genome version 3.0.

5.5.3 Detecting RT insertions

To find RT insertions we used the ITIS algorithm (Jiang et al., 2015). Briefly, ITIS is a pipeline written in Perl, using different external software packages to find existing or new transposable element insertions in a given sequence data. The pipeline includes (1) finding all transposons present in the reference genome based on a given representative transposon sequence using BLASTN, and masking the reference genome where these copies are found; (2) whole genome alignment of short reads against the masked reference genome combined with the transposon sequence by BWA-mem, then processing the alignments to report a set of candidate insertions together with supporting reads; (3) filtering candidate insertions based on user-provided settings and categorizing the insertions as copy (i.e. nearby an existing copy) or not. As suggested by the ITIS developers, we used BWA version 0.7.7 (Li, 2013), blast+ 2.6.0 (Camacho *et al.*, 2009), samtools 0.1.19 (Li *et al.*, 2009), R 3.4.2 (R Core Team, 2018), bedtools 2.17 (Quinlan and Hall, 2010), python 2.7 (Python Software Foundation, http://www.python.org) and Bioperl 1.6.924 (Stajich *et al.*, 2002).

We made some minor changes in the ITIS pipeline to accommodate the needs of the current study. In the first part of the pipeline, to find the copies of RT in the reference genome, *S. lycopersicum* Heinz version 3.0, we relaxed the setting of the BLAST search from the default 28 bp word size to a 14 bp word size. This allows the pipeline to detect more distantly related RT copies. The second part of the pipeline was used as is; in the third part, we implemented some rules and settings as detailed below.

As a gap in the reference genome potentially includes an RT insertion, we added another category. Insertions in this category may not be new but rather a copy of an existing one not reported in the reference assembly due to a gap. We extracted assembly gaps (sequences of N's) from the reference genome using a custom script and used these to label insertions as near a gap when sufficiently close (similar to the "nearby copy"). As a result, there were three categories: "copy", a copy of an RT present in the reference genome; "gap", an RT insertion near a gap in the reference genome; and "new", an RT insertion not present in the reference genome and not near a gap.

ITIS filters candidate insertions by requiring that all of the user-provided thresholds regarding supporting reads have to be passed. The original thresholds are on the following measures.

I. The total number of supporting reads (t)

II. The number of *clipped* reads covering the start/end of the candidate insertion (CS/CE)

III. The number of *crossing* reads covering the start/end of the candidate insertion (cs/ce)

IV. The *total* number of reads covering the start/end of the candidate insertion (TS/TE)

We modified the filtering script of ITIS so that some combinations of thresholds will be required instead of all thresholds. The modified code is available on github https://github.com/sdemirci/ITIS. The minimum supporting read requirements were defined based on manual inspection to allow ITIS report insertions even if one end (either 5' or 3') of the insertions is not supported by any reads. The requirements for the copy and gap categories were t >= 3 and TS or TE >= 3, while we used rather stringent settings for new insertions. For the latter, the requirements were either t >=6 and TE = TS >=1 or t >= 30 and CS or CE >=1 and TS or TE >= 1 (to allow insertions that have at least one clipped read and supported by many reads if the insertion is supported by one side; otherwise, the insertion should be supported by both ends (5' and 3') by at least 6 reads in total).

For all categories (copy, gap, new), the minimum mapping quality of reads was set to 30. Insertions were labelled as copy resp. gap if the distance between the insertion and the RT resp. gap in the reference was less than 400 bp, i.e. library insert size minus the read length. Insertions were labelled as new if the distance between an insertion and the nearest reference copy or gap is more than 1000 bp, i.e. twice the insert size, to make sure the insertion is unique with respect to the reference genome.

To find the copies of insertions in the reference genome, we merged BLAST hits, which were found by the modified ITIS pipeline step 1, if they were less than 100 bp apart from each other using the bedtools v2.25 merge tool (Quinlan and Hall, 2010). The resulting region was labelled as "copy". Where there was BLAST evidence in the reference that a transposon is present at this location in the reference, but the sequence has likely not been assembled (resulting in a gap), and the BLAST hit was less than 100 bp from an assembly gap (NN's) we merged the gap and BLAST hit and labelled the resulting region as "copy,gap".

Since ITIS reports the start/end position of the supporting reads which could be different for each studied genome (sample), we replaced the positions of the reported region with the positions of the nearest reference genome "copy". In this way the positions of a copy would be the same for all genomes and comparable to each other for polymorphism analysis. We did not filter insertions based on the depth of all reads to be able to report more insertions. As a result, for some insertions we cannot determine whether they are homozygous or heterozygous insertions. Since our study doesn't depend on whether the insertions are homo/heterozygous, the filtering does not have any negative impact on our study.

5.5.4 Benchmarking ITIS on *S. lycopersicum* cv Heinz

To measure how well ITIS method identifies LTR-RT insertions from short read data, we benchmarked the reported ITIS calls based on LTR-RT insertions found in the reference genome, *S. lycopersicum* cv Heinz. The copy insertions for each LTR-RT family are found by RepeatMasker version 4.1 (Smit et al, 2013-2015). As the alignment algorithm in RepeatMasker we used modified blast for Repeatmasker, rmblast version 2.10. As a library, we used all reference TE sequences to search similar sequences

in the reference genome, version SL3.0. The results of the Repeatmasker run were filtered based on sequence identity; sequences that are at least 80% identical to the reference LTR-RT sequences are retained and merged (extended) if they are less than 100 bp apart from each other for each LTR-RT family match. This set is regarded as the ground truth for that LTR-RT. We then classified each ITIS call as one of:

TP: True positive, an ITIS call that overlaps with a ground truth call

FP: False positive, an ITIS call that does not overlap with a ground truth call

FN: False negative, an insertion found in the ground truth that does not overlap any ITIS call, i.e. an insertion missed by ITIS

FDR: FP / (TP + FP), number of FP divided by total number of ITIS calls

FNR: FN / (TP + FN), i.e. miss rate, number of misses divided by the sum of TP and FN. Note that the TP and FN generally sum to the number of positives in the ground truth; however, in this case, it might not. This difference is due to the overlap counts. Since a sequence can overlap multiple other sequences - for instance, a full copy ITIS call can overlap with two solo LTRs in the ground truth - the number of TP can be lower than the number of overlaps in the ground truth (see Figure S1).

We calculated the FDR and FNR for ITIS calls labelled as (1) `copy' and `copy,gap' calls, and (2) only `copy' calls. We also measured these rates for a subset of ITIS calls by using a subset of ground truth where - in both subsets - the sizes of LTR-RT insertions were at least 80% of the reference LTR-RT length. For the merged LTR-RT calls, the sizes of merged calls were at least 80% of the shortest reference LTR-RT.

5.5.5 Annotation of RT insertions

We annotated all RT insertions as CDS, UTR or intron if they interrupt one of these gene elements. The insertions were annotated as intergenic if the insertion does not overlap with genes. In the latter case, the distance from the insertion to the closest gene is calculated. We compared the locations of RT insertions with the gene models of reference genome annotation (ITAG) version 3.2, downloaded from https://solgenomics.net. We also annotated the insertions whether they fall in euchromatin or heterochromatin. We re-calculated the borders of eu/heterochromatin for tomato reference genome version 3.0 according to method defined in (Demirci *et al.*, 2017).

5.5.6 Finding similarity between RT sequences

The limited insert size of the NGS data limits our ability to distinguish between transposons that are highly similar in the start and end regions. To quantify this, we aligned the 500 bp from both ends of each pair of transposon sequences and their reverse complements (pairwise global alignment, calculated using the pairwise2 function Biopython version 1.72. The maximum score of these pairwise comparisons was then used a similarity score:

$$\text{Similarity}(Ti,Tj) = \text{maximum}\left(s(Ui, Uj),\ s(rc(Di), Uj),\ s(Di, Dj),\ s(rc(Ui), Dj) \right)$$

where Ti and Tj are two transposons, U is the upstream (5') end of a transposon, D is the downstream (3') end of a transposon, *rc* is the reverse complement function and *s* is the percent identity function.

5.5.7 Evolutionary relationships

We constructed a cluster map as a combination of a heat map and two trees, one based on clustering based on absence/presence of RT insertions, the other an evolutionary tree based on whole-genome SNP information. We used Euclidean distance

to cluster the insertions via the clustermap function in the seaborn module version 0.9.0 (Waskom *et al.*, 2017) of Python 3.

The Newick-formatted tree for the evolutionary relations between tomato species and cultivars was obtained from Aflitos et al.'s study (2014) by personal communication. We pruned the tree to summarize the relations between the 60 tomato genomes used in this study using the ETE3 (Huerta-Cepas *et al.*, 2016) module in Python 3. To compare the clustering based dendrogram to the whole genome SNP-based dendrogram, we used the compare function in the ETE3 module, which uses the Robinson-Foulds metric to calculate the differences between two trees based on their topology.

5.5.8 RT copies in Solanaceae

The sources of genomes of Solanaceae species are given in Table S6. We searched for sequences similar to RT in the Solanaceae genomes by BLASTN (Camacho *et al.*, 2009) with a word size of 14 bp (the same setting used our modified ITIS pipeline), minimum percent identity 80 and otherwise default settings.

5.5.9 Data availability

All tomato (Solyc) gene sequences can be retrieved from the SOL Genomics Network website (https://solgenomics.net). Sources of RT family sequences are listed in Table S1. Sources of Solanaceae genomes are listed in Table S6.

5.6 Acknowledgements

The work presented here is supported by the EU FP7 COMREC Marie Curie Initial Training Networks Programme project number 606956. We thank Wytze Gelderloos for his contribution in the ITIS pipeline testing and modification, Guangnan Chen for Figure S5 and Rens Holmer and Baojian Chen for commenting on the manuscript. We thank Dr. Richard Finkers (Plant Breeding, Wageningen University) for resequencing data of tomato accessions deposited under project number PRJEB29506, which was funded by Dutch Topsector project TKI EZ-2012-19 and the breeding companies Bejo seeds, Semillas Fito and BHN seeds.

5.7 Supporting Information

Supplementary files are available at Zenodo, http://doi.org/10.5281/zenodo.4939935

5.8 References

Aflitos, S., Schijlen, E., Jong, H. de, et al. (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.*, **80**, 136–148.

Aflitos, S.A., Sanchez-Perez, G., Ridder, D. de, Fransz, P., Schranz, M.E., Jong, H. de and Peters, S.A. (2015) Introgression browser: high-throughput whole-genome SNP visualization. *Plant J.*, **82**, 174–182.

Bolger, A.M., Lohse, M. and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics*, **30**, 2114–2120.

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.

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K. and Madden, T.L. (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Cao, K., Cui, L., Zhou, X., Ye, L., Zou, Z. and Deng, S. (2015) Four Tomato FLOWERING LOCUS T-Like Proteins Act Antagonistically to Regulate Floral Initiation. *Front. Plant Sci.*, **6**, 1213.

Cao, Y., Hu, S., Dai, Q. and Liu, Y. (2014) Tomato terpene synthases TPS5 and TPS39 account for a monoterpene linalool production in tomato fruits. *Biotechnol. Lett.*, **36**, 1717–1725.

Chen, G., Hackett, R., Walker, D., Taylor, A., Lin, Z. and Grierson, D. (2004) Identification of a specific isoform of tomato lipoxygenase (TomloxC) involved in the generation of fatty acid-derived flavor compounds. *Plant Physiol.*, **136**, 2641–2651.

Cheng, X., Zhang, D., Cheng, Z., Keller, B. and Ling, H.Q. (2009) A new family of Ty1-copia-like retrotransposons originated in the tomato genome by a recent horizontal transfer event. *Genetics*, **181**, 1183–1193.

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.

Contreras, B., Vives, C., Castells, R. and Casacuberta, J.M. (2015) The Impact of Transposable Elements in the Evolution of Plant Genomes: From Selfish Elements to Key Players. In P. Pontarotti, ed. *Evolutionary Biology: Biodiversification from Genotype to Phenotype*. Cham: Springer International Publishing, pp. 93–105.

Daraselia, N.D., Tarchevskaya, S. and Narita, J.O. (1996) The promoter for tomato 3-hydroxy-3-methylglutaryl coenzyme a reductase gene 2 has unusual regulatory elements that direct high-level expression. *Plant Physiol.*, **112**, 727–733.

Demirci, S., Dijk, A.D.J.J. van, Sanchez Perez, G., Aflitos, S.A., Ridder, D. de and Peters, S.A. (2017) Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between Solanum lycopersicum and Solanum pimpinellifolium. *Plant J.*, **89**, 554–564.

Ewing, A.D. (2015) Transposable element detection from whole genome sequence data. Mob. DNA, 6, 24.

Fernandez-Moreno, J.-P., Tzfadia, O., Forment, J., Presa, S., Rogachev, I., Meir, S., Orzaez, D., Aharoni, A. and Granell, A. (2016) Characterization of a New Pink-Fruited Tomato Mutant Results in the Identification of a Null Allele of the SlMYB12 Transcription Factor. *Plant Physiol.*, **171**, 1821–1836.

Fray, R.G. and Grierson, D. (1993) Molecular genetics of tomato fruit ripening. *Trends Genet.*, **9**, 438–443.

Galindo-González, L., Mhiri, C., Deyholos, M.K. and Grandbastien, M.-A. (2017) LTR-retrotransposons in plants: Engines of evolution. *Gene*, **626**, 14–25.

Gao, Y., Liu, J., Zhang, Z., Sun, X., Zhang, N., Fan, J., Niu, X., Xiao, F. and Liu, Y. (2013) Functional characterization of two alternatively spliced transcripts of tomato ABSCISIC ACID INSENSITIVE3 (ABI3) gene. *Plant Mol. Biol.*, **82**, 131–145.

Goulet, C., Mageroy, M.H., Lam, N.B., Floystad, A., Tieman, D.M. and Klee, H.J. (2012) Role of an esterase in flavor volatile variation within the tomato clade. *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 19009–19014.

Grzebelus, D. (2018) The functional impact of transposable elements on the diversity of plant genomes. *Diversity*, **10**, 18.

Hileman, L.C., Sundstrom, J.F., Litt, A., Chen, M., Shumba, T. and Irish, V.F. (2006) Molecular and phylogenetic analyses of the MADS-box gene family in tomato. *Mol. Biol. Evol.*, **23**, 2245–2258.

Huang, C.R.L., Burns, K.H. and Boeke, J.D. (2012) Active transposition in genomes. *Annu. Rev. Genet.*, **46**, 651–675.

Huang, Z., Houten, J. Van, Gonzalez, G., Xiao, H. and Knaap, E. Van Der (2013) Genome-wide identification, phylogeny and expression analysis of SUN, OFP and YABBY gene family in tomato. *Mol. Genet. Genomics*, **288**, 111–129.

Huerta-Cepas, J., Serra, F. and Bork, P. (2016) ETE 3: Reconstruction, Analysis, and Visualization of Phylogenomic Data. *Mol. Biol. Evol.*, **33**, 1635–1638.

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.

Jiang, C., Chen, C., Huang, Z., Liu, R. and Verdier, J. (2015) ITIS, a bioinformatics tool for accurate identification of transposon insertion sites using next-generation sequencing data. *BMC Bioinformatics*, **16**, 72.

Jiang, N., Gao, D., Xiao, H. and Knaap, E. Van Der (2009) Genome organization of the tomato sun locus and characterization of the unusual retrotransposon Rider. *Plant J.*, **60**, 181–193.

Jiang, N., Visa, S., Wu, S. and Knaap, E. Van Der (2012) Rider transposon insertion and phenotypic change in tomato. In M. A. Grandbastien and J. . Casacuberta, eds. *Plant Transposable Elements.Topics in Current Genetics*. pp. 297–312.

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

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P.B., Ouyang, S., Jiang, J., Robin Buell, C. and Baker, B. (2009) Identification of miniature inverted-repeat transposable elements (MITEs) and biogenesis of their siRNAs in the Solanaceae: New functional implications for MITEs. *Genome Res.*, **19**, 42–56.

Lanfermeijer, F.C., Warmink, J. and Hille, J. (2005) The products of the broken Tm-2 and the durable Tm-22 resistance genes from tomato differ in four amino acids. *J. Exp. Bot.*, **56**, 2925–2933.

Lashbrooke, J., Adato, A., Lotan, O., et al. (2015) The Tomato MIXTA-Like Transcription Factor Coordinates Fruit Epidermis Conical Cell Development and Cuticular Lipid Biosynthesis and Assembly. *Plant Physiol.*, **169**, 2553–71.

Lashbrooke, J., Cohen, H., Levy-Samocha, D., et al. (2016) MYB107 and MYB9 Homologs Regulate Suberin Deposition in Angiosperms. *Plant Cell*, **28**, 2097–2116.

Li, H. (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. , arXiv:1303.3997 [q-GN].

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R. (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

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

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.

Makałowski, W., Gotea, V., Pande, A. and Makałowska, I. (2019) Transposable elements: Classification, identification, and their use as a tool for comparative genomics. In M. Anisimova, ed. Evolutionary Genomics. Methods in Molecular Biology, vol 1910. New York, NY: Humana, pp. 177–207.

Maloney, G.S., Kochevenko, A., Tieman, D.M., Tohge, T., Krieger, U., Zamir, D., Taylor, M.G., Fernie, A.R. and Klee, H.J. (2010) Characterization of the Branched-Chain Amino Acid Aminotransferase Enzyme Family in Tomato. *Plant Physiol.*, **153**, 925–936.

McClintock, B. (1984) The Significance of Responses of the Genome to Challenge. *Science*, **226**, 792–801.

Meissner, R., Chague, V., Zhu, Q.H., Emmanuel, E., Elkind, Y. and Levy, A.A. (2000) A high throughput system for transposon tagging and promoter trapping in tomato. *Plant J.*, **22**, 265–274.

Morata, J., Tormo, M., Alexiou, K.G., Vives, C., Ramos-Onsins, S.E., Garcia-Mas, J. and Casacuberta, J.M. (2018) The Evolutionary Consequences of Transposon-Related Pericentromer Expansion in Melon. *Genome Biol. Evol.*, **10**, 1584–1595.

Moy, M., Dai, N., Cohen, S., Hadas, R., Granot, D., Petrikov, M., Yeselson, Y., Shen, S. and Schaffer, A.A. (2007) The presence of a retrotransposon in the promoter region of the TIV gene encoding for soluble acid invertase distinguishes between the sucrose and hexose accumulating species of Lycopersicon. *Acta Hortic.*, **745**, 429–436.

Narita, J.O. and Gruissem, W. (1989) Tomato Hydroxymethylglutaryl-CoA Reductase Is Required Early in Fruit Development but Not during Ripening. *Plant Cell*, **1**, 181.

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.

Olmstead, R.G., Bohs, L., Migid, H.A., Santiago-Valentin, E., Garcia, V.F. and Collier, S.M. (2008) A molecular phylogeny of the Solanaceae. *Taxon*, **57**, 1159–1181.

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**, 533–7.

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

Paz, R.C., Kozaczek, M.E., Rosli, H.G., Andino, N.P. and Sanchez-Puerta, M.V. (2017) Diversity, distribution and dynamics of full-length Copia and Gypsy LTR retroelements in Solanum lycopersicum. *Genetica*, **145**, 417–430.

Quinlan, A.R. and Hall, I.M. (2010) BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, **26**, 841–842.

Rishishwar, L., Wang, L., Clayton, E.A., Mariño-Ramírez, L., McDonald, J.F. and Jordan, I.K. (2017) Population and clinical genetics of human transposable elements in the (post) genomic era. *Mob. Genet. Elements*, **7**, 1–20.

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.

Schmidt, M.H., Vogel, A., Denton, A.K., et al. (2017) De Novo Assembly of a New Solanum pennellii Accession Using Nanopore Sequencing. *Plant Cell*, **29**, 2336–2348.

Smit, A.F.A., Hubley, R. and Green, P. RepeatMasker Open-4.0. 2013-2015 <http://www.repeatmasker.org>

Soyk, S., Lemmon, Z.H., Oved, M., et al. (2017) Bypassing Negative Epistasis on Yield in Tomato Imposed by a Domestication Gene. *Cell*, 1–14.

Stajich, J.E., Block, D., Boulez, K., et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. *Genome Res.*, **12**, 1611–8.

Takisawa, R., Nakazaki, T., Nunome, T., Fukuoka, H., Kataoka, K., Saito, H., Habu, T. and Kitajima, A. (2018) The parthenocarpic gene Pat-k is generated by a natural mutation of SlAGL6 affecting fruit development in tomato (Solanum lycopersicum L.). *BMC Plant Biol.*, **18**, 72.

Tam, S.M., Mhiri, C., Vogelaar, A., Kerkveld, M., Pearce, S.R. and Grandbastien, M.-A. (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. *Theor. Appl. Genet.*, **110**, 819–831.

TheTomatoGenomeConsortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635–641.

Tsukahara, S., Kobayashi, A., Kawabe, A., Mathieu, O., Miura, A. and Kakutani, T. (2009) Bursts of retrotransposition reproduced in Arabidopsis. *Nature*, **461**, 423–6.

Vandenbussche, M., Zethof, J. and Gerats, T. (2013) Massive indexed parallel identification of transposon flanking sequences. *Methods Mol. Biol.*, **1057**, 251–264.

Wang, Y., Tang, X., Cheng, Z., Mueller, L., Giovannoni, J. and Tanksley, S.D. (2006) Euchromatin and pericentromeric heterochromatin: Comparative composition in the tomato genome. *Genetics*, **172**, 2529–2540.

Waskom, M., Botvinnik, O., O'Kane, D., et al. (2017) mwaskom/seaborn: v0.8.1 (September 2017).

Wicker, T., Sabot, F., Hua-Van, A., et al. (2007) A unified classification system for eukaryotic transposable elements should reflect their phylogeny. *Nat. Rev. Genet.*, **8**, 973–982.

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, Y. and Du, J. (2014) Young but not relatively old retrotransposons are preferentially located in gene-rich euchromatic regions in tomato (Solanum lycopersicum) plants. *Plant J.*, **80**, 582–591.

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.

Ye, J., Wang, X., Hu, T., et al. (2017) An InDel in the Promoter of Al-ACTIVATED MALATE TRANSPORTER9 Selected during Tomato Domestication Determines Fruit Malate Contents and Aluminum Tolerance. *Plant Cell*, **29**, 2249–2268.

Yin, H., Liu, J., Xu, Y., Liu, X., Zhang, S., Ma, J. and Du, J. (2013) TARE1, a Mutated Copia-Like LTR Retrotransposon Followed by Recent Massive Amplification in Tomato. *PLoS One*, **8**, e68587.

Chapter 6

Discussion

As genomes constantly change during natural evolution and human directed evolution, the sources and consequences of genetic diversity are interconnected and influence each other. In this thesis, I investigated both these sources and consequences from a genome organization point of view, such as rearrangements due to recombination and a wide range of structural variations, including transposons. This thesis contributes to the fundamental knowledge of how recombination works in plants as a source of genetic diversity and sheds light on the genetic composition of key crops and their wild relatives.

In the sections below, recombination, structural variation and transposons are discussed in terms of the challenges encountered during the project, the current advances and future perspectives of the field. Afterwards, issues are discussed, such as the need for a high-quality reference genome and the ability of sequencing technologies to detect genome rearrangements. Finally, the potential application of our findings in plant breeding are discussed.

6.1 Recombination

6.1.1 Challenges in detecting recombination

In chapter 2, we have detected recombination in a tomato population. During the setup of the experiment to generate recombination breakpoints, recombinant inbred lines were used, as these contain mosaics of parental genomes which are clearly distinguishable from each other (Broman, 2005). In order to make the hybrid lines homozygous, they were self-crossed several times, which eventually fixed the haplotypes. Crossovers in homozygous domains that might have occurred during self-crossing cannot be detected due to lack of sequence variation, possibly causing an underrepresentation of recombination frequency. Hence, it is difficult to tell whether the lower recombination frequency that we observed for interspecific hybrids is due to biological issues, such as recombination machineries that work slightly different between species, or gene recombinations that are detrimental for hybrid offspring (Tepfer *et al.*, 2020), or alternatively due to the lack of sequence diversity, preventing detection of COs. One solution could be to use F2 offspring and include the heterozygous regions to represent the recombination frequency more accurately.

Another biological issue in the representation of recombination is the choice of using *S. pimpinellifolium X S. lycopersicum* to study interspecific recombination. We chose *S. pimpinellifolium* because it is easy to cross with *S. lycopersicum*, unlike *S. lycopersicoides* which shows linkage drag (Canady *et al.*, 2006). Nevertheless, this raises the question whether we indeed study interspecific recombination rather than intraspecific recombination., The ease with which crosses can be made suggest the latter; this could be due to heavy introgression of *S. pimpinellifolium* alleles into *S. lycopersicum* in the past (Wang *et al.*, 2020). As a result, *S. pimpinellifolium* possibly behaved more like a landrace or feral tomato during chromosome pairing with domesticated tomato, possibly resulting in more homologous rather than homeologous recombination. Although *S. pimpinellifolium* may not be the best candidate to study the limitations of homeologous recombination, considering its excellent crossability with tomato, it is a good candidate to be used as a bridging species - that is, an intermediate species between two divergent lines - between elite tomato lines and wild relatives of tomato which are more difficult to cross directly.

Besides the biological component, there is also a bioinformatics factor influencing recombination detection. Following the sequencing of parental lines and inbred lines, use of a single coordinate reference system to detect the recombination breakpoints is common practice. Then, depending on the quality of the reference and the variation between the reference and the parental genomes, several thresholds on the num-

ber of discriminative markers and filters on candidate breakpoints are applied. These thresholds and filters aim to eliminate false detection due to sequencing errors and read alignment artifacts. Since there is no standard for these filtering steps, different studies tend to apply slightly different thresholds on parameters even for the same dataset resulting in different numbers of COs as seen in two studies using the same set of tomato RIL data (chapter 2, de Haas *et al.*, 2017). To determine the recombination breakpoint, in chapter 2, we applied a threshold of 80% marker assignment to either parent in a window of 200 markers, while de Haas *et al.* used a similar approach, but applied a different threshold and window size. Beside differences in thresholds on the markers, one study used filters while the other did not use any filters on the candidate breakpoints to correct for read alignment errors resulting from an incomplete reference assembly, such as assembly gaps and/or SVs between the reference and the recombinant genome. In chapter 2, we filtered breakpoints near assembly gaps or in regions with copy number variation (CNV), while de Haas *et al.* filtered breakpoints overlapping with SVs between the reference used in mapping the reads and the newest available reference assembly. Another example of different filtering approaches is seen in detection of gene conversion events in Arabidopsis (Qi *et al.*, 2014). Due to the presence of transposable elements and CNV between reference and parental genome, false SNPs were called which led to false identification of gene conversions in studies that did not filter SNPs for the resulting read alignment artifacts. In summary, common challenges such as read alignment artifacts and sequencing errors can be solved by adjusting methods based on variation between reference and the parents, whereas study-specific challenges such as the limited number of available markers have to be solved in an *ad hoc* manner, usually by adjusting the window size and the number of reliable markers to take into account in that window. Failing to do the former and/or making different choices on the latter one results in different numbers of recombination. Taking such differences into consideration, caution should be taken when comparing studies on the same or similar data sets, focussing on global trends rather than overinterpreting comparisons of results from different studies.

6.1.2 Recent findings on CO associated features and advances in detecting recombination

After the publication of our work on crossovers (chapter 2-3), more evidence on the association of open chromatin and crossovers has surfaced. In *Arabidopsis*, COs are highly associated with open chromatin as CO sites overlap with DNase I hypersensitivity sites and ATAC-Seq sites, where both assays are indicators of accessible DNA (Rowan *et al.*, 2019). Similarly, open chromatin was also shown to be associated with CO in potato (Marand *et al.*, 2017), in rice via DNase HS (Marand *et al.*, 2019) and in maize via nucleosome occupancy assays (He *et al.*, 2017). Another feature associated with CO is sequence diversity between pairing hom(e)ologous chromosomes. SNP diversity is positively associated with CO in small regions up to a few kb in size in *Arabidopsis* (Blackwell *et al.*, 2020), while SVs which span larger regions are negatively associated with CO; COs were found to be suppressed inside inversions and transpositions in *Arabidopsis* (Rowan *et al.*, 2019). These new findings confirm and support our findings on the correlation of sequence diversity and open chromatin with COs.

Besides open chromatin, COs are associated with transposons, but in a rather conditional manner, depending on the location of the transposons. Transposable element (TE) content in euchromatin is positively correlated (Tc1/Pogo/Mariner and Helitrons in *Arabidopsis;* Mu/MuDR in maize; *Stowaway* in rice and potato) and pericentromeric transposons (En/Spm, Gypsy and Copia in *Arabidopsis*) are negatively correlated with COs (reviewed in Underwood & Choi, 2019). Interestingly, DNA methylation in CHH and CHG context were enriched in crossover hotspots in rice together with H3K4me3,

H3K9ac, H4K12ac, and H3K27me3 histone modifications, while DNA methylation in CpG context was not enriched in COs (Marand *et al.*, 2019). In chapter 3, we found that some classes of transposons are negatively correlated with COs while others, depending on the species, are positively correlated. The argument that TE correlation with CO depends on location could explain our findings of species dependence, as TEs are mobile and can be found throughout the genome (although certain classes of TE have a tendency to localize in specific regions (see section 1.1.3), for instance LTR retrotransposons are generally found on pericentromeric heterochromatin in various species, thus negatively correlated with COs in many species (chapter 3)). Similar to TE, DNA methylation also has a conditional correlation with CO; there may thus be more features that are not unequivocally correlated with COs.

Apart from new features, there are also new developments in the detection of recombination. Until 2019, the approaches to detect CO were not suited for genome-wide, high-throughput CO analyses in parallel. A new approach has been developed to detect CO breakpoints in directly pooled pollen data using linked-read technology (Sun *et al.*, 2019). This approach allows to screen thousands of fragments (of approximately 50 Kb) coming from individual pollen in parallel, so that recombinations can be identified at a high resolution. With this approach, a fine map of recombination has been revealed in *Arabidopsis* (Sun *et al.*, 2019) and tomato (Fuentes *et al.*, 2019). Moreover, an image processing technique applied on the tetrads was developed to accelerate the analyses of CO frequency as well as CO interference (Lim *et al.*, 2020). The whole-mount Immuno-FISH technique was adjusted to be applied on *Arabidopsis* meiocytes, which allows localization of meiotic proteins in their natural spatial organization without the need for chromatin spreading (Sims *et al.*, 2020). These different methods to detect CO will benefit meiotic recombination research by allowing to generate large-scale CO data.

6.1.3 Future perspectives

Over the past decade, we have accumulated remarkable knowledge on the genomic and genic features associated with CO (Wang *et al.*, 2019). However, this knowledge is mostly concentrated on the model organism *Arabidopsis (Rowan et al., 2019)* or a few crop species having large market value. Even for these crops, full data and analyses on the many aspects of CO are often not available. Moreover, some studies focus on initiation of CO via DSB and some others on non-CO via gene conversion. There are still major gaps in the big picture of CO occurrence in different plants. Thus, we should focus on closing these gaps by performing similar analyses in other crops to confirm earlier findings, as well as performing new analyses to clarify the influence of certain features such as different classes of transposons or different epigenetic modifications. For instance, DNA methylation was thought to be one of the features blocking recombination, but methylation in the CHH context is associated with crossovers (Marand *et al.*, 2019). To further extend our understanding, we should investigate DNA methylation and other epigenetic characteristics of CO in meiocytes, specifically in the pachytene stage of meiosis. To reveal the chromatin conformation in meiosis, single-cell Hi-C can be used; to uncover meiotic protein relations to the chromatin state, HiChIP - a protein-centric chromatin conformation method - (Mumbach *et al.*, 2016) seems promising. Altogether, research on these genomic features will help to further understand how CO localization works in plants.

Further advances are anticipated not only in the understanding of CO but also in its application in plant breeding. For instance, in introgression breeding, parental line selection for effective hybrid generation could be aided with recombination models where recombination locations and subsequently the genotypes of hybrids could be predicted. In chapter 3, we have built such a model with potential to predict

recombination in the next generation. This and similar models may be improved with more advanced machine learning algorithms (such as deep learning), and can be implemented with automatic parental selection based on the predicted genotypes. Another application in plant breeding is molecular manipulation of CO locations and frequency to obtain desired crops. It has already been shown that CO location and frequency can be altered by using certain manipulators such as site-directed nucleases, epigenetic modifiers or changing CO specific factors (Taagen *et al.*, 2020). Manipulating COs can bring many advantages in plant breeding (see section 6.5). Thus, it seems relevant for CO manipulation to be used more widely in future, although care must be taken due to rules and regulations on genetic manipulation in different parts of the world.

6.2 Structural variations

6.2.1 Challenges

In chapter 4, we used SVs to infer phylogenetic relationships among melon and wild relatives. We showed that SVs are sufficiently supportive to capture the same phylogenetic relations as SNPs and indels among sub-species and groups. However, these two types of variation give complementary information and therefore may be combined to obtain a more detailed overview of the total diversity of populations. However, combining these two types of variation is challenging due to different sequence lengths they affect. We need suitable algorithms and tools to combine these two genetic/genomic measures in a single analysis Until this issue is solved, analysing large and small variants separately is a proper choice.

Among the large variations, we were able to detect inversions, duplications and deletions. We could not detect large insertions, i.e. larger than the read length. However, insertions contribute to SV variation as well, so we are missing this type of variation in our analyses. Since we mapped short reads to a single reference genome, we miss the complex structures in large divergent regions (for instance Mb-scale insertions/deletions) especially in the landraces and wild related species of crops. The wild genomes possess more genetic variation and should ideally be *de novo* assembled so that re-arrangements can be better analysed.

6.2.2 Recent advances in SV diversity, effect and detection

In recent years, more population-wide SV data has accumulated. With the increased accessibility of long read (LR) technology, population wide studies started using this to support analyses. One example is the sequencing of 100 accessions out of a population of 700 tomatoes with LR to construct a set of diverse SVs i.e. panSV (Alonge *et al.*, 2020). Similarly, genetic diversity including PAV and CNV was revealed in almost three thousand deep sequenced soybean samples with the guidance of 26 selected representative soybean genomes (including 3 wild relatives of soybean) which were *de novo* assembled using LR, optical mapping and Hi-C (Yucheng Liu *et al.*, 2020).

Besides cataloguing more SV diversity, there are also studies revealing its connection with phenotypic variation. For instance, the effect of SV on the metabolic pathway of glyphosate tolerance (related to herbicide resistance) in two maize lines has been revealed by combining Illumina and PacBio data (Mahmoud *et al.*, 2020). In soybean, seed luster (an agronomic trait) was proposed to be associated with a 10kb PAV via a genome-wide association study (Yucheng Liu *et al.*, 2020). More associations between SVs and phenotypes could be revealed by linking gene expression to SV as done in human studies (Eteleeb *et al.*, 2020) and in tomato (Alonge *et al.*, 2020).

There are developments on the detection of SV via both short and long read technology. Even though LR technology is getting more accessible, for large scale population

studies it is more practical to do an SV screening with short reads followed by SV identification on selected lines with long reads (Sedlazeck *et al.*, 2018; Alonge *et al.*, 2020), since short reads offer affordable large-scale screening, but are not accurate enough to find novel SVs. In the meantime, new methods for SV detection based on LR are becoming available. Various SV detection tools based on LR have been benchmarked in human (De Coster *et al.*, 2019; De Coster and Broeckhoven, 2019) and pear (Yueyuan Liu *et al.*, 2020). They consistently report that using multiple tools is advantageous. Merging the results of a combination of methods based on custom thresholds via tools like SURVIVOR (Jeffares *et al.*, 2017) is currently the most popular way to do so. Given the vast amount of available short read datasets on plant populations, it has been shown feasible to combine the results with machine learning (Wijfjes *et al.*, 2019), which automatically merges multiple SV calls of different tools for the same accession and gives more reliable variations. In short, while new tools are being developed to detect SVs from LR data; for short reads, combining existing tools is popular and both technologies are being used as each has its own strength.

6.2.3 Future perspectives

Since the importance of SV in the study of evolutionary history and in plant breeding has been established, more SV is expected to be reported in major crops and their wild relatives. Especially with improving LR technology, i.e. higher throughput machines which can sequence many plant samples at low cost, we will see more plant genomes sequenced for i) genotyping with small and large variants ii) *de novo* assembly for more complete genomes and iii) generating pangenomes, a population representation of multiple genomes. It is also very likely that with increasing read accuracy or with better post-sequencing error correction (Karst *et al.*, 2021), we will be able to use a single technology to detect both large and small variants. In the short term, a candidate technology is PacBio HiFi, although the data generation step needs to be scaled up to meet the demands of population studies of large plant genomes.

When enough SV data is accumulated, we can start calculating population statistics on SVs, for example to estimate SV age and SV diversity. Also, we will need to define and generate new concepts to explain SV based diversity in populations as well as reuse existing concepts, for instance allele frequency. We can start calculating linkage disequilibrium between SV pairs instead of (or in combination with) SNPs, which can be used to learn more about the evolutionary dynamics of wild and crop (breeding) populations.

6.3 Transposons

6.3.1 Challenges in detecting TEs

In chapter 5, we provided an inventory of active retrotransposons in tomato cultivars, landraces and few wild related species. The project started in 2015 and we chose a tool which was appropriate for plants at that time. Better tools have been published since then (Kosugi *et al.*, 2019). Yet, these still cannot fully overcome the limitations imposed by the use of short reads, as long repetitive parts of TEs cannot be resolved. Since most TEs have a length in the range of 5-20kb, a high quality LR technology, such as Pacbio HiFi, would suit the job better. Moreover, TE sequences that considerably diverged from the initial TE copy would not be detected due to high nucleotide diversity. Therefore, finding old copies of TE is challenging and still an issue, even with LR.

Besides the length of TE, TE detection is very hard from only short reads due to the structure of TEs, combinations of genes and repetitive elements, and the mutations they experience over time. As with any other repetitive element, it is hard to *de novo*

assemble the TE from such reads. The only possibility is to map the reads to the reference genome and look whether the covered TE is present in the reference genome to infer PAV of that TE. This of course very much depends on the presence of TEs in the reference genome, which is an issue when a reference genome diverges from the target genome from which the reads were generated. Similar to SV detection in wild related species, the usage of crop genetic background reference genome is not sufficient to detect transposon insertion polymorphisms (TIPs) in large insertions, simply due to lack of anchor sequences in the reference. For TEs that are part of larger insertions, short or long reads may not suffice to anchor the TE to the reference genome sequence. Also, since wild relatives are genetically distant to crops, they potentially have more SVs and thus more such large insertions. This hampers detection of TIPs in wild species by mapping reads to crop reference genomes.

6.3.2 Recent advances in TE diversity, effect and detection

In line with the SV studies, the availability of small scale and large scale population-wide TIP data is increasing in major crops. In maize, 85% othe 2.3 Gb genome is composed of TEs, and 1.6 Gb of TE sequence was found to be polymorphic between four maize lines (Anderson *et al.*, 2019). The authors compared the published assemblies of these lines and then validated the TEs with available short reads. In another example, more than six hundred tomato accessions have been scanned for TIPs and over 300 TE families were reported to have TIPs (Domínguez *et al.*, 2020). Moreover, there are potentially many TE polymorphisms waiting to be detected in SV data, given reported associations of SV with repetitive elements in rice (Fuentes *et al.*, 2019) as well as in other plants. Thus, likely there is more TE data than explicitly reported in the literature waiting for us to be explored.

Among the reported TIPs studies, an interesting TIP-GWAS approach, which associates TIPs with a phenotype, uncovered TIP associations with in total five traits, including two previously known associations: fruit colour and leaf morphology (Domínguez *et al.*, 2020). The same approach also revealed an intronic TE insertion affecting tomato flavour. The authors also revealed that the expression of immune- and stress-responsive genes were affected by TIPs, by demonstrating an association between TIPs and gene expression (Domínguez *et al.*, 2020).

TIPs detection focuses on active TEs since it has more impact on phenotypes. As a start, a pilot study in *Arabidopsis* attempted to detect active TEs by using low coverage (<1x) Oxford Nanopore LRs (Debladis *et al.*, 2017). Even though LR technology is now available, applying it to a large number of samples is still costly. Therefore some studies still choose to use short reads and validate part of the results with LR. One such study found polymorphisms in three active TEs (*Hopi*, *Tos17*, and *Karma)* in 3000 rice accessions, based on mapping short reads using the TRACKPOSON tool (Carpentier *et al.*, 2019). The authors tested the performance of the tool with Oxford Nanopore LR on one rice accession for these three TEs. With the new tools, even the known associations of TEs to phenotypes can be understood in more detail with the new studies, as shown for the association of the Jointless2 gene and Rider TE family (Alonge *et al.*, 2020).

6.3.3 Future perspectives

In the near future we will be able to discover TIPs in many more plant species either by directly searching for TE activity or inferring them from SVs in general. The obvious next step then is to look for effects of TIPs on gene expression, as well as to associate TIPs with agricultural traits via TE-GWAS (Akakpo *et al.*, 2020). When directly searching for TE activity, TE detection is not trivial. Due to the unique structure of TEs, we need improved detection methods, either using LR or another technology

which can take into account the structure of different TE families, and mapping to alternative TE sequences of the same family rather than a consensus. With that, we could detect TE locations more accurately.

Accurate TE locations could give us clues on how TEs select new insertion locations in the genome, which could be advantageous for breeding strategies. This might help to develop targeted TE insertion strategies or to direct TE activities for plant breeding. It may also be worthwhile to investigate how to control TE activity via epigenetic regulators, or chromatin remodelers. Moreover, we should investigate to what extent recombination contributes to the dispersal of TEs and how DNA transposons open up areas for recombination to occur. Altogether, knowledge gained in these areas can help to improve breeding strategies for improving crops.

6.4 Overarching issues in detecting genome rearrangements

6.4.1 Need for good quality reference genomes

Since the first plant genome, that of *A. thaliana,* was published in 2000 (*Arabidopsis* Genome Initiative), 1,510 plant genome assemblies have been made available (https://www.ncbi.nlm.nih.gov/assembly) by March 31st, 2021. Among these, 377 are at contig level, 625 are scaffolded and 508 are complete at chromosome level. Contig/scaffold-level reference genomes are sub-optimal to provide a basis for large variant calling such as recombinations and SVs. For successful study of genome rearrangement, such assemblies should be improved using LR, Hi-C and/or optical mapping techniques (Michael and Van Buren, 2020). These improvements should focus on resolving assembly gaps and repetitive regions as well as centromeres to construct contiguous chromosomes. Also, genomes should be fully annotated for centromeres, genes, epigenetic and open chromatin status.

Once we are able to construct genomes properly, and in a high-throughput fashion, we can consider multi-reference genome(s) or pangenomes which potentially include wild related species. This is necessary since using a single reference genome is insufficient to identify variants from wild species or distant accessions. There have been attempts to generate such pangenomes from available short read-based resequencing data in tomato (Gao *et al.*, 2019), in *Brassica napus* (rapeseed) (Dolatabadian *et al.*, 2020) and from *de novo* assembled genomes in soybean (Yucheng Liu *et al.*, 2020). However, the pangenomes are better constructed from LR-based whole genome assemblies, as these have more contiguity than assemblies based on short reads and higher potential to represent large variations. Already several tools have been published to do this in bacteria (PGAweb, Chen *et al.*, 2018) and small plant genomes (PanTools, Sheikhizadeh *et al.*, 2016), but their applicability to larger genomes still has to be demonstrated.

Recently, the challenge of constructing pangenomes from larger genomes was picked up for human genomes. The Minigraph toolkit was released to combine multiple human genomes into a graph-based reference pangenome which can hold structural variations (Li *et al.*, 2020). The toolkit can also be used to map reads to a pangenome and detect variations. As the tool can handle the human genome, it should be usable on diploid plant genomes. Tools which can work with graphs are already available; for instance, structural variations can be genotyped in pangenome graphs using the vg toolkit (Hickey *et al.*, 2020). These developments show that the field is on its way to handle the construction of large (diploid or polyploid) plant genomes which then could be scaled up to hold a population.

Besides mapping complete variation among accessions via pangenomes, we also need to accurately map variation within accessions by separating the haplotypes of

chromosome sets. This refers to phasing genomes of diploids as well as polyploids (Zhang *et al.*, 2020). Most current reference genomes have a crop genetic background derived from inbred or double haploid lines. Since most plants (crop or otherwise) are heterozygous and polyploid, such references do not fully represent their actual genomes. In recent attempts to assemble wild and landrace heterozygous genomes, the haplotypes are collapsed with purging (i.e. removing duplicated content from the draft assemblies; Roach *et al.*, 2018) to obtain haploid representation of di/poly-ploid genomes. This results in mosaics of haplotypes in the reference genomes. However, in phased genomes, each copy of chromosomes will be represented fully and alleles on the same chromosome will be linked to each other (Cheng *et al.*, 2021). In phased genomes we can easily detect recombination breakpoints, where the setup used in this thesis, based on heterozygous alleles on both parents, cannot differentiate shared alleles on a marker. For example, in our current setup, the hybrid genotype A/G cannot be assigned as originating from one parent or as heterozygous when the parental genotypes are A/T and A/G. To overcome this limitation, we use a combination of markers to be sure that a certain region indeed originated from a single parent or not, at the cost of resolution. However, if all alleles were phased, all markers could be informative which would increase the resolution, thus quality, of recombination detection. In a similar way, phased genomes will facilitate genome comparison-based methods of detecting SV (including TEs), as the heterozygous alleles will not be lost as when using a consensus genome.

6.4.2 Need for improvements of sequencing technologies and algorithms

The nature of sequencing technology has a huge effect on the detection of all kinds of genomic variation, including rearrangements. As discussed above, the currently commonly used short read technology is insufficient when it comes to detecting large SV, especially large insertions from TEs and T-DNA. Moreover, due to the heavy dependence on short reads, variations accumulated in databases are biased towards short indels and SNPs, while LRs could provide variation at a larger scale (De Coster *et al.*, 2019). LR technology has already been introduced for SV studies in human and a few plant species. In the near future, more studies will switch to LR technology. The limitations of LR, such as high error rate and higher cost per sample, are being reduced at the time of writing this thesis. This will allow researchers to screen rearrangements in large populations. Besides SV, high accuracy LR (for instance PacBio HiFi) can provide haplotype-level resolution for CO breakpoints or translocations. Moreover, when applied on pollen pools (section 6.1.2), the location of COs can be detected in an efficient way for an entire population. If non-PCR amplified fragments are used in combination with a low-error sequencing technology, even the frequency of recombination can be obtained from pollen pools - since every read would represent an initial DNA fragment - as a follow up of linked-read based detection. This could be possible only for regions with high marker density i.e. if enough markers are present in the reads. Thus, longer reads than offered by the current PacBio HiFi technology, with similar or higher accuracy, are needed to fully resolve the rearrangements.

6.5 Application of genomic rearrangements in plant breeding

So far we have approached SV diversity and recombination in plants from a fundamental point of view. However, the knowledge also contributes to many areas of plant breeding. Plant breeding uses many techniques which depend on genetic diversity in natural populations. As this thesis reveals recombination distribution, recombination-associated features and genetic diversity at the genome structure level, it is worthwhile to consider how this information can help to direct the plant breeding.

First, the information we reveal on recombination-associated genomic features can contribute to manipulating recombination during plant crossing, aiding precision breeding and developing designer crops. Also, the prediction model that we constructed can aid the breeders to select which parental lines to cross. Second, our inventories of SVs in melon (chapter 4) and of TEs in tomato (chapter 5) provide insights into diversity within these populations which can be associated with phenotypes. Since specific SVs/TEs can affect the phenotype by disrupting or enhancing the expression of (trait related) genes, the inventory of such variation aids detection of gene associations with SV. These associations then might be used by plant breeders for manipulation of the phenotypes, by introducing or removing this diversity from crops. Finally, our methods to detect rearrangements are exemplary to other researchers detecting SV and TE in other crops/plants, contributing to revealing more plant diversity.

Given the direct relevance of recombination to breeding, the first point deserves more attention. As stated in section 6.1.2, certain genomic features (i.e. factors) were found to be associated with CO (Rowan *et al.*, 2019). Presence or absence of factors promoting CO such as open chromatin, SV depleted areas, certain TE classes, or factors blocking CO such as inversions in the crop lines can be used by breeders to estimate the probability of a successful cross. This can then be used to prevent lengthy crossing experiments. To aid the breeders in that direction, in fact, we have made a prediction model in chapter 3 which uses both positive and negative features to predict COs. With this model, breeders can estimate the possible outcomes for multiple pairs of candidate parental accessions and select the best matching plant accessions for their introgression breeding programs.

In addition to a screening approach such as outlined above, knowledge of CO associated genomic features creates an opportunity to develop methods for altering the CO location and/or frequency in hybrids during breeding. This way of alteration is advantageous over classical breeding, where massive numbers of crop lines are crossed to obtain desired products. There are several paths to follow in CO manipulation depending on the crop and trait of interest. One path could be the manipulation of the (epi)genome - such as reducing CO associated DNA methylation, opening chromatin - to allow more recombination to increase the genetic diversity in the hybrids (Underwood *et al.*, 2018). Another path could be via increasing controlled TE activity (Paszkowski *et al.*, 2015; Thieme *et al.*, 2017). As TEs can jump to different locations in the genome, by promoting this activity especially for CO prone TE, for instance by freeing them from epigenetic silencers, we can create new phenotypes and evaluate whether they have value for the market.

On the whole, the inventory of rearrangements and recombination-associated genomic features revealed in this thesis directly or indirectly contributes to plant breeding efforts.

6.6 References

Akakpo, R., Carpentier, M.C., Ie Hsing, Y. and Panaud, O. (2020) The impact of transposable elements on the structure, evolution and function of the rice genome. *New Phytol.*, **226**, 44–49.

Alonge, M., Wang, X., Benoit, M., *et al.* (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**, 145-161.e23.

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.

Blackwell, A.R., Dluzewska, J., Szymanska-Lejman, M., *et al.* (2020) MSH2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. *EMBO J.*, e104858.

Broman, K.W. (2005) The Genomes of Recombinant Inbred Lines. *Genetics*, **169**, 1133–1146.

Canady, M. a., Ji, Y. and Chetelat, R.T. (2006) Homeologous recombination in Solanum lycopersicoides introgression lines of cultivated tomato. *Genetics*, **174**, 1775–1788.

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

Chen, X., Zhang, Y., Zhang, Z., *et al.* (2018) PGAweb: A Web Server for Bacterial Pan-Genome Analysis. *Front. Microbiol.*, **9**.

Cheng, H., Concepcion, G.T., Feng, X., Zhang, H. and Li, H. (2021) Haplotype-resolved de novo assembly using phased assembly graphs with hifiasm. *Nat. Methods*, **18**, 170–175.

Coster, W. De and Broeckhoven, C. Van (2019) Newest Methods for Detecting Structural Variations. *Trends Biotechnol.*, **37**, 973–982.

Coster, W. De, Rijk, P. De, Roeck, A. De, Pooter, T. De, D'Hert, S., Strazisar, M., Sleegers, K. and Broeckhoven, C. Van (2019) Structural variants identified by Oxford Nanopore PromethION sequencing of the human genome. *Genome Res.*, **29**, 1178–1187.

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**, 537.

Dolatabadian, A., Bayer, P.E., Tirnaz, S., Hurgobin, B., Edwards, D. and Batley, J. (2020) Characterization of disease resistance genes in the Brassica napus pangenome reveals significant structural variation. *Plant Biotechnol. J.*, **18**, 969–982.

Domínguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jiménez-Gómez, J.M., Colot, V. and Quadrana, L. (2020) The impact of transposable elements on tomato diversity. *Nat. Commun.*, **11**.

Eteleeb, A.M., Quigley, D.A., Zhao, S.G., *et al.* (2020) SV-HotSpot: detection and visualization of hotspots targeted by structural variants associated with gene expression. *Sci. Rep.*, **10**, 15890.

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

Haas, L.S. de, Koopmans, R., Lelivelt, C.L.C., Ursem, R., Dirks, R. and Velikkakam James, G. (2017) Low-coverage resequencing detects meiotic recombination pattern and features in tomato RILs. *DNA Res.*, **24**, 549–558.

He, Y., Wang, M., Dukowic-Schulze, S., *et al.* (2017) Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proc. Natl. Acad. Sci.*, **114**, 12231–12236.

Hickey, G., Heller, D., Monlong, J., *et al.* (2020) Genotyping structural variants in pangenome graphs using the vg toolkit. *Genome Biol.*, **21**, 35.

Jeffares, D.C., Jolly, C., Hoti, M., *et al.* (2017) Transient structural variations have strong effects on quantitative traits and reproductive isolation in fission yeast. *Nat. Commun.*, **8**, 1–11.

Karst, S.M., Ziels, R.M., Kirkegaard, R.H., Sørensen, E.A., McDonald, D., Zhu, Q., Knight, R. and Albertsen, M. (2021) High-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *Nat. Methods*, **18**, 165–169.

Kosugi, S., Momozawa, Y., Liu, X., Terao, C., Kubo, M. and Kamatani, Y. (2019) Comprehensive evaluation of structural variation detection algorithms for whole genome sequencing. *Genome Biol.*, **20**, 8–11.

Li, H., Feng, X. and Chu, C. (2020) The design and construction of reference pangenome graphs with minigraph. *Genome Biol.*, **21**, 265.

Lim, E.C., Kim, Jaeil, Park, J., *et al.* (2020) DeepTetrad: high-throughput image analysis of meiotic tetrads by deep learning in Arabidopsis thaliana. *Plant J.*, **101**, 473–483.

Liu, Yucheng, Du, H., Li, P., *et al.* (2020) Pan-Genome of Wild and Cultivated Soybeans. *Cell*, **182**, 162-176. e13.

Liu, Yueyuan, Zhang, M., Sun, J., Chang, W., Sun, M., Zhang, S. and Wu, J. (2020) Comparison of multiple algorithms to reliably detect structural variants in pears. *BMC Genomics*, **21**, 61.

Mahmoud, M., Gracz-Bernaciak, J., Żywicki, M., Karłowski, W., Twardowski, T. and Tyczewska, A. (2020) Identification of Structural Variants in Two Novel Genomes of Maize Inbred Lines Possibly Related to Glyphosate Tolerance. *Plants*, **9**, 523.

Marand, A.P., Jansky, S.H., Zhao, H., *et al.* (2017) Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biol.*, **18**, 203.

Marand, A.P., Zhao, H., Zhang, W., Zeng, Z., Fang, C. and Jiang, J. (2019) Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *Plant Cell*, **31**, 645–662.

Michael, T.P. and VanBuren, R. (2020) Building near-complete plant genomes. *Curr. Opin. Plant Biol.*, **54**, 26–33.

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

Qi, J., Chen, Y., Copenhaver, G.P. and Ma, H. (2014) Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping. *Proc. Natl. Acad. Sci. U. S. A.*, **111**, 10007–12.

Roach, M.J., Schmidt, S.A. and Borneman, A.R. (2018) Purge Haplotigs: allelic contig reassignment for third-gen diploid genome assemblies. *BMC Bioinformatics*, **19**, 460.

Rommel Fuentes, R., Hesselink, T., Nieuwenhuis, R., *et al.* (2020) Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. *Plant J.*, **102**, 480–492.

Rowan, B.A., Heavens, D., Feuerborn, T.R., Tock, A.J., Henderson, I.R. and Weigel, D. (2019) An Ultra High-Density Arabidopsis thaliana Crossover. *Genetics*, **213**, 771–787.

Sedlazeck, F.J., Lemmon, Z., Soyk, S., Salerno, W.J., Lippman, Z. and Schatz, M.C. (2018) SVCollector: Optimized sample selection for validating and long-read resequencing of structural variants. *bioRxiv*, 342386.

Sheikhizadeh, S., Schranz, M.E., Akdel, M., Ridder, D. de and Smit, S. (2016) PanTools: representation, storage and exploration of pan-genomic data. *Bioinformatics*, **32**, i487–i493.

Sims, J., Chouaref, J. and Schlögelhofer, P. (2020) Whole-Mount Immuno-FISH on Arabidopsis Meiocytes (WhoMI-FISH). In pp. 59–66.

Sun, H., Rowan, B.A., Flood, P.J., Brandt, R., Fuss, J., Hancock, A.M., Michelmore, R.W., Huettel, B. and Schneeberger, K. (2019) Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nat. Commun.*, **10**, 1–9.

Taagen, E., Bogdanove, A.J. and Sorrells, M.E. (2020) Counting on Crossovers: Controlled Recombination for Plant Breeding. *Trends Plant Sci.*, **25**, 455–465.

Tepfer, M., Hurel, A., Tellier, F. and Jenczewski, E. (2020) Evaluation of the progeny produced by interspecific hybridization between Camelina sativa and C. microcarpa. *Ann. Bot.*, **125**, 993–1002.

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsisthaliana. *Nature*, **408**, 796–815.

Underwood, C.J. and Choi, K. (2019) Heterogeneous transposable elements as silencers, enhancers and targets of meiotic recombination. *Chromosoma*, **128**, 279–296.

Underwood, C.J., Choi, K., Lambing, C., *et al.* (2018) Epigenetic activation of meiotic recombination near Arabidopsis thaliana centromeres via loss of H3K9me2 and non-CG DNA methylation. *Genome Res.*, **28**, 519–531.

Wijfjes, R.Y., Smit, S. and Ridder, D. de (2019) Hecaton: reliably detecting copy number variation in plant genomes using short read sequencing data. *BMC Genomics*, **20**, 818.

Zhang, X., Wu, R., Wang, Y., Yu, J. and Tang, H. (2020) Unzipping haplotypes in diploid and polyploid genomes. *Comput. Struct. Biotechnol. J.*, **18**, 66–72.

Summary

In plants, genetic diversity is important for species adaptation in nature and for crop improvement through breeding. For decades, genetic diversity was measured at the nucleotide and gene level. With the abundance of whole genome sequences and chromosome level genome assemblies, the importance of diversity at the genome level - due to rearrangements within genomes - became more obvious given its effect on phenotypic diversity. In genetics research it has become crucial to understand the role of rearrangements in genome evolution and environmental adaptation. Likewise in plant breeding, it is essential to inventory diversity in a wide range of plant varieties for crop improvement as well as to understand how to create new varieties. In this thesis, we touch upon three different genomic rearrangements: recombination, structural variations (SVs) and transposon insertions. We approach them from a fundamental point of view and consider their application to plant breeding.

As a first step, we studied genomic recombination investigating the position, distribution and genomic patterns of recombination (crossovers) in offspring from interspecific crosses of tomato (chapter 2). Interspecific recombination is used to introgress wild alleles into elite crops, but this is often problematic, among others due to linkage drag. We learned that recombination in tomato occurs preferentially at the distal ends of chromosomes in open chromatin or euchromatin. On a closer look, at the sequence level, the crossover breakpoints showed preferential occurrence near certain sequence motifs and transcription start sites. The next step was to develop a genome-wide predictor for regions prone to recombination. In chapter 3, we developed a classifier based on genomic patterns associated with recombination based on results presented in the second chapter as well as from studies on other plants (such as Arabidopsis, rice and maize). This model not only accurately predicted recombination probability, it also linked new genomic features such as DNA shape to recombination. With the application of this model on different plants we found that particular genomic features were predictive across the plant kingdom. Breeders can use this information to estimate the outcome of breeding programs, prior to the beginning of lengthy crossing experiments.

Given the known effects of collinearity and SVs on recombination, we turned our focus to SVs. In chapter 4, we identified and inventoried SVs in another model crop, melon, due to its intriguing phenotypic and genetic diversity and cross-pollinating behavior. We revealed genetic diversity in 94 melon lines and 6 wild relatives of melon from an SV point of view and unraveled the history of melon breeding. Due to significant overlap between certain SVs and agronomic traits such as fruit ripening, fragrance, and stress response, we were able to see footprints of selective breeding in melon subspecies. Studying inversions in melon in detail showed that they likely resulted from meiotic recombination events, as their breakpoints share the same sequence motif.

A specific subset of SVs, transposon insertions, were studied next. Transposons can create genetic diversity by moving along the genome (semi)autonomously. For the study of transposons, we returned to tomato, as knowledge on retrotransposons and related phenotypic changes is well established. In chapter 5, we surveyed the activity of retrotransposons in 60 cultivated tomato lines, with a focus on the *Rider* family of retrotransposons. We created an inventory of transposon insertion polymorphisms and reported the possible effect of insertions on the expression of agronomy-related genes.

The thesis concludes with a general discussion on the genomic rearrangements studied, insights obtained to further fundamental research, the consequences for plant breeding and what is necessary from the research community to improve the fields of bioinformatics, genetics and plant breeding.

Acknowledgements

I would like to thank several people who contributed either directly or indirectly to the work presented in this thesis.

Firstly, I would like to thank my supervisors *Sander, Dick* and *Aalt-Jan* for hiring me to the position; for teaching me how to be more critical; for curving my perfectionism; for patiently correcting my English and for their overall guidance during my PhD. Their knowledge on different subjects and the combination of their unique expertise made this thesis special and cross disciplinary. You were kind to share your knowledge and wisdom which I am grateful for.

I also would like to thank *Sandra Smit* and *Ruud de Maagd* for their contribution and guidance on my last research chapter on the TIPs project; *Hans de Jong* and *Erik Wijnker* for their guidance and discussions on meiotic recombination; *Paul Fransz, Ian Henderson* and *Peter Schlögelhofer* for hosting me and providing a learning opportunity about ChIP-Seq analyses which unfortunately didn't make it to this thesis, but deeply contributed to my understanding of factors influencing meiotic recombination.

I was blessed to be part of two groups, Bioscience Applied Bioinformatics and the chair group Bioinformatics. It was double the fun as I got to join yearly outings and X-mas celebrations twice and double the feedback as I got to present my work twice. For all the scientific discussions, coffee breaks and random chats that relaxed us and at the same time stimulated us, I would like to thank *Jan, Linda, Henri, Gabino, Sara, Paul, Ronald, Jan-Peter* and *Jennifer* of the AB group; *Judith, Harm, Sandra, Justin, Miguel, Raul, Vittorio, Mehmet, Janani, Sander R., Rens, Roven, Siavash, Carlos, Martijn, Ronald, Nando, Victoria, Barbara, Eef* of the BIF group and many other members of these groups whom I may have forgotten to include here. Also, thanks to our secretaries *Hana Nobels, Marie-Jose* and *Maria* for administrative support and our system admins *Jennifer* and *Gwen* for keeping cluster up and running. More thanks to the members of the Bioscience department for keeping me in the loop on the plant biology studies and their feedback on the plant genomics aspect of this thesis. With some of my fellow PhD colleagues, our friendship extended from Radix to living rooms. I am glad that I was part of movie and game nights.

I would like to thank especially my beloved office mates *Sven* and *Saulo* for conversations over cultures, life and work; for helping me to integrate into the Netherlands and answering my random questions about anything. You made the days pass in the office ^.^.

I am proud to be a part of 13 early stage researchers in COMREC Marie Curie PhD fellowship. We meet almost every time in a different country and attended conferences, meetings, workshops and pubs. With the group I gained lots of knowledge on meiosis, extended my network and developed friendships. I would be an orphan without you, my dear COMREC fellas *Jihed, Marina, Jason, Maria, Mikel, Mateusz, Gunjita, Divya, Vanesa, Amy, Adrian* and *Pablo*.

Being and working in the application field in KeyGene, I got to learn the aspects that shaped my view of the field and thus the discussion chapter. I am glad that I am working at KeyGene not just because all the cutting-edge technology projects but also because very friendly working environment.

It is not possible to finish a PhD project spanning several years without emotional, moral and mental support. I was lucky that I was in several friendship circles.

Firstly, I want to thank the *Thursday group*, and *Saulo* for introducing me to them. Over the years we enjoyed several trips and games over special beers together. Their consistent presence on every Thursday was an assurance of a strong laughter. I would like to thank all the former, present and future members of Thursday group *Paul, Erik I. Petra, Rico, Inge, Viola, Martijn* for their friendship but especially *Erik R.* and *Heleen* for sailing to adventures in different realms.

As I got sick of loneliness, *Huismus* opened their arms for me and they gave a room to live in a house full of friendly faces and compassion. I would like to thank all of my ex-housemates *Rose, Eva, Jeroen, Maricella, Ananda, Sjoerd, Matthijs, Anneke* but especially *Myrthe* and *Rene* as they became more like a family.

I would like to thank *Elianne* for helping me to find strength in me to finish this thesis for the past two years.

I would like to thank *Heleen* and *Sven* for accepting the very important task of making me chill & cheer on the defence day and standing next to me as my knights in shining armor – my paranymphs.

I would like to thank my family for believing in me and supporting me and never giving up on me. I would like to thank my aunt *Hacer* for encouraging me to chase my dream of doing a PhD abroad and walking me through the initial culture shock. *Annecim, babacım,* sizlerin sonsuz desteği olmadan buralara gelemezdim. Benden sevginizi ve desteğinizi esirgemediğiniz için teşekkür ederim. *Meloşum,* canım kardeşim, varlığın bile yola devam etmeme yetti. Beni hiç yalnız bırakmadığın için teşekkür ederim.

And finally, I would like to thank my partner, *Hèctor*, for loving me, supporting me even when he disagrees and sacrificing holidays and weekends for the sake of completion of this thesis. Since my PhD contract ended three years ago, I always told him that I will be finishing in 6 months, I am glad to announce that this notorious 6 months period is over.

List of Publications

Demirci, S., Fuentes, R.R., Dooijeweert, W. van, Aflitos, S., Schijlen, E., Hesselink, T., Ridder, D. de, Dijk, A.D.J. van and Peters, S. (2021) Chasing breeding footprints through structural variations in *Cucumis melo* and wild relatives J. Wendel, ed. *G3 Genes|Genomes|Genetics*, **11**. (Chapter 4)

Demirci, S., Peters, S.A., Ridder, D. de and Dijk, A.D.J. van (2018) DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J.*, **95**, 686–699. (Chapter 3)

Demirci, S., Dijk, A.D.J. van, Sanchez Perez, G., Aflitos, S.A., Ridder, D. de and Peters, S.A. (2017) Distribution, position and genomic characteristics of crossovers in tomato recombinant inbred lines derived from an interspecific cross between *Solanum lycopersicum* and *Solanum pimpinellifolium*. *Plant J.*, **89**, 554–564. (Chapter 2)

Education Statement of the Graduate School

Experimental Plant Sciences

The Graduate School EXPERIMENTAL PLANT SCIENCES

Issued to: **Sevgin Demirci**
Date: **02 November 2021**
Group: **Bioinformatics**
University: **Wageningen University & Research**

1) Start-Up Phase	*date*	*cp*
► **First presentation of your project**		
Bioinformatic analyses of meiotic recombination in tomato hybrids and related species	16 Oct 2014	1,5
► **Writing or rewriting a project proposal**		
► **Writing a review or book chapter**		
► **MSc courses**		
Subtotal Start-Up Phase		1,5

2) Scientific Exposure	*date*	*cp*
► **EPS PhD student days**		
EPS Get2Gether event, Soest, NL	29-30 Jan 2015	0,6
EPS Get2Gether event, Soest, NL	9-10 Feb 2017	0,6
► **EPS theme symposia**		
EPS theme 4 'Genome Biology', Wageningen, NL	3 Dec 2014	0,3
EPS theme 4 'Genome Biology', Amsterdam, NL	15 Dec 2015	0,3
EPS theme 4 'Genome Biology', Wageningen, NL	16 Dec 2016	0,3
► **Lunteren Days and other national platforms**		
Annual meeting Experimental Plant Sciences (EPS), Lunteren, NL	13-14 Apr 2015	0,6
Annual meeting Experimental Plant Sciences (EPS), Lunteren, NL	11-12 Apr 2016	0,6
Annual meeting Experimental Plant Sciences (EPS), Lunteren, NL	10-11 Apr 2017	0,6
Annual meeting Experimental Plant Sciences (EPS), Lunteren, NL	9-10 Apr 2018	0,6
Netherlands Bioinformatics and Systems Biology Conference (BioSB), Lunteren, NL	20-21 May 2015	0,6
Netherlands Bioinformatics and Systems Biology Conference (BioSB), Lunteren, NL	19-20 Apr 2016	0,6
Netherlands Bioinformatics and Systems Biology Conference (BioSB), Lunteren, NL	4-5 Apr 2017	0,6
Netherlands Bioinformatics and Systems Biology Conference (BioSB), Lunteren, NL	15-16 May 2018	0,6
► **Seminars (series), workshops and symposia**		
Workshop: Data structures in Bioinformatics (COST SeqAhead), Montpellier, France	8-9 Dec 2014	0,6
Workshop: COMREC Bioinformatics Workshop, Wageningen, NL	4-6 Feb 2015	0,9
Workshop: COMREC Advanced Methodologies in Meiosis Research, Madrid, Spain	5 May 2015	0,2
Workshop: COMREC Commercial Plant Breeding	18 May 2016	0,1
Symposium: EPS Symposium - Omics Advances for Academia and Industry: Towards True Molecular Plant Breeding, Wageningen, NL	11 Dec 2014	0,3
Symposium: BioSB youngCB PhD Retreat, Lunteren, NL	18 Apr 2016	0,3
Symposium: 3rd Wageningen PhD Symposium - Diversity in Science, Wageningen, NL	26 Apr 2016	0,3
Symposium: WURomics Symposium, Wageningen, NL	15 Dec 2016	0,3
► **Seminar plus**		
► **International symposia and congresses**		
EMBO Conference on Meiosis, Oxford, UK	30 Aug - 4 Sep 2015	1,5
EMBO Conference on Meiosis, Hvar, Croatia	27 Aug - 1 Sep 2017	1,5
Control of Meiotic Recombination (COMREC) 1st annual meeting, Madrid, Spain	4 May 2015	0,3
Control of Meiotic Recombination (COMREC) mid-term meeting, Brussels, Belgium	25 Nov 2015	0,3
Control of Meiotic Recombination (COMREC) 2nd annual meeting, IPK, Gatersleben, Germany	17 May 2016	0,3
Control of Meiotic Recombination (COMREC) final meeting, Cambridge, UK	10-11 May 2017	0,6
► **Presentations**		
Poster: "Analysing meiotic recombination hotspots in the tomato genome" presented at BioSB and EPS annual meeting	13-14 Apr & 20-21 May 2015	1,0
Poster: "Analysing meiotic recombination hotspots in the tomato genome" presented at EMBO Meiosis conference	30 Aug 2015	1,0
Poster: "High resolution meiotic recombination profiles in recombinant inbred lines of tomato" presented at EPS annual meeting	11-12 & 19-20 Apr 2016	1,0
Poster: "Developing prediction models for meiotic crossovers in plant species" presented at BioSB and EPS annual meeting	4-5 & 10-11 Apr 2017	1,0
Poster: "Finding meiotic crossovers in tomato" presented at COMREC final meeting	11 May 2017	1,0
Poster: "Developing prediction models for meiotic crossovers in plant species" presented at EMBO Meiosis conference	27 Aug 2017	1,0
Talk: "Bioinformatic analyses of meiotic recombination in tomato hybrids and related species", COMREC 1st annual meeting	4 May 2015	1,0
Talk: "Developing a prediction model for meiotic crossovers in tomato", EPS Theme 4 Symposium	16 Dec 2016	1,0
Talk: "Bioinformatic analyses of meiotic recombination in tomato hybrids and related species", COMREC final meeting & invited at Amsterdam University	10 May & 6 Jul 2017	1,0
Talk: "Genomic features predictive of meiotic recombination in plant species", Annual meeting EPS	10 Apr 2018	1,0
Talk: "Revealing structural variations and recombination breakpoints in plant genomes", B-wise Seminar	8 Jan 2019	1,0
► **IAB interview**		
► **Excursions**		
Subtotal Scientific Exposure		25,4

Continued on the next page

3) In-Depth Studies	date	cp
► **Advanced scientific courses & workshops**		
Pattern Recognition (BioSB)	23-27 Mar 2015	3,0
Principles, Statistical and Computational Tools for Reproducible Science (edX, online)	Feb - Apr 2018	1,0
► **Journal club**		
Bioinformatics journal club held every other week	2015-2018	3,0
► **Individual research training**		
COMREC Secondment: ChIP analyis, University of Amsterdam, NL	4-5 Nov 2015	0,3
COMREC Secondment: ChIP experiment steps, University of Vienna, Austria	4-8 Apr 2016	1,5
COMREC Secondment: ChIP-Seq data analysis, University of Cambridge, UK	9-13 Oct 2017	1,2
Subtotal In-Depth Studies		10,0

4) Personal Development	date	cp
► **General skill training courses**		
WGS PhD Competence Assessment	8 Apr 2015	0,3
Writing scientific papers, reports and grant proposals, COMREC	6 May 2015	0,2
WGS Course Scientific Writing	Oct - Dec 2015	1,8
EPS Introduction Course	11 Feb 2016	0,3
WGS Course Interpersonal Communication for PhD candidates	21-22 Apr 2016	0,6
Business Skills and Entrepreneurship, COMREC	18 May 2016	0,1
WGS PhD Workshop Carousel	7 Apr 2017	0,3
WGS Course Efficient Writing Strategies	Apr - Jun 2017	1,3
► **Organisation of meetings, PhD courses or outreach activities**		
Outreach activity: “Meet the Scientist”, Thinktank Museum, Birmingham, UK	19 Feb 2015	0,3
Convener at 3rd Wageningen PhD Symposium, Wageningen, NL	26 Apr 2016	1,0
Helper at Fascination of Plants Day, Belmonte Arboretum, Wageningen, NL	16 May 2015	0,1
Helper at EPS Get2Gether event	29-30 Jan 2015	0,1
► **Membership of EPS PhD Council**		
Subtotal Personal Development		6,4

TOTAL NUMBER OF CREDIT POINTS*	**43,3**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS with a minimum total of 30 ECTS credits.

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by EU FP7 COMREC Marie Curie Initial Training Networks program project number 606956. Part of this research was financed by the Topsector Horticulture and Propagation Materials project 100 Melon Genome Project (https://topsectortu.nl/nl/100-meloen-genoom-project) project number 1310-034.

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by Evelien Jagtman || https://evelienjagtman.com

Printed by ProefschriftMaken || www.proefschriftmaken.nl