Meiotic recombination in natural and experimental plant populations

Roven Rommel Fuentes

Propositions

1. Crop domestication is a threat to food security.
   (this thesis)

2. Performing recombination profiling as standard leads to more efficient plant breeding.
   (this thesis)

3. All genome assemblies are wrong, but some are useful.

4. Lack of interdisciplinary training hampers independence of bioinformaticians.

5. Increased immigration in Europe is manifested by growing popularity of eating rice.

6. GMO can be effectively promoted in a wider audience by strategically demonizing it in social media.

Propositions belonging to the thesis, entitled

Meiotic recombination in natural and experimental plant populations

Roven Rommel Fuentes
Wageningen, 28 August 2023

Meiotic recombination in natural and experimental plant populations

Roven Rommel Fuentes

Thesis committee

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

Co-promotors
Dr S.A. Peters
Senior Scientist, Bioscience
Wageningen University & Research

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

Other members
Dr L. Fokkens, Wageningen University & Research
Dr C. Lelivelt, Rijk Zwaan Breeding B.V., Fijnaart
Dr R. Mercier, Max Planck Institute for Plant Breeding Research, Cologne, Germany
Prof. Dr B. Zwaan, Wageningen University & Research

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

Meiotic recombination in natural and experimental plant populations

Roven Rommel Fuentes

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 Monday 28 August 2023
at 11 a.m. in the Omnia Auditorium.

Roven Rommel Fuentes
Meiotic recombination in natural and experimental plant populations,
173 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2023)
With references, with summary in English

ISBN 978-94-6447-748-1
DOI https://doi.org/10.18174/633063

Table of contents

Chapter 1 Introduction 7

Chapter 2 Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing 31

Chapter 3 Phasing HiFi reads to detect high-resolution crossovers in a pollen gamete pool 59

Chapter 4 Domestication shapes recombination patterns in tomato 73

Chapter 5 Aberrant patterns of recombination in interspecific tomato hybrids reveal breeding bottlenecks 107

Chapter 6 Discussion 151

Summary 167

List of publications 169

Acknowledgements 171

Chapter 1

Introduction

Most eukaryotes propagate via sexual reproduction, involving the reduction of chromosome copies during meiosis and the restoration of ploidy level during fertilization (Mercier *et al.*, 2015). Meiosis is a cellular mechanism essential for halving the number of chromosomes and shuffling genetic information to produce progenies with unique allele and chromosome combinations. In particular, meiotic recombination involves the physical exchange of DNA between parental chromosomes. The concept of recombination was first formulated with the theory of gene linkage and crossing-over of Hunt Morgan (1913) and first experimentally demonstrated by Harriet Creighton and Barbara McClintock (1931) by correlating cytological and genetic exchanges in maize. The recombination process is conserved among animals, plants, and fungi but there are also distinct variations between species (Gerton and Hawley, 2005; Ramesh *et al.*, 2005).

Next to random mutations, transposon activity, whole genome duplication, chromosomal breaks and horizontal gene transfer, the shuffling of genetic materials during meiosis provides organisms a way to increase genetic diversity. Recombination ensures the generation of offspring that are distinct from the parents. This biological process has become integral in understanding many living organisms, specifically their reproduction and inheritance. The ability to generate genetic diversity also makes meiotic recombination a tool to breed different animals and plants. Breeders rely on recombination to produce favorable gene combinations, conferring better traits to progenies such as to overcome threats and diseases. Here, we focus on plants because of the urgency of improving breeding practices for many crops. With the steady increase of the human population and global warming that intensify food insecurity and reduce crop yield (Seck *et al.*, 2012; Ray *et al.*, 2013; Zhao *et al.*, 2017), finding new and better ways to develop novel crop varieties has become a serious research endeavor. Plant breeding involves the generation of elite varieties with higher yield, better fruits, disease resistance and durability in changing environments. To improve plants, favorable traits from the parents are combined through crossbreeding and multiple rounds of selection. Traditionally, breeders generate a large offspring population and select the progeny exhibiting the set of desired traits from the parents. This shuffling of traits from the parents is mediated by meiotic recombination. Many of the processes and components of meiotic recombination are shared among plants and some are unique from other organisms, making plants an excellent system for studying conservation and diversity in the process of meiosis (Wang and Copenhaver, 2018). Tinkering with the recombination mechanism has been a subject of much research for years because of its potential to revolutionize breeding. This makes it important to uncover the nature of meiotic recombination and the factors contributing to its frequency and localization in the genome.

1.1 The mechanism of recombination

Meiosis involves two rounds of nuclear division. Homologous chromosomes separate during meiosis I and sister chromatids separate during meiosis II, resulting in four daughter cells. In animals, these cells can differentiate directly into gametes; in plants, these cells undergo postmeiotic mitotic divisions to produce gametophytes (Wang and Copenhaver, 2018). The female gametophyte, or embryo sac, contains the egg cell while the male gametophyte, or pollen grain, contains the sperm cell. During meiosis I, specifically in the prophase I stage (Figure 1.1), homologous chromosomes are paired and connected by the synaptonemal complex (SC) (Mercier *et al.*, 2015; Wang *et al.*, 2021). Afterwards, reciprocal exchange or crossover (CO) occurs between the chromosome pairs, forming bivalents. An obligatory CO per bivalent ensures that chromosomes segregate equally to opposite poles during the first meiotic division. Disruption in crossover may lead to imbalanced disjoining of chromosomes (aneuploidy), which may yield unviable gametes or defects in offspring.

The process of recombination is initiated after the introduction of DNA double strand breaks (DSBs) along the chromosomes. DSBs and the corresponding protein, SPO11, implicated for their formation, were discovered in budding yeast Saccharomyces cerevisiae (de Massy *et al.*, 1995; Keeney and Kleckner, 1995; Keeney *et al.*, 1997). In plants, DSB formation is catalyzed by the SPO11-containing complex at about 100-500 genomic sites (Pawlowski *et al.*, 2003; Sanchez-Moran *et al.*, 2007), but only 1 to 2 DSBs are repaired into COs per chromosome pair, suggesting that the number of DSB does not limit CO formation. For example, among the ~500 DSBs in maize, only about 20 form COs (Franklin *et al.*, 1999). During meiosis, DSB-containing DNA is processed to invade either the intact sister chromatid or one of the homologous chromatids to form a displacement loop structure (D-loop) (Mercier *et al.*, 2015). This intermediate can be processed to form a CO through two homologous recombination (HR) pathways conserved in most eukaryotes. These pathways facilitate use of homolog as template for repair, resulting in reciprocal exchanges between non-sister chromatids. One of the pathways relies on a group of proteins collectively called ZMMs (named after the S. cerevisiae proteins Zip1-4, Mer3, Msh4, and Msh5) to form a double-Holliday junction (dHJ) and produce class I COs, which exhibit interference or the reduced likelihood of CO occurring in adjacent intervals (Mercier *et al.*, 2015; Lambing *et al.*, 2017a; Wang and Copenhaver, 2018). This interference disperses class I COs along the chromosome. The other pathway is ZMM-independent and results in class II COs that are independently distributed, showing no interference. Class II COs rely on MUS81 for their formation and its mutation reduces COs by 10% (Berchowitz *et al.*, 2007; Higgins *et al.*, 2008). Most COs are class I, only 5-30% are class II (Anderson *et al.*, 2014; Mercier *et al.*, 2015).

Intermediates may also be resolved into noncrossovers (NCOs) through mechanisms like synthesis-dependent strand annealing (SDSA) and dHJ dissolution. Anti-crossover proteins such as FANCM, RECQ4 and FIGL1 negatively regulate the

frequency of CO by directing DSBs into NCOs pathways (Crismani *et al.*, 2012; Fernandes *et al.*, 2018; Girard *et al.*, 2014; Girard *et al.*, 2015; Seguela-Arnaud *et al.*, 2015). Even though many DSBs are resolved into NCOs, detection and characterization of NCOs in plants is still extremely difficult. One reason for the low detection rate of NCOs is the lack of genetic evidence left, as DSBs may be repaired using sister chromatids or there may be insufficient polymorphisms between parental chromosomes. If an NCO overlaps a polymorphism, it can be discovered as a gene conversion (GC). However, detecting GCs is hard since they span less than 1 kbps and they mostly overlap one polymorphism only (Drouaud *et al.*, 2013; Wijnker *et al.*, 2013; Si *et al.*, 2015; Xu *et al.*, 2008).

1.2 The role of recombination in evolution

1.2.1 Speciation and domestication

Recombination generates genetic diversity upon which natural selection can operate, but it can also break favorable combinations of alleles, possibly resulting in reduced fitness (Otto, 2009; Ortiz-Barrientos *et al.*, 2016). Changes to recombination distributions may consequently affect not just the genetic diversity of an organism but its survival and adaptation. It is then important to understand any changes to the recombination landscape, i.e. the genome-wide CO distribution. The dynamics of these recombination landscapes have been explored through the comparison of recombination events between and within species. Recombination rates in different genomic regions may be estimated for a population and be used to account for changes relative to other populations. There are genomic regions with conserved recombination rates within species or between closely related species, but there are also highly variable patterns (local distributions) of rates of recombination (Smukowski and Noor, 2011).

The observed changes in recombination patterns between related species raises the question whether these changes influenced the evolution of these organisms. It was reported that variations in recombination rates play a major role in adaptation and speciation (Ortiz-Barrientos *et al.*, 2016). Selection in regions with low recombination rates resulted in a heterogeneous landscape of differentiation and diversity (Burri *et al.*, 2015; Ortiz-Barrientos and James, 2017; Dumont, 2020). If a genomic region has a low recombination rate, selection on one site reduces genetic diversity in linked sites through hitchhiking of neighboring alleles. The selection of adaptive alleles may allow other alleles, unfavorable for the organism to thrive in different environments, to hitchhike along. On the other hand, regions with higher levels of recombination are expected to exhibit less evidence of linked selection, occurring in populations where selected alleles interfere with each other (Roze and Barton, 2006), or a certain genetic background hinders allele fixation (Keightley and Otto, 2006). This implies that changes in recombination may influence the patterns of diversity and divergence.

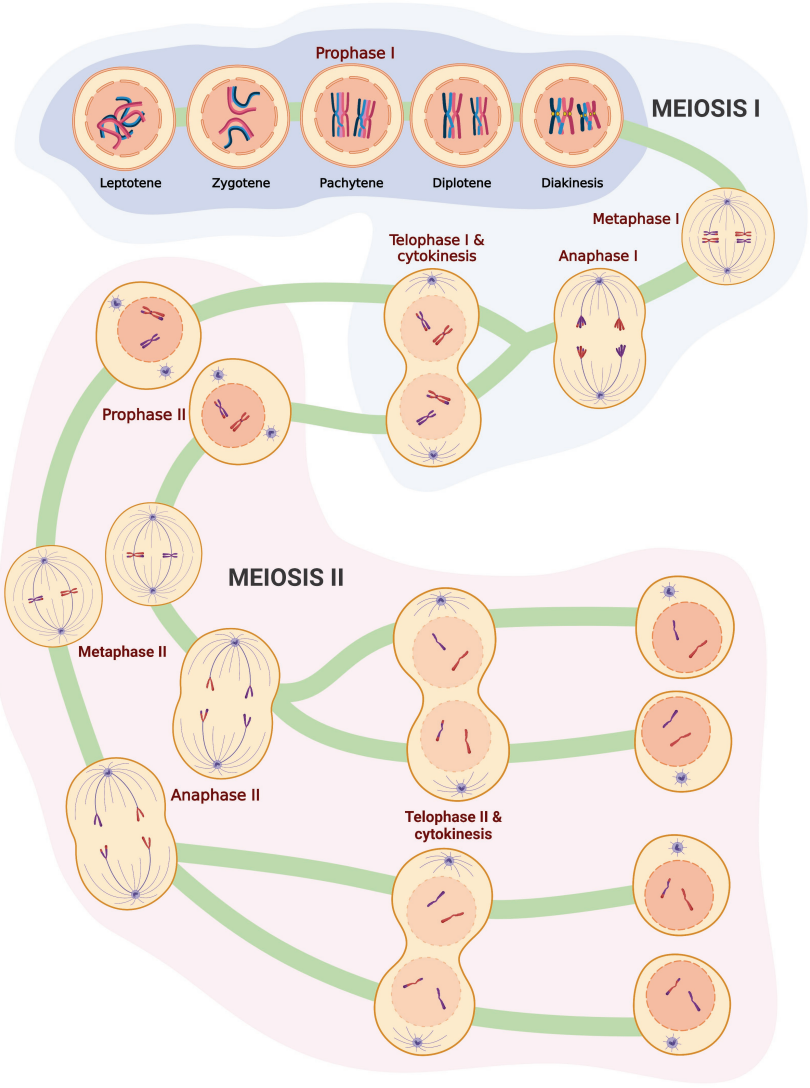


Figure 1.1. Stages of meiosis. After replication, the chromosomes arrange into long thin strands. This is followed by the synapsis of homologous chromosomes in zygotene, facilitated by the synaptonemal complex (SC). Meiotic crossover occurs in pachytene, then the SC disassembles in diplotene with visible chiasma. In the last stage of prophase I (diakinesis), chromosomes become condensed and are recognized as bivalents. This is followed by the movement of bivalents to opposite poles and the reduction of chromosome numbers in the first division in telophase I. Further reduction occurs in meiosis II which forms four haploid gametes. This figure was created with BioRender.com.

Aside from natural selection, differences in recombination patterns and rates were observed between wild and domesticated species, suggesting that domestication was accompanied by modification in recombination patterns. Domestication is a process of human-directed evolution, selecting favorable traits to create forms suitable for human needs (Doebley *et al.*, 2006; Yang *et al.*, 2019). High recombination rates are found in the distal and interstitial chromosome regions of domesticated barley and a wild relative, respectively (Dreissig *et al.*, 2019). Another example is the higher recombination rate in the domesticated population of cacao compared to that in wild populations (Schwarzkopf *et al.*, 2020) and the higher recombination rate in genes associated with agronomic traits such as fiber quality in domesticated cotton (Shen *et al.*, 2019). Domestication is accompanied by significant loss of genetic diversity, which makes recombination less *effective* (Moyers *et al.*, 2018): if homologous chromosomes contain identical sequences or limited heterozygosity, recombination produces no changes in the resulting gametes. In contrast to effective recombination, *actual* recombination measures the total number of COs regardless of whether they resulted in a new haplotype or not. It was estimated that domestication resulted in an average of ~40% loss in diversity in fruit crops, which consequently affects recombination (Miller and Gross, 2011).

1.2.2 Adaptation and changing climate

In evolution, recombination mechanisms respond to selective pressures or external stressors, demonstrated by shifts in recombination rates. A study on two natural populations of *Drosophila pseudoobscura* revealed that changes in CO rates have been driven by adaptive evolution (Samuk *et al.*, 2020). Moreover, exposure to pathogens increased the recombination rate in Arabidopsis and tobacco and shifted chiasmata toward the interstitial regions in tomato and barley (Kovalchuk *et al.*, 2003; Boyko *et al.*, 2007; Andronic, 2012). Recombination hotspots, i.e. genomic regions with high recombination rates, are enriched with resistance (R) genes. This is favorable for rapid diversification to combat new or highly variable pathogens (Nieri *et al.*, 2017). In rice, increased recombination rates are found in genes involved in stress response and environmental stimuli (Si *et al.*, 2015), which suggests that recombination may play a role in adaptive evolution. Similar to biotic stresses, increase in temperature has also been reported to change the frequency and pattern of recombination in varying degrees in most eukaryotes (Dowrick, 1957; Higgins *et al.*, 2012; Phillips *et al.*, 2015; Morgan *et al.*, 2017). Elevated temperature increases recombination frequency in *Arabidopsis thaliana* and rice (Francis *et al.*, 2007; Si *et al.*, 2015) and reduces recombination rates in *Endymion nonscriptus*, *Rhoeo spathacea*, and barley (Lin, 1982; Higgins *et al.*, 2012; Wilson, 2004). Alteration of CO frequency in regions with normally low frequency, such as the interstitial regions, could bring about phenotypic changes, breaking adaptive linkages or introducing new haplotypes that could confer survival in higher temperatures. In populations of barley, recombination rates were observed to be

influenced by environmental conditions such as temperature and temperature oscillation (Dreissig *et al.*, 2019). Extreme temperature changes may cause structural failure in the synaptonemal complex, which can lead to unsuccessful recombination or chromosome missegregation (Morgan *et al.*, 2017). In fission yeast, changes in global CO frequency distribution in changing environments are found to be mediated by a DNA sequence motif, activating recombination hotspots through transcription factor binding (Protacio *et al.*, 2022). However, such adaptive responses to various stressors or the ability to rapidly cope with environmental pressures are not found in every organism, which can result in damage, death or even species extinction. With the sensitivity of the meiotic recombination process to environmental perturbations, the changing climate may pose a threat to the reproductive systems of many species and the stability of major food crops.

1.3 Recombination in plant breeding

Recombination produces numerous allele combinations that breeders can utilize for developing new plant varieties, but the unpredictable occurrence across the genome slows down the breeding process (Wijnker and de Jong, 2008). Understanding the mechanisms and factors affecting meiotic recombination may provide ways to control it for breeding applications, specifically for characterizing and manipulating haplotypes associated with the desired phenotypes. A breeding strategy called introgressive hybridization allows the transfer of desirable traits like disease resistance from donor species/accessions to the recipient individual. It relies on the recombination mechanism to produce gametes, combining the donor locus with the recipient genome and performing successive backcrossing to remove unwanted alleles. Given the random nature of recombination, this requires screening large populations and multiple generations for an offspring possessing the desired allele combinations. During crop domestication, the pursuit of crop stability and agronomic traits led to a loss of genetic diversity, by selecting alleles that confer the desired observable phenotypes. This substantial reduction of diversity is nowadays reversed by the introgression of alleles from wild relatives into crops, but introgression is challenging for highly divergent species due to incompatibility or reproductive issues.

Increased genetic linkage has been observed in domesticated species and its effect is exacerbated by the reduction of outcrossing or by increasing self-fertilization. This switching of the mating system in many cultivated plants like barley, soybeans and rice allows agronomic and morphological consistency in the crop (Allard, 1999). However, linked selection in selfers is prone to fixate detrimental alleles linked with advantageous alleles (Cutter and Payseur, 2013). This linked selection is particularly evident in domesticated crops, which are selected for fruit and yield phenotypes at the expense of losing resistance to diseases. Inbreeding also increases homozygosity, which consequently reduces the effective recombination rate because crossovers result in gametes without appreciable exchange of alleles between the homologous chromosome, and thus do not

generate genetic diversity. In contrast, the actual recombination rate in many domesticated species is higher than their wild progenitors (Ross-Ibarra, 2004; Wilfert *et al.*, 2007; Groenen *et al.*, 2009). This increased inbreeding in many major crops further accentuates the contrasting patterns of genomic variation and effective recombination rate between wild and domesticated crops.

Since breeding thus faces the stochastic nature of meiotic recombination, increasing CO frequency can produce more combinations of alleles in recombinant offspring, consequently accelerating the breeding processes. There are at least 80 plant genes involved in meiosis (Mercier *et al.*, 2015), some of which have been the subject of meticulous research to manipulate CO frequency and distribution. The average number of COs rarely exceeds three per bivalent, irrespective of the size of the chromosome (Mercier *et al.*, 2015). It was demonstrated in Arabidopsis that mutations in anti-CO pathway genes RECQ4, FANCM and FIGL1 increased CO frequency by at most eight fold (Fernandes *et al.*, 2018; Serra *et al.*, 2018). FANCM was mutated in crops, Brassicas and rice, significantly raising the CO frequency (Blary and Jenczewski, 2019; Mieulet *et al.*, 2018). Similarly, mutation in the anti-CO RECQ4 also increases CO frequency in different crops but is more problematic than mutating FANCM as it has multiple copies in some polyploid crops (Blary and Jenczewski, 2019). Another way to increase CO frequency is to increase the number of copies of HEI10, a ZMM protein, resulting in a more than two fold increase in Arabidopsis (Ziolkowski *et al.*, 2017). Similar attempts to manipulate meiosis-associated genes were performed in rice, pea and tomato to demonstrate application of enhanced recombination in crops (Mieulet *et al.*, 2018). Despite these efforts, the CO frequency is never elevated uniformly along the genome, leaving regions with extremely low recombination rates (Mieulet *et al.*, 2018; Serra *et al.*, 2018). It is thus important to understand the factors limiting recombination in specific genomic regions and promoting it in other regions. Increasing CO frequency may, however, disrupt gene combinations favorable to elite cultivars, indicating that it requires careful use. The elevated frequency may be more helpful in pre-breeding, such as in finding the desired trait or gene and introgressing it into a breeding line.

Aside from efforts to increase COs in specific genomic regions, avoiding COs has also been explored. Some breeding applications may require fixation of certain alleles and chromosome combinations to produce stable breeding materials, requiring the prevention of recombination in certain genomic regions or chromosomes. Such prevention of recombination is in particular visible in domesticated crops, where genomic regions with reduced recombination rates are associated with favorable traits for farming (Allard, 1999; Morrell *et al.*, 2003; Kovach *et al.*, 2007; Lin *et al.*, 2014). Fruit size and quality are important agronomic traits that have been selected for during domestication. Another breeding practice is to suppress recombination to preserve unique allele combinations in heterozygous F1 hybrids, which exhibit hybrid vigor and outperform their parents (Wijnker and de Jong, 2008). This may be done through reverse breeding or the formation of clonal seeds via apomixis (d'Erfurth *et al.*, 2009; Wijnker *et al.*, 2012;

Mieulet *et al.*, 2016; Underwood *et al.*, 2022; Underwood and Mercier, 2022). These technologies have been implemented in Arabidopsis and rice (d'Erfurth *et al.*, 2009; Wijnker *et al.*, 2012; Mieulet *et al.*, 2016) but are not yet ready to substitute the current hybridization techniques due to governmental restrictions on engineered crops.

1.4 Crossover profiling

1.4.1 Early methods

To characterize recombination and better understand the factors contributing to or influencing its occurrence, we first need to profile the recombination landscape or distribution across the genome. The earliest methods of detecting recombination events used microscopy or cytogenetic techniques, counting chiasmata during the pachytene stage and recombination nodules during diakinesis (Rahn and Solari, 1986; Stack *et al.*, 1989; Lawrie *et al.*, 1995; Anderson *et al.*, 2003). The two-dimensional spread of synaptonemal complex (SC) in tomato allowed to more precisely locate recombination nodules than by inspecting chiasmata, enabling a comparison of CO distributions in euchromatin and heterochromatin (Sherman and Stack, 1995). A similar technique to identify recombination nodules was applied in maize to produce high-resolution CO maps (Anderson *et al.*, 2003). To improve the throughput of CO detection, a visual assay was developed that enables the segregation of fluorescent alleles in thousands of pollen tetrads (Francis *et al.*, 2007). The most recent improvements to this visual assay are based on deep learning to analyze the pollen tetrads and avoid manual scoring (Lim *et al.*, 2020). However, these cytological methods detect recombination sites at coarse resolutions and are relatively low throughput.

Another widely-used method to profile recombination in the genome is the building of genetic maps based on a set of markers like SNPs, intronic polymorphic markers and simple sequence repeats. Earlier maps relied on markers detected using restriction fragment length polymorphism (RFLP) probes, expressed sequence tags (ESTs) and other polymerase chain reaction (PCR) based markers (Tanksley *et al.*, 1992; Shirasawa *et al.*, 2010; Robbins *et al.*, 2011). However, such maps contain only a few hundred markers across the whole genome, limiting the resolution of genomic analyses. Another method relying on single-feature polymorphisms (SFPs) increased the number of markers compared to SNP-based markers, prioritizing quantity over quality of markers (Kim *et al.*, 2006).

1.4.2 Genomics methods

With the advent of high-throughput genotyping and next generation sequencing (NGS), more genetic markers can now be used to build high-density linkage maps. In tomato and maize, arrays of several thousands of SNPs were used to generate linkage maps for offspring populations (Ganal *et al.*, 2011; Sim *et al.*, 2012b). The resulting maps, however, still have limited numbers of markers which prevents their

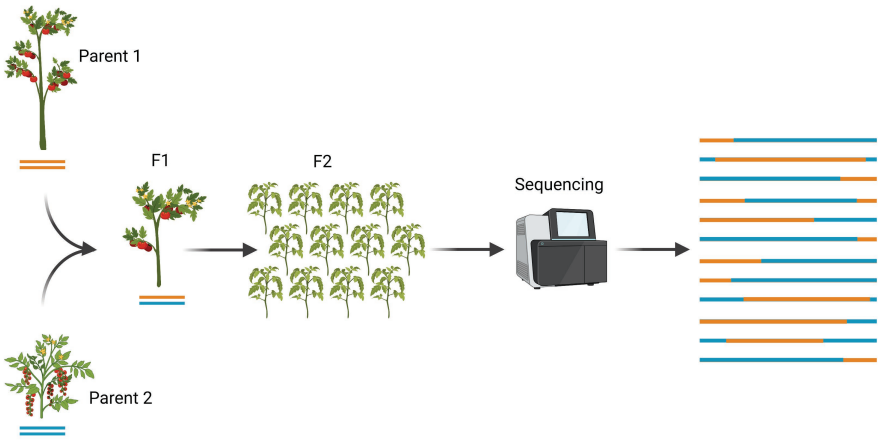


Figure 1.2. Crossover detection. An F1 plant is generated from a cross between two distinct parents. From this F1 plant, an F2 population can be produced. The F2 plants are sequenced and based on the informative markers in the parental genome, recombination sites can be detected. This figure was created with BioRender.com.

use for fine-scale genomics such as comparing recombination events with the occurrence of genome elements like genes, TEs and sequence motifs. Whole genome sequencing of offspring populations allowed construction of recombination maps in different species with even better resolution (Rowan *et al.*, 2015; Si *et al.*, 2015; Pan *et al.*, 2016; Demirci *et al.*, 2017; Kianian *et al.*, 2018; Rowan *et al.*, 2019). In Arabidopsis and maize, recombination events were analyzed in meiotic tetrads, which allowed the detection of genomic exchanges among the four chromatids of the same tetrad such as COs and CO-associated gene conversion (Wijnker *et al.*, 2013; Li *et al.*, 2015). The use of NGS and bioinformatics tools not only helped detect more COs but also enabled genomic analyses that require higher resolution CO regions. Moreover, high-density maps have been useful for breeding, specifically in trait discovery, population analyses and genome assembly. However, NGS-based methods rely on laborious screening and sequencing of offspring plants and cannot account for recombination that leads to or occurs in unviable gametes.

Aside from detecting recombination in an offspring population of a selected cross, it is also possible to estimate recombination rates in a natural population (Auton and McVean, 2007; Chan *et al.*, 2012). This method is linkage disequilibrium-(LD) or coalescence-based, and derives a population-scaled recombination rate from the association of alleles. This rate is an average between both sexes, integrated over many generations (Penalba and Wolf, 2020). Due to the long period of successive recombination events accounted for in this method, it can help answer questions relating to the evolution and plasticity of recombination in a species. It

was found that historical recombination hotspots are drivers of evolutionary divergence in rice, showing association with lineage-specific variations and nucleotide diversity (Marand *et al.*, 2019). LD-based methods helped elucidate the dynamics of recombination rates in wild and domestication populations in barley, cotton and cocoa (Dreissig *et al.*, 2019; Shen *et al.*, 2019; Schwarzkopf *et al.*, 2020).

NGS-based methods for CO detection rely on markers such as SNPs to infer reciprocal exchanges between homologous chromosomes, making them ideal for crosses with divergent parental genomes (Figure 1.2). However, detecting COs in inbreds or crosses of highly similar genomes is challenging due to the lack of segregating markers. Estimating the actual recombination in a population of domesticated accessions of a species faces the same problem, as it also relies on genetic polymorphisms. With low levels of polymorphism between homologous chromosomes or members of a natural population, CO detection or estimation of historical recombination rates using resequencing data cannot account for recombination that occurred in runs of homozygosity. Alternatively, counting chiasmata per bivalent or other cytogenetic methods may provide a more accurate estimate of the actual recombination rate in a highly homozygous genome. A recent study in Arabidopsis inbred lines utilized genetic markers introduced through mutagenesis to profile recombination (Lian *et al.*, 2022), although the resolution of CO regions depends on the number of introduced mutations.

1.4.3 Model crop system

Many early studies on plant recombination have been performed in *Arabidopsis thaliana*, the most popular model plant with relatively small size and short generation time (Koornneef and Meinke, 2010). The first linkage groups for Arabidopsis were constructed in 1965, but it was only after 18 years that the complete genetic map was published, covering only 76 markers (Koornneef *et al.*, 1983; Koornneef and Meinke, 2010). Further developments helped to improve the maps in Arabidopsis (Chang *et al.*, 1988) and to construct maps for other species such as lettuce, rice, barley, tomato, potato and maize (Landry *et al.*, 1987; McCouch *et al.*, 1988; Graner *et al.*, 1991; Tanksley *et al.*, 1992; Gardiner *et al.*, 1993). From a few hundred markers to several thousand markers now, genetic maps have facilitated quantitative trait mapping, marker-assisted breeding and evolutionary studies in different crops. One of the well-studied crops for meiosis is tomato, based on cytological observations and several high-density genetic maps (Peters and Underwood, 2023). The availability of at least 12 wild relative species of tomato enabled the construction of several genetic maps for different interspecific crosses and the identification of many quantitative trait loci such as flowering time and seed infertility (Sim *et al.*, 2012a; Moyle and Graham, 2005; Jimenez-Gomez *et al.*, 2007). NGS-based detection of crossovers was also performed in 52 F6 recombinant inbred lines obtained from an interspecific tomato hybrid (Demirci *et al.*, 2017). The application of genome editing to study meiotic genes and processes has been employed on tomato, such as the mutation in RECQ4 gene and FIGL1 (de Maagd *et*

al., 2020; Mieulet *et al.*, 2018). As one of the most cultivated vegetable crops and the model system for genetic, developmental and physiological studies for fleshy fruits, tomato is proposed as a model system for studying meiosis in eudicot plants (Peters and Underwood, 2023; Giovannoni, 2004).

Several projects led to the resequencing of at least 900 tomato accessions, mostly representing domesticated tomato *Solanum lycopersicum* and its wild progenitor *Solanum pimpinellifolium* (Aflitos *et al.*, 2014; Lin *et al.*, 2014; Tieman *et al.*, 2017; Zhu *et al.*, 2018; Razifard *et al.*, 2020). These data were used to study the diversity of tomato, previous introgressions, fruit flavor and metabolome, and tomato domestication. More recently, 100 tomato accessions were sequenced and 32 genomes were assembled to profile large genetic variations (Alonge *et al.*, 2020; Zhou *et al.*, 2022). These data are useful for the further study of recombination in tomato, particularly through genomics. Tomato has been the focus of many studies, but there is still much to learn on harnessing the recombination mechanism to accelerate tomato breeding.

1.5 Patterns in the genome

1.5.1 Chromosomal distribution

CO occurrences are stochastic, but are found to cluster in hotspots within the subtelomeric regions (Lambing *et al.*, 2017b; Wang and Copenhaver, 2018). Reported in many plant species, including tomato and maize (Li *et al.*, 2015; Demirci *et al.*, 2017), this skewed CO distribution toward the chromosome ends may be explained by the earlier synapsis and DSB formation in those regions compared to the interstitial segments (Anderson *et al.*, 2014; Wang and Copenhaver, 2018; Osman *et al.*, 2021). At a finer scale, the majority of COs cluster in hotspots and are confined to less than a quarter of the genome (Choi *et al.*, 2013; Kianian *et al.*, 2018), limiting the shuffling of alleles and reducing breeding efficiency in the proximal regions. Furthermore, conserved across eukaryotes, recombination is suppressed in centromeres to prevent missegregation and reproductive defects (Copenhaver *et al.*, 1999; Rockmill *et al.*, 2006; Salome *et al.*, 2012; Si *et al.*, 2015; Nambiar and Smith, 2016). Pericentromeric regions exhibit reduced recombination rates but are not completely devoid of it (Li *et al.*, 2015; Demirci *et al.*, 2017; Dreissig *et al.*, 2020). In rice, mutations were introduced in specific anti-CO genes in an attempt to elevate CO frequency in the proximal regions of the genome. However, an increase in COs was found only in regions where wild-type COs already occur, not in coldspots (Mieulet *et al.*, 2018). Further studies must be done to understand factors that limit recombination in the proximal genomic regions.

1.5.2 Genome features

Chromatin structure also contributes to the recombination landscape. To access its DNA substrate, SPO11 requires open chromatin. CO regions are associated with different histone features such as H3K4me3 and H2A.Z, and AT-rich, poly-A and CTT-repeat motifs (Pan *et al.*, 2011; Choi *et al.*, 2013; Wijnker *et al.*, 2013; Demirci *et al.*, 2017; Wang and Copenhaver, 2018). Poly-A motifs are known to prevent nucleosomes from binding DNA, making a genomic region accessible for DSB formation (Field *et al.*, 2008; Choi *et al.*, 2013; Wijnker *et al.*, 2013; Shilo *et al.*, 2015). In Arabidopsis, loss of CG methylation redistributed CO in the distal regions while the disruption of H3K9me2 and non-CG methylation increased pericentromeric COs (Yelina *et al.*, 2012; Underwood *et al.*, 2018). Regulation by methylation is proposed as the underlying mechanism that facilitates changes in recombination patterns during exposure to pathogens or elevated temperature (Melamed-Bessudo and Levy, 2012; Yelina *et al.*, 2012). Thus, exploring the connection between epigenetic profiles and CO localization may reveal ways to manipulate CO patterns by changing the epigenetic landscape.

The preferential occurrence of COs in the distal chromosome is also associated with higher gene density in this genome region. Previous studies in different plant species reported that COs tend to localize in gene promoters and terminators (Pan *et al.*, 2011; Choi *et al.*, 2013; Wijnker *et al.*, 2013; Demirci *et al.*, 2017; Marand *et al.*, 2017; Kianian *et al.*, 2018). Both gene transcription start sites (TSSs) and termination sites (TTSs) are associated with chromatin marks that promote RNA polymerase II (Pol II) transcription and with hypomethylated DNA (Choi *et al.*, 2013; Mercier *et al.*, 2015). In Arabidopsis, the CO-associated motifs mentioned above are often flanking or overlapping TSSs, further corroborating the enrichment of COs in genic regions (Choi *et al.*, 2013). The occurrence of COs in several different plant genomes could reasonably well be predicted using a machine learning model based on DNA sequence information, indicating the relevance of chromosomal features in determining CO positioning (Demirci *et al.*, 2018).

1.5.3 Transposable elements

Apart from enrichment in genic regions, recombination is also associated with transposable elements (TEs), the mobile, repetitive DNA sequences occupying a large part of eukaryotic genomes. TEs are grouped into two classes depending on their transposition mechanism. Class I TEs transpose in a copy-and-paste manner via RNA intermediates, while class II TEs transpose in cut-and-paste via DNA intermediates (Wicker *et al.*, 2007). DSB hotspots in plants have been reported to overlap class II TEs (DNA transposons), specifically miniature inverted-repeat transposable elements (MITEs) (Choi *et al.*, 2018). Examination of COs at high resolution and of population-scaled recombination rates revealed enrichment of specific TE families like *Stowaway*, *SINE* and *Helitrons* in Arabidopsis, potato, rice and maize (Marand *et al.*, 2017; Pan *et al.*, 2017; Choi *et al.*, 2018; Marand *et al.*, 2019). These TE families preferentially insert near genes and it was proposed that

they may act as meiotic recombination enhancers (Kent *et al.*, 2017; Choi *et al.*, 2018). Conversely, class I TEs (retrotransposons), specifically *Gypsy, Copia* and *L1,* show negative association with global recombination rates and exhibit lower DSB levels than expected by chance (Tiley and Burleigh, 2015; Darrier *et al.*, 2017; Marand *et al.*, 2017; Choi *et al.*, 2018). In contrast to class II TEs, retrotransposons are more abundant in the pericentromeric heterochromatin, which exhibit less recombination, and have higher levels of DNA methylation, H3K9me2 and nucleosome occupancy (Dworkin *et al.*, 2017; Underwood *et al.*, 2018; Quesneville, 2020; Liu *et al.*, 2021). CO suppression in these elements may be imposed by their inaccessible chromatin state or their tendency to induce ectopic recombination, resulting in deleterious genome rearrangements. Given this association with specific TE families, the difference in overall recombination landscape between Arabidopsis and other plant species like tomato and maize may be influenced by the distribution of TEs. The Arabidopsis genome has fewer TEs than other plants and they are mostly confined in the pericentromeric and centromeric regions. In contrast, other plant species have more extensive pericentromeric heterochromatin with high retrotransposon content, confining most of the COs to the distal chromosome regions.

1.5.4 Genomic rearrangements

Another feature that shapes the landscape of recombination is structural rearrangement or variation (SV). SV is defined as a change in copy number, orientation or chromosomal location of a genomic segment relative to a reference genome (Medvedev *et al.*, 2009; Escaramís *et al.*, 2015). In many species, SVs account for more base pairs than SNPs and have greater potential impact to gene structure, dosage and location (Alkan *et al.*, 2011; Sudmant *et al.*, 2015; Layer *et al.*, 2014; Fuentes *et al.*, 2019; Alonge *et al.*, 2020). Aside from being a source of genetic diversity, these genomic rearrangements have shown to directly contribute to the evolution of different plant species, particularly in rapid adaptation and domestication (Lye and Purugganan, 2019). Compared with wild relatives, domesticated species have reduced SV diversity, partly due to the fixation of adaptive alleles (Swanson-Wagner *et al.*, 2010; Munoz-Amatriain *et al.*, 2013).

As early as 1921 it was proposed that rearrangements, such as inversions, in heterozygous state can suppress COs (Sturtevant, 1921). The excision site of a 70-kb transposition and the 1.17Mb inversion in Arabidopsis indeed show suppressed recombination (Fransz *et al.*, 2016; Alhajturki *et al.*, 2018). Several explanations for this suppression of COs in SV regions have been suggested, such as the preferential repair of DSBs to NCOs, absence of repair template, higher DNA methylation, incomplete synapsis, and tendency to create inviable gametes, making them unobservable in progenies (Rowan *et al.*, 2019). DSBs in regions with copy number variations cannot be resolved using the non-sister chromatid as template if no matching sequence is available in the homologous chromosome. On the other hand, if a template is available but with multiple copies, it can result in unequal crossovers

(illegitimate crossovers) where one of the resulting gametes loses copies, possibly causing a detrimental effect to the progenies. It was observed in Arabidopsis that induced structural changes resulted in recombination suppression in the affected region and redistribution of COs to other chromosomal regions (Ederveen *et al.*, 2015). Recently, reversal and introduction of inversion to restore or suppress CO, respectively, have been demonstrated using CRISPR/Cas mediated chromosome engineering, proving the direct role of SVs in CO positioning (Schmidt *et al.*, 2020; Ronspies *et al.*, 2022).

SV regions may span so-called "supergenes", i.e. tightly linked genes, which tend to be inherited as a single locus due to suppressed recombination. These supergenes are an important concern for breeders, because many are linked with local adaptation and reproductive isolation (Kirkpatrick, 2010; Thompson and Jiggins, 2014; Schwander *et al.*, 2014). As modern breeding relies on hybridization to restore or introduce genetic diversity in domesticated crops from wild relatives, structural divergence causing heterozygosity and CO suppression between the parental genomes may result in linkage drag, preventing the recombination machinery from breaking undesirable combinations of alleles within SV regions. Rearrangements such as inversions alter the recombination landscape in hybrids, but when fixed between populations do not alter the recombination landscape (Ortiz-Barrientos and James, 2017). Examination of pachytene SC spreads from interspecific tomato hybrids revealed synaptic irregularities due to SVs (Anderson *et al.*, 2010). However, there is still limited information on how rearrangements affect the local variability of recombination in interspecific hybrid crops, mainly because of the prohibitive cost of generating recombination data.

Given the challenges and limitations in the existing methods, alternative methods for CO detection requiring less labor and cost must be implemented. There is a need for high-resolution CO maps that will allow comparisons with small genomic features and large rearrangements. Furthermore, it is crucial to know how the landscape of recombination evolves and consequently affects breeding. In this thesis, we investigated methods to detect CO and explore recombination patterns and their evolution.

1.6 Contributions of this thesis

The aim of this thesis is to probe meiotic recombination in natural and experimental populations of tomato, uncovering factors that influence recombination patterns and providing insights that can help accelerate breeding. I focus on tomato because of its economic and food security importance, and its properties as an excellent model crop for the study of plant meiosis. In **Chapter 2**, I propose a method based on linked read technology for high-throughput detection of CO events from a pool of pollen gametes. It profiles high-resolution maps of COs and gene conversions without the laborious and costly development and screening of a large offspring population, relying only on the germinated pollen from a single interspecific F1 hybrid plant. I identify recombination hotspots and coldspots and uncover genomic

features associated with CO sites. With the development of other sequencing technologies, in **Chapter 3** I explore the use of PacBio hi-fidelity (HiFi) long reads in detecting COs in pollen. This further improves the resolution of COs and demonstrates the extensibility of my method to other sequencing technologies. After profiling the recombination landscape in an experimental hybrid, in **Chapter 4** I explore historical recombination in natural populations of wild and domesticated tomato. I examine the role of domestication to that of evolution in terms of the recombination landscape and reveal the link between genetic changes and structural variants to recombination hotspot divergence during domestication. To further understand the dynamics of recombination and how it can affect breeding, I also compare the recombination landscape of five interspecific tomato hybrids in **Chapter 5**. I uncover the recombination barriers in hybrids and their association with metabolic processes and disease resistance. I also expound on the impact of CO coldspots on the development of better tomato varieties and the solutions to resolve recombination suppression in regions of interest. This chapter exhibits the benefit of the CO-detection method introduced in Chapter 2 in studying meiotic recombination in multiple crosses, which addresses the prohibitive cost of this type of comparative study. Finally, **Chapter 6** describes the crucial importance of studying meiotic recombination to help accelerate breeding and the use of computational methods to decipher new insights on recombination-related processes. I also discuss future prospects and suggest research directions.

1.7 References

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L.*, et al.* 2014. Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing. *Plant J,* 80**,** 136-48.

Alhajturki, D., Muralidharan, S., Nurmi, M., Rowan, B. A., Lunn, J. E., Boldt, H., Salem, M. A., Alseekh, S., Jorzig, C., Feil, R.*, et al.* 2018. Dose-dependent interactions between two loci trigger altered shoot growth in BG-5 x Krotzenburg-0 (Kro-0) hybrids of *Arabidopsis thaliana*. *New Phytol,* 217**,** 392-406.

Alkan, C., Coe, B. P. & Eichler, E. E. 2011. Genome structural variation discovery and genotyping. *Nature Reviews. Genetics,* 12**,** 363-76.

Allard, R. W. 1999. History of plant population genetics. *Annu Rev Genet,* 33**,** 1-27.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D.*, 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., Covey, P. A., Larsen, L. R., Bedinger, P. & Stack, S. M. 2010. Structural differences in chromosomes distinguish species in the tomato clade. *Cytogenet Genome Res,* 129**,** 24-34.

Anderson, L. K., Doyle, G. G., Brigham, B., Carter, J., Hooker, K. D., Lai, A., Rice, M. & Stack, S. M. 2003. High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics,* 165**,** 849-65.

Anderson, L. K., Lohmiller, L. D., Tang, X., Hammond, D. B., Javernick, L., Shearer, L., Basu-Roy, S., Martin, O. C. & Falque, M. 2014. Combined fluorescent and electron microscopic imaging unveils the specific properties of two classes of meiotic crossovers. *Proc Natl Acad Sci U S A,* 111**,** 13415-20.

Andronic, L. 2012. Viruses as triggers of DNA rearrangements in host plants. *Canadian Journal of Plant Science,* 92**,** 1083-1091.

Auton, A. & McVean, G. 2007. Recombination rate estimation in the presence of hotspots. *Genome Res,* 17**,** 1219-27.

Berchowitz, L. E., Francis, K. E., Bey, A. L. & Copenhaver, G. P. 2007. The role of AtMUS81 in interference-insensitive crossovers in *A. thaliana*. *PLoS Genet,* 3**,** e132.

Blary, A. & Jenczewski, E. 2019. Manipulation of crossover frequency and distribution for plant breeding. *Theor Appl Genet,* 132**,** 575-592.

Boyko, A., Kathiria, P., Zemp, F. J., Yao, Y., Pogribny, I. & Kovalchuk, I. 2007. Transgenerational changes in the genome stability and methylation in pathogen-infected plants: (virus-induced plant genome instability). *Nucleic Acids Res,* 35**,** 1714-25.

Burri, R., Nater, A., Kawakami, T., Mugal, C. F., Olason, P. I., Smeds, L., Suh, A., Dutoit, L., Bures, S., Garamszegi, L. Z.*, et al.* 2015. Linked selection and recombination rate variation drive the evolution of the genomic landscape of differentiation across the speciation continuum of *Ficedula* flycatchers. *Genome Res,* 25**,** 1656-65.

Chan, A. H., Jenkins, P. A. & Song, Y. S. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genet,* 8**,** e1003090.

Chang, C., Bowman, J. L., DeJohn, A. W., Lander, E. S. & Meyerowitz, E. M. 1988. Restriction fragment length polymorphism linkage map for *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A,* 85**,** 6856-60.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C.*, 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., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J.*, et al.* 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in *Arabidopsis thaliana* transposons and gene regulatory regions. *Genome Research,* 28**,** 532-546.

Copenhaver, G. P., Nickel, K., Kuromori, T., Benito, M. I., Kaul, S., Lin, X., Bevan, M., Murphy, G., Harris, B., Parnell, L. D.*, et al.* 1999. Genetic definition and sequence analysis of Arabidopsis centromeres. *Science,* 286**,** 2468-74.

Creighton, H. B. & McClintock, B. 1931. A Correlation of Cytological and Genetical Crossing-Over in Zea Mays. *Proc Natl Acad Sci U S A,* 17**,** 492-7.

Crismani, W., Girard, C., Froger, N., Pradillo, M., Santos, J. L., Chelysheva, L., Copenhaver, G. P., Horlow, C. & Mercier, R. 2012. FANCM limits meiotic crossovers. *Science,* 336**,** 1588-90.

Cutter, A. D. & Payseur, B. A. 2013. Genomic signatures of selection at linked sites: unifying the disparity among species. *Nat Rev Genet,* 14**,** 262-74.

d'Erfurth, I., Jolivet, S., Froger, N., Catrice, O., Novatchkova, M. & Mercier, R. 2009. Turning meiosis into mitosis. *PLoS Biol,* 7**,** e1000124.

Darrier, B., Rimbert, H., Balfourier, F., Pingault, L., Josselin, A. A., Servin, B., Navarro, J., Choulet, F., Paux, E. & Sourdille, P. 2017. High-Resolution Mapping of Crossover Events in the Hexaploid Wheat Genome Suggests a Universal Recombination Mechanism. *Genetics,* 206**,** 1373-1388.

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

de Massy, B., Rocco, V. & Nicolas, A. 1995. The nucleotide mapping of DNA double-strand breaks at the CYS3 initiation site of meiotic recombination in *Saccharomyces cerevisiae*. *EMBO J,* 14**,** 4589-98.

Demirci, S., Peters, S. A., de Ridder, D. & van Dijk, A. D. J. 2018. DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J*.

Demirci, S., van Dijk, A. J., Sanchez Perez, G., Aflitos, S. A., de Ridder, D. & 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 Journal,* 89**,** 554-564.

Doebley, J. F., Gaut, B. S. & Smith, B. D. 2006. The molecular genetics of crop domestication. *Cell,* 127**,** 1309-21.

Dowrick, G. J. 1957. The influence of temperature on meiosis. *Heredity,* 11**,** 37-49.

Dreissig, S., Mascher, M. & Heckmann, S. 2019. Variation in Recombination Rate Is Shaped by Domestication and Environmental Conditions in Barley. *Mol Biol Evol,* 36**,** 2029-2039.

Dreissig, S., Maurer, A., Sharma, R., Milne, L., Flavell, A. J., Schmutzer, T. & Pillen, K. 2020. Natural variation in meiotic recombination rate shapes introgression patterns in intraspecific hybrids between wild and domesticated barley. *New Phytol,* 228**,** 1852-1863.

Drouaud, J., Khademian, H., Giraut, L., Zanni, V., Bellalou, S., Henderson, I. R., Falque, M. & Mezard, C. 2013. Contrasted patterns of crossover and non-crossover at *Arabidopsis thaliana* meiotic recombination hotspots. *PLoS Genet,* 9**,** e1003922.

Dumont, B. L. 2020. Evolution: Is Recombination Rate Variation Adaptive? *Curr Biol,* 30**,** R351-R353.

Dworkin, M., Xie, S., Saha, M., Thimmapuram, J. & Kalavacharla, V. K. 2017. Analyses of methylomes of upland and lowland switchgrass (*Panicum virgatum*) ecotypes using MeDIP-seq and BS-seq. *BMC Genomics,* 18**,** 851.

Ederveen, A., Lai, Y., van Driel, M. A., Gerats, T. & Peters, J. L. 2015. Modulating crossover positioning by introducing large structural changes in chromosomes. *BMC Genomics,* 16**,** 1-12.

Escaramís, G., Docampo, E. & Rabionet, R. 2015. A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics,* 14**,** 305-314.

Fernandes, J. B., Seguela-Arnaud, M., Larcheveque, C., Lloyd, A. H. & Mercier, R. 2018. Unleashing meiotic crossovers in hybrid plants. *Proc Natl Acad Sci U S A,* 115**,** 2431-2436.

Field, Y., Kaplan, N., Fondufe-Mittendorf, Y., Moore, I. K., Sharon, E., Lubling, Y., Widom, J. & Segal, E. 2008. Distinct modes of regulation by chromatin encoded through nucleosome positioning signals. *PLoS Comput Biol,* 4**,** e1000216.

Francis, K. E., Lam, S. Y., Harrison, B. D., Bey, A. L., Berchowitz, L. E. & Copenhaver, G. P. 2007. Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proc Natl Acad Sci U S A,* 104**,** 3913-8.

Franklin, A. E., McElver, J., Sunjevaric, I., Rothstein, R., Bowen, B. & Cande, W. Z. 1999. Three-dimensional microscopy of the Rad51 recombination protein during meiotic prophase. *Plant Cell,* 11**,** 809-24.

Fransz, P., Linc, G., Lee, C. R., Aflitos, S. A., Lasky, J. R., Toomajian, C., Ali, H., Peters, J., van Dam, P., Ji, X.*, et al.* 2016. Molecular, genetic and evolutionary analysis of a paracentric inversion in *Arabidopsis thaliana*. *Plant J,* 88**,** 159-178.

Fuentes, R. R., Chebotarov, D., Duitama, J., Smith, S., De la Hoz, J. F., Mohiyuddin, M., Wing, R. A., McNally, K. L., Tatarinova, T., Grigoriev, A.*, et al.* 2019. Structural variants in 3000 rice genomes. *Genome Res,* 29**,** 870-880.

Ganal, M. W., Durstewitz, G., Polley, A., Berard, A., Buckler, E. S., Charcosset, A., Clarke, J. D., Graner, E. M., Hansen, M., Joets, J.*, et al.* 2011. A large maize (*Zea mays L.*) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One,* 6**,** e28334.

Gardiner, J. M., Coe, E. H., Melia-Hancock, S., Hoisington, D. A. & Chao, S. 1993. Development of a core RFLP map in maize using an immortalized F2 population. *Genetics,* 134**,** 917-30.

Gerton, J. L. & Hawley, R. S. 2005. Homologous chromosome interactions in meiosis: diversity amidst conservation. *Nat Rev Genet,* 6**,** 477-87.

Giovannoni, J. J. 2004. Genetic regulation of fruit development and ripening. *Plant Cell,* 16 Suppl**,** S170-80.

Girard, C., Chelysheva, L., Choinard, S., Froger, N., Macaisne, N., Lemhemdi, A., Mazel, J., Crismani, W. & Mercier, R. 2015. AAA-ATPase FIDGETIN-LIKE 1 and Helicase FANCM Antagonize Meiotic Crossovers by Distinct Mechanisms. *PLoS Genet,* 11**,** e1005369.

Girard, C., Crismani, W., Froger, N., Mazel, J., Lemhemdi, A., Horlow, C. & Mercier, R. 2014. FANCM-associated proteins MHF1 and MHF2, but not the other Fanconi anemia factors, limit meiotic crossovers. *Nucleic Acids Res,* 42**,** 9087-95.

Graner, A., Jahoor, A., Schondelmaier, J., Siedler, H., Pillen, K., Fischbeck, G., Wenzel, G. & Herrmann, R. G. 1991. Construction of an RFLP map of barley. *Theor Appl Genet,* 83**,** 250-6.

Groenen, M. A., Wahlberg, P., Foglio, M., Cheng, H. H., Megens, H. J., Crooijmans, R. P., Besnier, F., Lathrop, M., Muir, W. M., Wong, G. K.*, et al.* 2009. A high-density SNP-based linkage map of the chicken genome reveals sequence features correlated with recombination rate. *Genome Res,* 19**,** 510-9.

Higgins, J. D., Buckling, E. F., Franklin, F. C. & Jones, G. H. 2008. Expression and functional analysis of AtMUS81 in Arabidopsis meiosis reveals a role in the second pathway of crossing-over. *Plant J,* 54**,** 152-62.

Higgins, J. D., Perry, R. M., Barakate, A., Ramsay, L., Waugh, R., Halpin, C., Armstrong, S. J. & Franklin, F. C. 2012. Spatiotemporal asymmetry of the meiotic program underlies the predominantly distal distribution of meiotic crossovers in barley. *Plant Cell,* 24**,** 4096-109.

Jimenez-Gomez, J. M., Alonso-Blanco, C., Borja, A., Anastasio, G., Angosto, T., Lozano, R. & Martinez-Zapater, J. M. 2007. Quantitative genetic analysis of flowering time in tomato. *Genome,* 50**,** 303-15.

Keeney, S., Giroux, C. N. & Kleckner, N. 1997. Meiosis-specific DNA double-strand breaks are catalyzed by Spo11, a member of a widely conserved protein family. *Cell,* 88**,** 375-84.

Keeney, S. & Kleckner, N. 1995. Covalent protein-DNA complexes at the 5' strand termini of meiosis-specific double-strand breaks in yeast. *Proc Natl Acad Sci U S A,* 92**,** 11274-8.

Keightley, P. D. & Otto, S. P. 2006. Interference among deleterious mutations favours sex and recombination in finite populations. *Nature,* 443**,** 89-92.

Kent, T. V., Uzunovic, J. & Wright, S. I. 2017. Coevolution between transposable elements and recombination. *Philos Trans R Soc Lond B Biol Sci,* 372.

Kianian, P. M. A., Wang, M., Simons, K., Ghavami, F., He, Y., Dukowic-Schulze, S., Sundararajan, A., Sun, Q., Pillardy, J., Mudge, J.*, et al.* 2018. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat Commun,* 9**,** 2370.

Kim, S., Zhao, K., Jiang, R., Molitor, J., Borevitz, J. O., Nordborg, M. & Marjoram, P. 2006. Association mapping with single-feature polymorphisms. *Genetics,* 173**,** 1125-33.

Kirkpatrick, M. 2010. How and why chromosome inversions evolve. *PLoS Biol,* 8.

Koornneef, M. & Meinke, D. 2010. The development of Arabidopsis as a model plant. *Plant J,* 61**,** 909-21.

Koornneef, M., van Eden, J., Hanhart, C. J., Stam, P., Braaksma, F. J. & Feenstra, W. J. 1983. Linkage map of Arabidopsis thaliana. *Journal of Heredity,* 74**,** 265-272.

Kovach, M. J., Sweeney, M. T. & McCouch, S. R. 2007. New insights into the history of rice domestication. *Trends Genet,* 23**,** 578-87.

Kovalchuk, I., Kovalchuk, O., Kalck, V., Boyko, V., Filkowski, J., Heinlein, M. & Hohn, B. 2003. Pathogen-induced systemic plant signal triggers DNA rearrangements. *Nature,* 423**,** 760-2.
Lambing, C., Franklin, F. C. & Wang, C. R. 2017a. Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiol,* 173**,** 1530-1542.
Lambing, C., Franklin, F. C. H. & Wang, C.-J. R. 2017b. Understanding and Manipulating Meiotic Recombination in Plants. *Plant Physiology,* 173**,** 1530-1542.
Landry, B. S., Kesseli, R. V., Farrara, B. & Michelmore, R. W. 1987. A Genetic Map of Lettuce *(Lactuca sativa L.*) with Restriction Fragment Length Polymorphism, Isozyme, Disease Resistance and Morphological Markers. *Genetics,* 116**,** 331-7.
Lawrie, N. M., Tease, C. & Hulten, M. A. 1995. Chiasma frequency, distribution and interference maps of mouse autosomes. *Chromosoma,* 104**,** 308-14.
Layer, R. M., Chiang, C., Quinlan, A. R. & Hall, I. M. 2014. LUMPY: A probabilistic framework for structural variant discovery. *Genome biology,* 15**,** R84-R84.
Li, X., Li, L. & Yan, J. 2015. Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize. *Nat Commun,* 6**,** 6648.
Lian, Q., Solier, V., Walkemeier, B., Durand, S., Huettel, B., Schneeberger, K. & Mercier, R. 2022. The megabase-scale crossover landscape is largely independent of sequence divergence. *Nat Commun,* 13**,** 3828.
Lim, E. C., Kim, J., Park, J., Kim, E. J., Kim, J., Park, Y. M., Cho, H. S., Byun, D., Henderson, I. R., Copenhaver, G. P.*, et al.* 2020. DeepTetrad: high-throughput image analysis of meiotic tetrads by deep learning in *Arabidopsis thaliana*. *Plant J,* 101**,** 473-483.
Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X.*, et al.* 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet,* 46**,** 1220-6.
Lin, Y. J. 1982. Temperature and chiasma formation in *Rhoeo spathacea var. variegata*. *Genetica,* 60**,** 25-30.
Liu, Z., Zhao, H., Yan, Y., Wei, M. X., Zheng, Y. C., Yue, E. K., Alam, M. S., Smartt, K. O., Duan, M. H. & Xu, J. H. 2021. Extensively Current Activity of Transposable Elements in Natural Rice Accessions Revealed by Singleton Insertions. *Front Plant Sci,* 12**,** 745526.
Lye, Z. N. & Purugganan, M. D. 2019. Copy Number Variation in Domestication. *Trends Plant Sci,* 24**,** 352-365.
Marand, A. P., Jansky, S. H., Zhao, H., Leisner, C. P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A. J., Veilleux, R. E.*, 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. & Jiang, J. 2019. Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *Plant Cell,* 31**,** 645-662.
McCouch, S. R., Kochert, G., Yu, Z. H., Wang, Z. Y., Khush, G. S., Coffman, W. R. & Tanksley, S. D. 1988. Molecular mapping of rice chromosomes. *Theor Appl Genet,* 76**,** 815-29.
Medvedev, P., Stanciu, M. & Brudno, M. 2009. Computational methods for discovering structural variation with next-generation sequencing. *Nature Methods,* 6**,** 13-20.
Melamed-Bessudo, C. & Levy, A. A. 2012. Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in Arabidopsis. *Proc Natl Acad Sci U S A,* 109**,** E981-8.
Mercier, R., Mezard, C., Jenczewski, E., Macaisne, N. & Grelon, M. 2015. The molecular biology of meiosis in plants. *Annu Rev Plant Biol,* 66**,** 297-327.
Mieulet, D., Aubert, G., Bres, C., Klein, A., Droc, G., Vieille, E., Rond-Coissieux, C., Sanchez, M., Dalmais, M., Mauxion, J. P.*, et al.* 2018. Unleashing meiotic crossovers in crops. *Nat Plants,* 4**,** 1010-1016.
Mieulet, D., Jolivet, S., Rivard, M., Cromer, L., Vernet, A., Mayonove, P., Pereira, L., Droc, G., Courtois, B., Guiderdoni, E.*, et al.* 2016. Turning rice meiosis into mitosis. *Cell Res,* 26**,** 1242-1254.
Miller, A. J. & Gross, B. L. 2011. From forest to field: perennial fruit crop domestication. *Am J Bot,* 98**,** 1389-414.
Morgan, C. H., Zhang, H. & Bomblies, K. 2017. Are the effects of elevated temperature on meiotic recombination and thermotolerance linked via the axis and synaptonemal complex? *Philos Trans R Soc Lond B Biol Sci,* 372.

Morgan, T. H. 1913. *Heredity and Sex*, Columbia University Press.
Morrell, P. L., Lundy, K. E. & Clegg, M. T. 2003. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare ssp. spontaneum*) despite migration. *Proc Natl Acad Sci U S A,* 100**,** 10812-7.
Moyers, B. T., Morrell, P. L. & McKay, J. K. 2018. Genetic Costs of Domestication and Improvement. *J Hered,* 109**,** 103-116.
Moyle, L. C. & Graham, E. B. 2005. Genetics of hybrid incompatibility between *Lycopersicon esculentum* and *L. hirsutum*. *Genetics,* 169**,** 355-73.
Munoz-Amatriain, M., Eichten, S. R., Wicker, T., Richmond, T. A., Mascher, M., Steuernagel, B., Scholz, U., Ariyadasa, R., Spannagl, M., Nussbaumer, T.*, et al.* 2013. Distribution, functional impact, and origin mechanisms of copy number variation in the barley genome. *Genome Biol,* 14**,** R58.
Nambiar, M. & Smith, G. R. 2016. Repression of harmful meiotic recombination in centromeric regions. *Semin Cell Dev Biol,* 54**,** 188-97.
Nieri, D., Di Donato, A. & Ercolano, M. R. 2017. Analysis of tomato meiotic recombination profile reveals preferential chromosome positions for NB-LRR genes. *Euphytica,* 213.
Ortiz-Barrientos, D., Engelstadter, J. & Rieseberg, L. H. 2016. Recombination Rate Evolution and the Origin of Species. *Trends Ecol Evol,* 31**,** 226-236.
Ortiz-Barrientos, D. & James, M. E. 2017. Evolution of recombination rates and the genomic landscape of speciation. *J Evol Biol,* 30**,** 1519-1521.
Osman, K., Algopishi, U., Higgins, J. D., Henderson, I. R., Edwards, K. J., Franklin, F. C. H. & Sanchez-Moran, E. 2021. Distal Bias of Meiotic Crossovers in Hexaploid Bread Wheat Reflects Spatio-Temporal Asymmetry of the Meiotic Program. *Front Plant Sci,* 12**,** 631323.
Otto, S. P. 2009. The evolutionary enigma of sex. *Am Nat,* 174 Suppl 1**,** S1-S14.
Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H. G., Tischfield, S. E., Zhu, X., Neale, M. J., Jasin, M., Socci, N. D.*, et al.* 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell,* 144**,** 719-31.
Pan, Q., Deng, M., Yan, J. & Li, L. 2017. Complexity of genetic mechanisms conferring nonuniformity of recombination in maize. *Sci Rep,* 7**,** 1205.
Pan, Q., Li, L., Yang, X., Tong, H., Xu, S., Li, Z., Li, W., Muehlbauer, G. J., Li, J. & Yan, J. 2016. Genome-wide recombination dynamics are associated with phenotypic variation in maize. *New Phytol,* 210**,** 1083-94.
Pawlowski, W. P., Golubovskaya, I. N. & Cande, W. Z. 2003. Altered nuclear distribution of recombination protein RAD51 in maize mutants suggests the involvement of RAD51 in meiotic homology recognition. *Plant Cell,* 15**,** 1807-16.
Penalba, J. V. & Wolf, J. B. W. 2020. From molecules to populations: appreciating and estimating recombination rate variation. *Nat Rev Genet,* 21**,** 476-492.
Peters, S. A. & Underwood, C. J. 2023. Technology-driven approaches for meiosis research in tomato and wild relatives. *Plant Reprod,* 36**,** 97-106.
Phillips, D., Jenkins, G., Macaulay, M., Nibau, C., Wnetrzak, J., Fallding, D., Colas, I., Oakey, H., Waugh, R. & Ramsay, L. 2015. The effect of temperature on the male and female recombination landscape of barley. *New Phytol,* 208**,** 421-9.
Protacio, R. U., Mukiza, T. O., Davidson, M. K. & Wahls, W. P. 2022. Molecular mechanisms for environmentally induced and evolutionarily rapid redistribution (plasticity) of meiotic recombination. *Genetics,* 220.
Quesneville, H. 2020. Twenty years of transposable element analysis in the *Arabidopsis thaliana* genome. *Mob DNA,* 11**,** 28.
Rahn, M. I. & Solari, A. J. 1986. Recombination nodules in the oocytes of the chicken, *Gallus domesticus*. *Cytogenet Cell Genet,* 43**,** 187-93.
Ramesh, M. A., Malik, S. B. & Logsdon, J. M., Jr. 2005. A phylogenomic inventory of meiotic genes; evidence for sex in Giardia and an early eukaryotic origin of meiosis. *Curr Biol,* 15**,** 185-91.
Ray, D. K., Mueller, N. D., West, P. C. & Foley, J. A. 2013. Yield Trends Are Insufficient to Double Global Crop Production by 2050. *PLoS One,* 8**,** e66428.

Razifard, H., Ramos, A., Della Valle, A. L., Bodary, C., Goetz, E., Manser, E. J., Li, X., Zhang, L., Visa, S., Tieman, D.*, et al.* 2020. Genomic Evidence for Complex Domestication History of the Cultivated Tomato in Latin America. *Mol Biol Evol,* 37**,** 1118-1132.

Robbins, M. D., Sim, S. C., Yang, W., Van Deynze, A., van der Knaap, E., Joobeur, T. & Francis, D. M. 2011. Mapping and linkage disequilibrium analysis with a genome-wide collection of SNPs that detect polymorphism in cultivated tomato. *J Exp Bot,* 62**,** 1831-45.

Rockmill, B., Voelkel-Meiman, K. & Roeder, G. S. 2006. Centromere-proximal crossovers are associated with precocious separation of sister chromatids during meiosis in *Saccharomyces cerevisiae*. *Genetics,* 174**,** 1745-54.

Ronspies, M., Schmidt, C., Schindele, P., Lieberman-Lazarovich, M., Houben, A. & Puchta, H. 2022. Massive crossover suppression by CRISPR-Cas-mediated plant chromosome engineering. *Nat Plants,* 8**,** 1153-1159.

Ross-Ibarra, J. 2004. The evolution of recombination under domestication: a test of two hypotheses. *Am Nat,* 163**,** 105-12.

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R. & Weigel, D. 2019. An Ultra High-Density *Arabidopsis thaliana* Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics,* 213**,** 771-787.

Rowan, B. A., Patel, V., Weigel, D. & Schneeberger, K. 2015. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3 (Bethesda),* 5**,** 385-98.

Roze, D. & Barton, N. H. 2006. The Hill-Robertson effect and the evolution of recombination. *Genetics,* 173**,** 1793-811.

Salome, P. A., Bomblies, K., Fitz, J., Laitinen, R. A., Warthmann, N., Yant, L. & Weigel, D. 2012. The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity (Edinb),* 108**,** 447-55.

Samuk, K., Manzano-Winkler, B., Ritz, K. R. & Noor, M. A. F. 2020. Natural Selection Shapes Variation in Genome-wide Recombination Rate in *Drosophila pseudoobscura*. *Curr Biol,* 30**,** 1517-1528 e6.

Sanchez-Moran, E., Santos, J. L., Jones, G. H. & Franklin, F. C. 2007. ASY1 mediates AtDMC1-dependent interhomolog recombination during meiosis in Arabidopsis. *Genes Dev,* 21**,** 2220-33.

Schmidt, C., Fransz, P., Ronspies, M., Dreissig, S., Fuchs, J., Heckmann, S., Houben, A. & Puchta, H. 2020. Changing local recombination patterns in Arabidopsis by CRISPR/Cas mediated chromosome engineering. *Nat Commun,* 11**,** 4418.

Schwander, T., Libbrecht, R. & Keller, L. 2014. Supergenes and complex phenotypes. *Curr Biol,* 24**,** R288-94.

Schwarzkopf, E. J., Motamayor, J. C. & Cornejo, O. E. 2020. Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. *BMC Genomics,* 21**,** 332.

Seck, P. A., Diagne, A., Mohanty, S. & Wopereis, M. C. S. 2012. Crops that feed the world 7: Rice. *Food Security,* 4**,** 7-24.

Seguela-Arnaud, M., Crismani, W., Larcheveque, C., Mazel, J., Froger, N., Choinard, S., Lemhemdi, A., Macaisne, N., Van Leene, J., Gevaert, K.*, et al.* 2015. Multiple mechanisms limit meiotic crossovers: TOP3alpha and two BLM homologs antagonize crossovers in parallel to FANCM. *Proc Natl Acad Sci U S A,* 112**,** 4713-8.

Serra, H., Lambing, C., Griffin, C. H., Topp, S. D., Nageswaran, D. C., Underwood, C. J., Ziolkowski, P. A., Seguela-Arnaud, M., Fernandes, J. B., Mercier, R.*, et al.* 2018. Massive crossover elevation via combination of HEI10 and recq4a recq4b during Arabidopsis meiosis. *Proc Natl Acad Sci U S A,* 115**,** 2437-2442.

Shen, C., Wang, N., Huang, C., Wang, M., Zhang, X. & Lin, Z. 2019. Population genomics reveals a fine-scale recombination landscape for genetic improvement of cotton. *Plant J,* 99**,** 494-505.

Sherman, J. D. & 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. & Levy, A. A. 2015. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in Arabidopsis. *Plant Cell,* 27**,** 2427-36.

Shirasawa, K., Isobe, S., Hirakawa, H., Asamizu, E., Fukuoka, H., Just, D., Rothan, C., Sasamoto, S., Fujishiro, T., Kishida, Y.*, et al.* 2010. SNP discovery and linkage map construction in cultivated tomato. *DNA Res,* 17**,** 381-91.

Si, W., Yuan, Y., Huang, J., Zhang, X., Zhang, Y., Zhang, Y., Tian, D., Wang, C., Yang, Y. & Yang, S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytol,* 206**,** 1491-502.

Sim, S. C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., Van Deynze, A., Hamilton, J. P., Buell, C. R., Causse, M., Wijeratne, S.*, et al.* 2012a. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One,* 7**,** e40563.

Sim, S. C., Van Deynze, A., Stoffel, K., Douches, D. S., Zarka, D., Ganal, M. W., Chetelat, R. T., Hutton, S. F., Scott, J. W., Gardner, R. G.*, et al.* 2012b. High-density SNP genotyping of tomato (*Solanum lycopersicum L.*) reveals patterns of genetic variation due to breeding. *PLoS One,* 7**,** e45520.

Smukowski, C. S. & Noor, M. A. 2011. Recombination rate variation in closely related species. *Heredity (Edinb),* 107**,** 496-508.

Stack, S. M., Anderson, L. K. & Sherman, J. D. 1989. Chiasmata and recombination nodules in *Lilium longiflorum*. *Genome,* 32**,** 486-498.

Sturtevant, A. H. 1921. A Case of Rearrangement of Genes in *Drosophila*. *Proc Natl Acad Sci U S A,* 7**,** 235-7.

Sudmant, P. H., Rausch, T., Gardner, E. J., Handsaker, R. E., Abyzov, A., Huddleston, J., Zhang, Y., Ye, K., Jun, G., Hsi-Yang Fritz, M.*, et al.* 2015. An integrated map of structural variation in 2,504 human genomes. *Nature,* 526**,** 75-81.

Swanson-Wagner, R. A., Eichten, S. R., Kumari, S., Tiffin, P., Stein, J. C., Ware, D. & Springer, N. M. 2010. Pervasive gene content variation and copy number variation in maize and its undomesticated progenitor. *Genome Res,* 20**,** 1689-99.

Tanksley, S. D., Ganal, M. W., Prince, J. P., de Vicente, M. C., Bonierbale, M. W., Broun, P., Fulton, T. M., Giovannoni, J. J., Grandillo, S., Martin, G. B.*, et al.* 1992. High density molecular linkage maps of the tomato and potato genomes. *Genetics,* 132**,** 1141-60.

Thompson, M. J. & Jiggins, C. D. 2014. Supergenes and their role in evolution. *Heredity (Edinb),* 113**,** 1-8.

Tieman, D., Zhu, G., Resende, M. F., Jr., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S., Taylor, M., Zhang, B.*, et al.* 2017. A chemical genetic roadmap to improved tomato flavor. *Science,* 355**,** 391-394.

Tiley, G. P. & Burleigh, J. G. 2015. The relationship of recombination rate, genome structure, and patterns of molecular evolution across angiosperms. *BMC Evol Biol,* 15**,** 194.

Underwood, C. J., Choi, K., Lambing, C., Zhao, X., Serra, H., Borges, F., Simorowski, J., Ernst, E., Jacob, Y., Henderson, I. R.*, 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.

Underwood, C. J. & Mercier, R. 2022. Engineering Apomixis: Clonal Seeds Approaching the Fields. *Annu Rev Plant Biol,* 73**,** 201-225.

Underwood, C. J., Vijverberg, K., Rigola, D., Okamoto, S., Oplaat, C., Camp, R., Radoeva, T., Schauer, S. E., Fierens, J., Jansen, K.*, et al.* 2022. A PARTHENOGENESIS allele from apomictic dandelion can induce egg cell division without fertilization in lettuce. *Nat Genet,* 54**,** 84-93.

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

Wang, Y., van Rengs, W. M. J., Zaidan, M. & Underwood, C. J. 2021. Meiosis in crops: from genes to genomes. *J Exp Bot,* 72**,** 6091-6109.

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O.*, et al.* 2007. A unified classification system for eukaryotic transposable elements. *Nat Rev Genet,* 8**,** 973-82.

Wijnker, E. & de Jong, H. 2008. Managing meiotic recombination in plant breeding. *Trends in Plant Science,* 13**,** 640-646.

Wijnker, E., van Dun, K., de Snoo, C. B., Lelivelt, C. L., Keurentjes, J. J., Naharudin, N. S., Ravi, M., Chan, S. W., de Jong, H. & Dirks, R. 2012. Reverse breeding in *Arabidopsis thaliana* generates homozygous parental lines from a heterozygous plant. *Nat Genet,* 44**,** 467-70.

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L.*, et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife,* 2**,** e01426.

Wilfert, L., Gadau, J. & Schmid-Hempel, P. 2007. Variation in genomic recombination rates among animal taxa and the case of social insects. *Heredity (Edinb),* 98**,** 189-97.

Wilson, J. Y. 2004. Temperature effect on chiasma frequency in the bluebell endymion nonscriptus. *Chromosoma,* 10**,** 337-354.

Xu, S., Clark, T., Zheng, H., Vang, S., Li, R., Wong, G. K., Wang, J. & Zheng, X. 2008. Gene conversion in the rice genome. *BMC Genomics,* 9**,** 93.

Yang, Z., Li, G., Tieman, D. & Zhu, G. 2019. Genomics Approaches to Domestication Studies of Horticultural Crops. *Horticultural Plant Journal,* 5**,** 240-246.

Yelina, N. E., Choi, K., Chelysheva, L., Macaulay, M., de Snoo, B., Wijnker, E., Miller, N., Drouaud, J., Grelon, M., Copenhaver, G. P.*, et al.* 2012. Epigenetic remodeling of meiotic crossover frequency in *Arabidopsis thaliana* DNA methyltransferase mutants. *PLoS Genet,* 8**,** e1002844.

Zhao, C., Liu, B., Piao, S., Wang, X., Lobell, D. B., Huang, Y., Huang, M., Yao, Y., Bassu, S., Ciais, P.*, et al.* 2017. Temperature increase reduces global yields of major crops in four independent estimates. *Proc Natl Acad Sci U S A,* 114**,** 9326-9331.

Zhou, Y., Zhang, Z., Bao, Z., Li, H., Lyu, Y., Zan, Y., Wu, Y., Cheng, L., Fang, Y., Wu, K.*, et al.* 2022. Graph pangenome captures missing heritability and empowers tomato breeding. *Nature,* 606**,** 527-534.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C.*, et al.* 2018. Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell,* 172**,** 249-261 e12.

Ziolkowski, P. A., Underwood, C. J., Lambing, C., Martinez-Garcia, M., Lawrence, E. J., Ziolkowska, L., Griffin, C., Choi, K., Franklin, F. C., Martienssen, R. A.*, et al.* 2017. Natural variation and dosage of the HEI10 meiotic E3 ligase control Arabidopsis crossover recombination. *Genes Dev,* 31**,** 306-317.

Chapter 2

Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing

Published as Roven Rommel Fuentes, Thamara Hesselink, Ronald Nieuwenhuis, Linda Bakker, Elio Schijlen, Willem van Dooijeweert, Sara Diaz Trivino, Jorn R. de Haan, Gabino Sanchez Perez, Xinyue Zhang, Paul Fransz, Hans de Jong, Aalt D.J. van Dijk, Dick de Ridder, Sander A. Peters. *Plant Journal* 88(2):159–178, 2020.

Abstract

Genome wide screening of pooled pollen samples from a single interspecific F1 hybrid obtained from a cross between tomato, *Solanum lycopersicum* and its wild relative, *Solanum pimpinellifolium* using linked read sequencing of the haploid nuclei, allowed profiling of the crossover (CO) and gene conversion (GC) landscape. We observed a striking overlap between cold regions of CO in the male gametes and our previously established F6 recombinant inbred lines (RILs) population. COs were overrepresented in non-coding regions in the gene promoter and 5'UTR regions of genes. Poly-A/T and AT rich motifs were found enriched in 1 kb promoter regions flanking the CO sites. Non-crossover associated allelic and ectopic GCs were detected in most chromosomes, confirming that besides CO, GC represents also a source for genetic diversity and genome plasticity in tomato. Furthermore, we identified processed break junctions pointing at the involvement of both homology directed and non-homology directed repair pathways, suggesting a recombination machinery in tomato that is more complex than currently anticipated.

2.1 Introduction

Meiotic recombination is a key fundamental biological process that ensures balanced chromosome distribution and the reshuffling of parental alleles into new combinations. In the current model, meiotic recombination is initiated by the formation of double-stranded DNA breaks (DSBs) that are processed to single stranded DNA, followed by single strand invasion into the intact sister chromatid or one of its non-sister chromatids. Upon break repair on the non-sister chromatids, DSBs are resolved either into crossovers (COs) or non-crossovers (NCOs). The resolution involves homology directed repair (HDR) by the ZMM and non-ZMM pathways leading to class I and class II COs respectively, and alternative pathways such as synthesis-dependent strand annealing (SDSA), dissolution of double Holiday junction (dHJ) and other pathways leading to NCOs (Mercier *et al.*, 2015). In general, CO involves reciprocal exchange of large DNA fragments, which can size up to whole chromosome arms. In contrast, a NCO usually comprises a small patch that is copied from the intact homologous chromosome (Mercier *et al.*, 2015). When the sister chromatid template is used as donor, the repair machinery restores the duplex. When the non-sister chromatid template is used as a donor, sequence divergence at the site of strand invasion may give rise to heteroduplex formation. The mismatch repair machinery recognizes and corrects meiotic heteroduplex DNA (Borts *et al.*, 2000), leading to a non-reciprocal exchange of genetic information or gene conversion (GC). Gene conversion can occur between donor and acceptor alleles of the same gene (allelic GC) or between different repetitive loci, the latter known as ectopic gene conversion (non-allelic GC) (Puchta, 2005; Duret and Galtier, 2009; Wijnker *et al.*, 2013; Trombetta *et al.*, 2016).

In plant breeding, the allelic reshuffling and gene conversion via meiotic recombination is a source of genetic diversity exploited by breeders that seek out new allele combinations for trait selection and crop improvement. As assorted alleles determine the inherent properties in progeny, it is important to precisely introgress chromosomal regions of interest, maintaining favorable genes yet preventing other regions harboring undesirable alleles to be recombined. Detection of genetic differences and monitoring of specific characters in segregating populations to assist trait selection is therefore crucial to successful breeding. Efficient detection and precise delineation of recombined alleles has seen many strategies. Microscopic analysis of meiotic events in tomato has provided valuable information on chromosome configurations (Moens, 1964; Havekes *et al.*, 1994; Anderson *et al.*, 2010), but the reduced throughput and low resolution limits genome wide and accurate measurements of recombined alleles. Currently, prevailing technologies to profile segregating alleles in populations include molecular marker-based screenings. In combination with high-throughput genome sequencing, such technologies are extremely powerful to delineate recombined alleles. Furthermore, such screenings provide valuable information on recombination frequency and genetic distances, which is used for genetic linkage map construction and marker assisted breeding. Previously, we profiled recombination events in a RIL population from a cross between tomato and its close relative S. *pimpinellifolium,* using genome

wide SNP analysis. The analysis provided insight into 'hot' and 'cold' regions of recombination, and genomic features involved in crossover recombination (Aflitos *et al.*, 2014; Demirci *et al.*, 2017; Demirci *et al.*, 2018). However, these methods require the production of offspring populations involving several generations and subsequent genetic marker screening for a large number of progeny plants (Paran *et al.*, 2004; van Os *et al.*, 2006; Ganal *et al.*, 2011; Baurens *et al.*, 2019). Often marker coverage is limited, causing low genetic resolution, and when integrated with physical maps, allows detection of crossovers in the kilo base range at best. Markers may even be absent leaving substantial genome portions uncovered and thus opaque to further genetic profiling. Furthermore, the accurate detection of COs and GCs by NGS technologies has long been hampered by short read length limitations, a relatively high base call error rate for long reads, and/or inadequate bioinformatics solutions. Therefore, more affordable quantitative methods to study CO and GC distributions are desired.

Here we present a method for profiling meiotic recombination in pollen nuclei from a single F1 interspecific hybrid between tomato and the closely related *S. pimpinellifolium*. The identification of polymorphisms marking COs and GCs is greatly improved by high-throughput long read sequencing with high base call accuracy, facilitating unambiguous read mapping and accurate identification of haplotype phases. We pinpoint COs and GCs at the SNP resolution level, using 10X Genomics linked read sequencing of gDNA from F1 pollen nuclei and subsequent detection of phase shifts in large gDNA molecules, providing further insight into gross similarities of recombination. Our sequence based approach can be applied to reliably profile meiotic recombination in a wide variety of crop species without the laborious and time-consuming production and screening of offspring populations, greatly benefitting introgression hybridization and precision breeding.

2.2 Experimental procedures

2.2.1 Selection of tomato accessions

An F1 plant was produced from an interspecific cross between *S. lycopersicum* cv. Heinz 1706 and *S. pimpinellifolium* CGN14498, which was used as the male and female parent respectively as described in Aflitos *et al.* (2014). The F1 plant was grown to flowering in a greenhouse under normal day night light (16h/8h) conditions at 24˚C.

2.2.2 High molecular weight DNA isolation from pollen

Mature pollen was released from flowers using a handheld milk frother device and collected in a 1.5 mL tube. Pollen were then stored at 4˚C at a relative humidity of 32% or immediately germinated after harvesting according to the protocol described by Lu *et al.* (2015) with modifications. Approximately 2-10 mg pollen were covered with a small gauze and positioned in a small petri dish containing saturated NA_2HPO_4 for 4-8 h at 25˚C without bringing the pollen into direct contact with the solution. The hydrated pollen were then incubated in 2.5 mL germination medium (20 mM

MES, 3mM $Ca(NO_3)_2.4H_2O$, 1 mM KCl, 0.8mM $MgSO_4.7H_2O$), 1.6 mM Boric acid (H_3BO_3), 24% ($^w/_v$) PEG4000, 2.5% ($^w/_v$) sucrose at pH 6.0 for 1.5 h at 25˚C, shaking at 90 rpm in the dark. Germination was monitored by assessing small samples with a light microscope at 20X magnification and was terminated when pollen tubes reached approximately 2X the diameter of the pollen grain. Germinated pollen were isolated from the medium with a 20 µM cell strainer (pluriStrainer, pluriSelect Life Sciences), collecting the flow through in a 50 mL tube by washing with 1 mL nuclei isolation buffer (0.3 M sucrose, 10 mM NaCl, 10 mM 2-(N-morpholine)-ethanesulphonic acid, 5 mM ethylenediamine-tetraacetic acid, 0.2 mM spermine, 0.5 mM spermidine, 0.2 mM phenylmethylsulphonyl fluoride, 5 mM dithiothreitol, and 0.2% (v/v) Triton-X-100, adjusted to pH 7.4 with 2 M NaOH). Nuclei were subsequently isolated from pollen according to Gao *et al.* (2015) with modifications. Subsequently, nuclei were collected by placing the 20 µM pluriStrainer with a connector ring on top of a 10 µM pluriStrainer and then put on top of a clean 50 mL tube. The 20 µM cell strainer was used to isolate the released nuclei and the non-germinated pollen from the germinated pollen, while the 10 µM cell strainer was used to separate the released nuclei from the non-germinated pollen. Next 1 mL nuclei isolation buffer was added to the pollen sample and nuclei were subsequently released by gentle homogenisation of the pollen tubes using a pestle for cell strainer (Sigma-Aldrich) at 60-100 rpm for 1 minute. Nuclei were rinsed from the 10 µM pluriStrainer and collected in a 50mL tube and centrifuged three times at 500g for 5 min to pellet the nuclei. The nuclei were then embedded in 2% liquid low melting point agar kept at 45˚C. The pellet was suspended in 10 µL nuclei isolation buffer, raising the temperature gently to 45˚C before adding 20 µL of 2% Low Melting Point (LMP) agarose at 45 ˚C and casting the solution into a mould. Plugs containing the nuclei can either be stored in TE buffer at 4°C, or directly be used for high molecular weight DNA isolation from the plugs according to the BioNano Genomics extraction protocol 'IrysPrep Plug Lysis Long DNA isolation' (30026D) as described by the manufacturer. The DNA concentration and molecule length was determined with a Qubit (Thermo Fisher Scientific) and a DNA Fragment Analyzer (Agilent).

2.2.3 10X Genomics library preparation and sequencing

The 10X Genomics libraries were constructed with the Chromium™ Genome Reagent Kits v2 (10X Genomics®) according to the Chromium™ Genome v2 Protocol (CG00043) as described by the manufacturer. Sequencing was carried out using Illumina HiSeq 2500 with 2 x 150bp paired-end reads to a genome coverage of approximately 100X. The LONG RANGER pipeline from 10X Genomics was used to process the sequencing output and align the reads to the tomato reference genome (The Tomato Genome Consortium, , 2012). The LOUPE genome browser summary view from 10X Genomics was used to retrieve the metrics on input DNA, sequencing, molecule length, coverage and gel beads in emulsion (GEM) performance. The number of F1 hybrid genome copies that were assessed was calculated using Lander-Waterman statistics (Lander and Waterman, 1988).

2.2.4 Sequence alignment and data processing

Illumina PE sequences from RIL parent *S. pimpinellifolium* CGN 14498 and from *S. lycopersicum* cv. Heinz from the 150 Tomato Genome Project (Aflitos *et al.*, 2014; Demirci *et al.*, 2017) were aligned to the *S. lycopersicum* cv. Heinz reference genome v3.0 using the read mapper tool from CLC Genomics workbench v11.0.1 (https://www.qiagenbioinformatics.com/). Reads were subsequently realigned to adjust for local mapping inconsistencies near insertions and deletions using the local read mapping alignment tool from CLC and the indel realignment GATK toolkit v.4. Homozygous and heterozygous single nucleotide variants (SNPs and indels) where then called using GATK UNIFIEDGENOTYPER and the fixed ploidy level variant detection algorithm from CLC separately. For GATK we used recommended parameter settings as described (https://software.broadinstitute.org/gatk/documentation/article.php?id=3225), whereas for the CLC fixed ploidy detection algorithm we used the following settings: Ploidy = 2; Required variant probability (%) = 80.0; Ignore positions with coverage above = 100,000; Ignore broken pairs = Yes; Minimum coverage = 8; Minimum count = 2; Minimum frequency (%) = 20.0; Neighborhood radius = 5; Minimum central quality = 20; Minimum neighborhood quality = 15; Relative read direction filter = Yes; Significance (%) = 1.0. Homozygous *S. pimpinellifolium* SNPs predicted by both GATK and CLC were then intersected and selected as segregating markers, excluding all positions found to be polymorphic between the tomato parent and the reference genome. Chromium barcoded sequences from the pollen sample were processed and aligned to the *S. lycopersicum* cv. Heinz v3.0 reference genome (https://solgenomics.net/organism/Solanum_lycopersicum/genome) using the LONG RANGER pipeline with default parameter settings. The output was processed by SAMTOOLS v1.3 *calmd* to generate the MD tags and then filtered per chromosome. Paired end (PE) reads without barcodes and molecule IDs, having a mapping quality below 60, showing an abnormal template length larger than 2 kb, or discordantly aligned to multiple chromosomes, were removed. The resulting alignment file was intersected with the list of *S. pimpinellifolium* homozygous SNPs to retrieve reads relevant to the phase shift detection. Reads having identical barcode (BX), molecule identifier (MI) and alignment start position were marked as clonal reads and discarded. Both CIGAR and MD tags of non-clonal reads were then parsed to assign the parental origin of the polymorphism that matched a segregating marker. The haplotype for each linked read molecule was determined using the pipeline illustrated in Supplementary Figure 2.1. Briefly, putative haploblock shifts were assigned to a linked read molecule if two adjacent SNP blocks, consisting of at least three uninterrupted and contiguous SNP markers, were assigned to different parents, i.e. ‘aaabbb’ or ‘bbbaaa’. CO recombination sites were then called taking the midpoint position between the two SNP markers flanking the phase shift. Molecules with candidate COs were subsequently filtered against tomato repeats (ITAG3.2_RepeatModeler_repeats), chloroplast insertions in the tomato nuclear genome, and tomato retrotransposons (tomato_retrotransposons_PRI_v2) retrieved from

ftp://ftp.solgenomics.net/genomes/Solanum_lycopersicum/repeats. To avoid aberrant haplotype shifts due to misalignment of tandem repeats aligned molecules overlapping at least 10% in length with repeats or candidate CO regions having a read coverage of flanking markers higher than 95% of all marker coverages across the genome were discarded (Supplementary Figure 2.2). To focus on the unambiguous occurrences of GC, molecules with multiple haplotype shifts were excluded. A block of at least two SNPs with the same parental origin flanked by SNPs from the other parent were identified as gene conversion, i.e. 'aaabbaaa' or 'bbbaabbb'. Molecules with phased regions flanked by two phase shifts were selected as candidate gene conversions (GCs). Candidate GCs were retained if (i) they were at most 400 bp long, supported by SNPs having a high base quality score (Phred-score > 30) matching with either parental allele, (ii) the supporting reads were not soft/hard-clipped, (iii) and did not overlap parental indel markers (to avoid false positives from aberrant indel marker calls). Remaining GCs containing indels that were not present in either homologous parental allele, but with supporting SNPs located within 3 bp distance of the indel, or GCs containing indels within homopolymeric regions of at least 4 bp were reported separately as they may have arisen from non-homologous directed repair mechanisms. Read alignments were visualized in the Intergrative Genome Browser (IGV) (Robinson *et al.*, 2011) and used to manually examine the linked read support for randomly selected COs and GCs, including a check on consistent BX and MI identifiers, and a verification of the SNP identity with a Phred score 30 or higher matching to either parental allele.

2.2.5 Sequence motif discovery and gene feature distribution

Sequences of 1kb, flanking both sides of the midpoint of high resolution (< 1 kb) CO regions (n=283) were selected for motif discovery with the MEME suite using default settings (Bailey *et al.*, 2009). Only overrepresented motifs at least 6 bp long, occurring in at least 50 CO regions, were reported. To find possibly enriched gene ontology terms, motifs with significant *E*-value < 0.05 were submitted to GOMO (Buske *et al.*, 2010).

2.2.6 Identification of hot and cold regions (CO-dense and CO-poor regions)

Using the R-package function 'density', we estimated the CO density using a Gaussian kernel (bandwidth = 20,000 nt) for every 1000 nt. Euchromatic and heterochromatic regions were delimited as previously described (Demirci *et al.*, 2017). We then computed the *p*-value for local maxima and minima detected by the R function 'turnpoints' based on density estimates in random recombination region sets generated by shuffling the observed euchromatic crossovers in each chromosome 10,000 times. Significant *p*-values of density peaks or hot regions were adjusted for multiple testing using a Bonferroni correction with a 0.05 cut-off threshold. For density pits or cold regions, we apply the *p*-value cut-off of 0.05 without Bonferroni correction and merged consecutive pits that are at most 10 kb apart. Only cold regions covering at least 100 kb were reported.

2.2.7 Genetic marker order

Sequences from linked reads containing haplotype shifts, marking a potential crossover recombination, were compared against marker sequences from the Tomato-EXPEN-2000.v1 genetic map (https://solgenomics.net) using BLASTN and then filtered using an identity threshold of 98%. Positional marker orders between the linked read genetic map and the tomato genetic maps were compared to assess marker collinearity.

2.3 Results

2.3.1 Detection of haplotype shifts and CO designation

Our aim was to localize recombination events at SNP resolution, taking advantage of linked-read sequencing of an F1 hybrid pollen grain population and the high density of informative polymorphisms between *S. lycopersicum* Heinz 1707 and *S. pimpinellifolium* CGN14498 parental genomes, allowing a detailed analysis of genomic sequences involved in CO recombination. To this end we constructed a linked read sequencing library from ~0.70 ng input gDNA and produced 781.3×10^6 paired end (PE) linked reads. With a median insert size of 391 bp, the data represented a mapped genome coverage of about 68X. Approximately 60% and 18% of the input gDNA molecules was longer than 20 kb and 100 kb, respectively, with a mean molecule length of 25.5 kb. We obtained 1.61×10^6 gel beads in emulsion (GEMS) with a mean DNA molecule length per GEM of 285 kbp, and an N50 linked reads per molecule (LPM) of 12. Given these GEM performance statistics, the molecule coverage for each genome copy was approximately 86%. Using the LONG RANGER pipeline from 10X Genomics, approximately 95.7% of the reads was successfully mapped to the tomato reference genome, leaving less than 0.14% of the reference bases uncovered. We detected polymorphisms in *S. pimpinellifolium* and *S. lycopersicum* PE reads that were aligned to the tomato SL3.0 reference genome, using SNP callers from CLC and GATK (McKenna *et al.*, 2010). Approximately 4.65 $\times 10^6$ CLC and 4.90 $\times 10^6$ GATK *S. pimpinellifolium* homozygous SNP calls were intersected, based on their corresponding genomic position, rendering 4.05 $\times 10^6$ high quality homozygous SNPs (Supplementary Figure 2.3). To remove possible base call errors and individual-specific sequence differences, approximately 210,000 SNPs between the tomato parent and the reference genome were excluded, leaving 3.84 $\times 10^6$ homozygous SNPs at a SNP density level of 1 per 210 bp. These SNP levels are in agreement with the SNP density level of 0.58% (~3.89×10^6 SNPs) as previously determined in genetic diversity study for the same parental lines (Aflitos *et al.*, 2014).

The SNP profiles in the genomic DNA molecules enabled us to assign COs according to the observed haplotype phase shifts. We first filtered ambiguous mappings, discarding discordant reads and molecules significantly overlapping with tomato genome repeats. Additionally, we removed recombinant molecules containing phase shifts with a skewed read coverage, that may have arisen from unknown repeat elements or copy number variation. Furthermore, we rejected multiple haplotype shifts, thereby avoiding cases where multiple putative COs or GCs

occurred in the same molecule. These filtering constraints rendered a total of 5,126 recombinant molecules on 12 chromosomes. We also anticipated on the occurrence of possible non-crossover associated (NCO-GC) and CO-associated gene conversions (CO-GCs) as possible products of recombination events. Indeed, a previous study on targeted recombination analysis in tomato reported on a single discontinuous GC of 5-6kb. This GC however, was observed in somatic tissue with artificially induced double stranded breaks (DSBs) by Cas9 (Filler Hayut *et al.*, 2017). In rice the average length of a gene conversion was reported to be 130 bp (Xu *et al.*, 2008). Considering that the observed average length of NCO-GCs and CO-GCs in *A. thaliana* ranges between 25 - 400 bp (Wijnker *et al.*, 2013), the additional filtering for tomato GCs left 3,819 recombinant molecules representing candidate CO regions with haploblock sizes larger than 400 bp, a resolution (i.e. distance between markers flanking a CO) ranging from 2 bp to 213 kb, and spanning distance (i.e. between the first and last markers within a recombinant molecule) (Figure 2.1A) from 1.9 kb to 297 kb (mean 55.9 kb, Figure 2.1B and S2.4A), respectively. The resolution of each molecule is negatively correlated with the coverage of reads overlapping a SNP marker (Pearson correlation coefficient r = -0.59, p-value < 2.2e-10) (Supplementary Figure 2.4B), suggesting that increasing the sequence coverage would likely further improve the resolution.

A single homologous CO event results into two reciprocal recombinant molecules. When mapped against the reference genome, such a pair of overlapping recombinant molecules would display reciprocal haplotype profiles and point to the same phase shift position. However, currently we cannot precisely establish the actual total number of CO events from recombinant molecules. Sizes may differ between molecules because of random DNA breaks during library preparation and sequencing. Furthermore, molecules are randomly primed, and when sequenced, likely will differ in their read profiles. Also, recombinant molecules originating from multiple gametes and generated by multiple CO events, may still map to overlapping positions in the reference genome. In addition, we cannot identify the recombinant chromosome pairs that originate from a single gamete. These limitations prevented to link recombinant molecules with overlapping CO regions to a distinct number of CO events that generated them. Nonetheless, to determine the minimum number of CO, we screened 3,819 CO regions, and assigned sets of recombinant molecules with overlapping CO regions to individual recombination events. This left a minimum of 3,169 putative CO recombination events. We then manually checked phase shift blocks for inconsistencies in 50 randomly selected candidate CO regions. Taking advantage of the Integrative Genome Viewer (IGV) displaying barcode (BX) and molecule identifier (MI) annotation tags called by LONG RANGER, we found a false positive rate of 2%. Furthermore, the flanking sequences for all randomly selected CO regions showed an unambiguous BlastN hit to their corresponding position in the tomato reference genome. We therefore anticipated assigned COs resulting from ambiguous read mappings less likely. We also found a significant overlap between COs in pollen and in RILs (Demirci *et al.*, 2017) (p-value = 8.44×10^{-23}, Fisher exact test) using *bedtools fisher*. While our stringent filtering method greatly improved

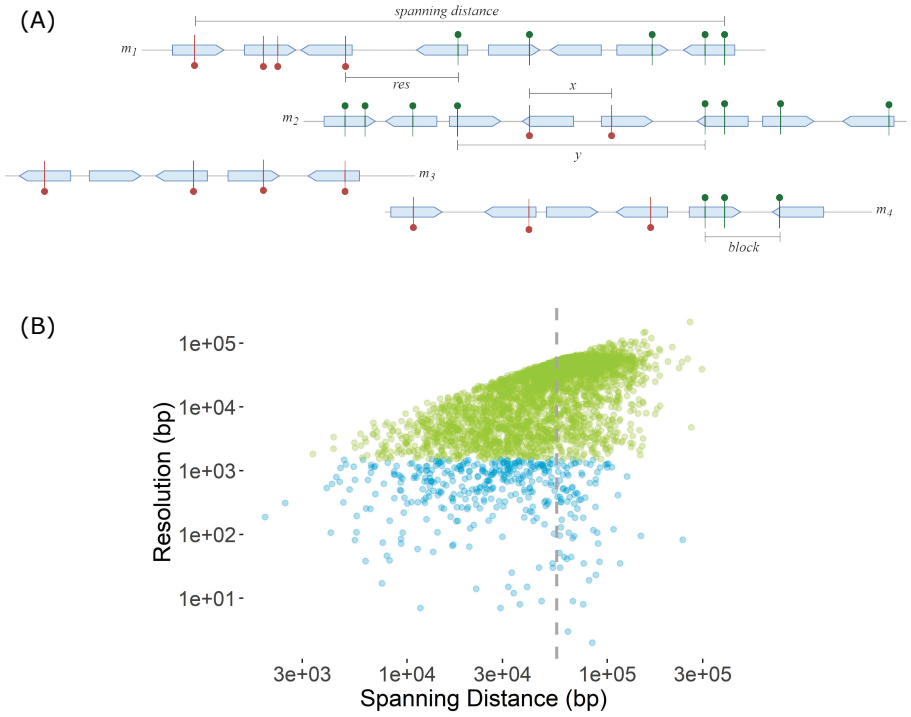


Figure 2.1. (A) Schematic representation of crossovers and gene conversions detected from haplotype shifts. Each molecule (m) consists of linked reads displayed as arrowed blocks connected by a horizontal line indicating their linkage. The red and green markers indicate the parental origin of the allele. Molecule m_1 represents a recombinant molecule showing phase shift marking a crossover. The resolution (res) of the crossover is defined as the length between two consecutive phase shifting markers, while spanning distance is the distance between the first and the last marker in the molecule. Molecule m_2 contains a gene conversion with two supporting markers with a minimal and maximal tract length denoted by x and y, respectively. For GC detection the maximum track length (x) was set to 400 bp. Molecule m_3 is non-recombinant, while in molecule m_4 the putative CO is discarded because its block length is shorter than 400 bp and displays only one phase shift. (B) Spanning distance versus crossover resolution with a CO represented either by a green or blue dot. Blue dots mark COs that were used to find correlations with sequence features. The dashed line represents the mean spanning distance of 55.9 kb.

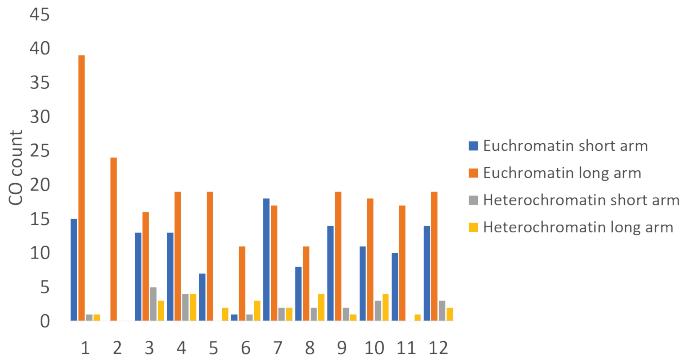
accurate detection, it likely decreased sensitivity though. Given the input gDNA, representing about 625 recombined F1 genome copies, and the previously reported average recombination rate in tomato of 1.8 per chromosome pair (Sherman and Stack, 1995), the theoretical number of COs in our pollen sample would be 13,500. The theoretical maximum number of COs, when compared to 3,169 detected COs in our pollen sample, suggests a considerable number of COs were left undetected.

However, considering that interspecies CO rates (i.e between the homeologous chromosomes of tomato and *pimpinellifolium*) can decrease to less than 20% compared to intraspecies recombination frequency (Canady *et al.*, 2006; Demirci *et al.*, 2017; Demirci *et al.*, 2018), the actual total number of COs in the F1 hybrid pollen presumably was lower. In the case of an interspecific hybrid, homoeologous chromosomes will to some extent be different in terms of genomic (repeat differences), and cytogenetic (heterozygosity for smaller structural chromosome variants) differences. Previously, chromosome structural differences have been revealed between *S. lycopersicum* and *S. pimpinellifolium*, including mismatched kinetochores foldbacks and other irregularities (Anderson *et al.*, 2010). Such differences between homeologous chromosomes lead to disturbances in the synapsis of the homoeologues, and hence may dramatically decrease formation of CO events. Depending on the species in the tomato clade used, genetic lengths of individual chromosomes have been found to vary substantially (Bernacchi and Tanksley, 1997). In addition, strong variation in recombination frequency has been shown dependent on the sex identity of the gametes. In plants, male and female meiosis differ considerably. Less recombination for male gametes was found in an F1 plant from a backcross to both the *S. lycopersicum* and *S. pennellii* parents (de Vicente and Tanksley, 1991). In contrast, Drouaud *et al.* (2007) found an elevated recombination frequency in male gametes from *Arabidopsis*, whereas Phillips *et al.* (2015) reported on a higher recombination frequency for female gametes of barley and showed recombination rates are temperature dependent. Thus, caution should be taken when comparing recombination frequencies and genetic map lengths such as for species in the tomato clade and for plants in general. Although the average frequency of ~0.42 COs per recombined F1 pollen chromosome represents 23% of the interspecies CO frequency, it nevertheless falls within the range as previously determined for an F6 RIL population (Demirci *et al.*, 2017).

2.3.2 Variation in CO frequencies

The frequency of COs between haploid nuclei can directly be compared since all genome copies originated from one F1 plant, eliminating a possible bias in recombination that would be due to differences in genetic background. To determine distributions, we assigned COs to 50 kb bins ordered along each of the 12 tomato chromosomes. We made a distinction between euchromatin and heterochromatin positions as previously established (Sherman and Stack, 1995; Demirci *et al.*, 2017). As expected, COs were not uniformly distributed. In all chromosomes a higher frequency was clearly observed for euchromatin portions, while pericentromeric heterochromatin regions were mostly devoid of COs (Figure 2.2A, 2.2B). As previously suggested, the low number of COs in pericentromere domains probably occurred within small euchromatin islands harbouring actively transcribed genes (Peters *et al.*, 2009; The Tomato Genome Consortium, , 2012). To further measure the propensities of the F1 pollen chromosomes toward COs, we compared sequences of 283 high-resolution CO regions (<1kb) against the tomato

(A)



(B)

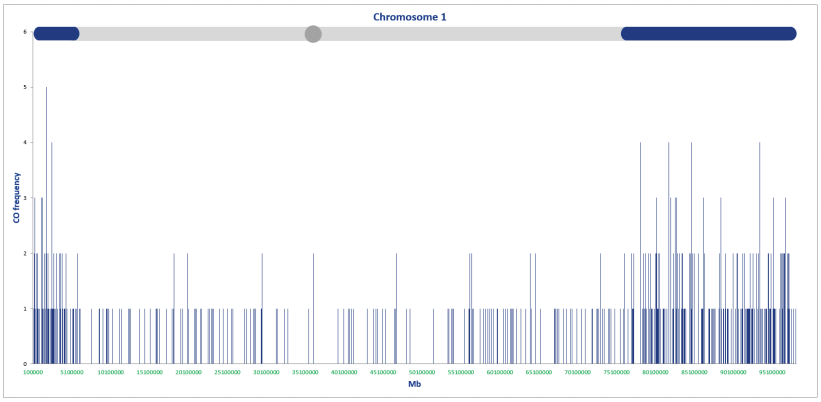


Figure 2.2. (A) Distributions of crossovers in the chromosome arms of pollen gametes. Positions of centromeres for division of chromosomes into long and short arms and euchromatin and heterochromatin sections were taken from Demirci *et al.* (2017). (B) Distribution of CO in pollen chromosome 1 in 100 kb bins. A schematic representation for tomato chromosome 1 is shown above the graph with euchromatin (dark blue) and heterochromatin (light grey), and centromere (dark grey).

bacterial artificial chromosome (BAC) fingerprint map. These high resolution COs do not share overlapping flanking markers and thus were likely generated by independent recombination events. The sequence comparison yielded 167 CO regions overlapping with tomato BACs anchored to the tomato EXPEN2000 genetic map. The order for 91 out 95 genetic markers associated with the anchored BACs coincided with the mapping order of all high-resolution COs to the tomato reference genome, indicating it is unlikely that these COs were called from aberrantly mapped

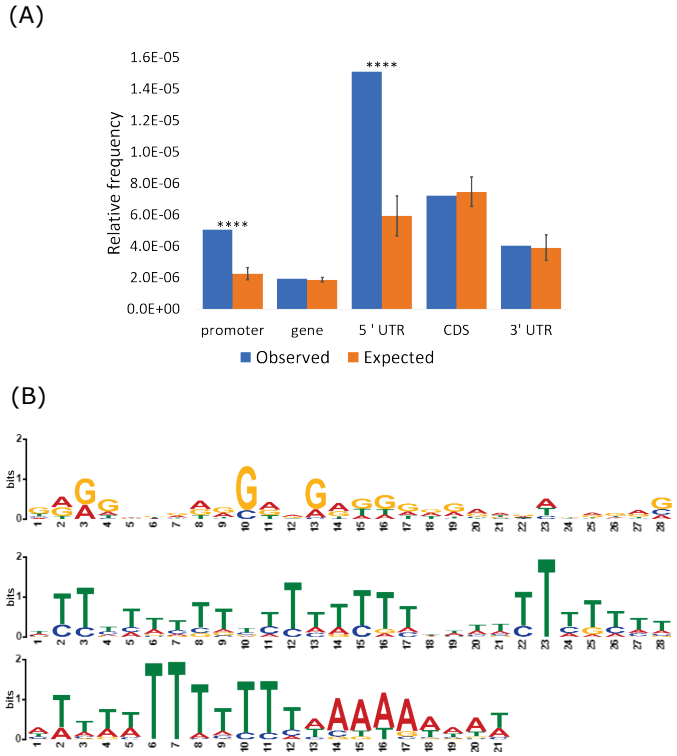


Figure 2.3. (A) Frequency distribution of COs observed in genic features. The observed and expected CO frequency per class is represented by blue and orange bars respectively. (B) Enriched sequence motifs in CO regions as discovered with MEME.

reads (Table S1). As expected, the frequency of COs in the heterochromatic regions is lower than in eurochromatic regions (Figure 2.2A, 2.2B). Furthermore, we observed a higher CO frequency for euchromatin regions toward the distal chromosome ends, except for chromosome 2 which showed a higher frequency towards the end of the long arm. These CO profiles in the F1 hybrid are in line with previously observed recombination profiles in tomato (Peters *et al.*, 2009; Peters *et al.*, 2012; Aflitos *et al.*, 2014; Demirci *et al.*, 2017; Demirci *et al.*, 2018).

2.3.3 Correlation with primary sequence and gene features and hot and cold CO regions

To obtain a detailed recombination profile, we assessed the sequence divergence and distribution of a subset of 400 CO regions (Figure 2.1B). In order to retrieve a sufficient number of datapoints, we relaxed the resolution threshold to 1500 bp, which was 500bp lower compared to the threshold for genomic sequence feature

enrichment detection. These COs regions are not overlapping and thus were also likely generated by independent recombination events. When comparing CO regions to the annotated reference genome, a permutation test showed a significantly higher rate of COs (p-value = 1.0 x 10^{-6}) in gene promoter regions (Figure 2.3A). This is in line with earlier results, reporting on COs biased toward regions upstream of transcription start sites (TSS) in tomato (Demirci *et al.*, 2017), rice (Wu *et al.*, 2003; Si *et al.*, 2015; Demirci *et al.*, 2018), maize (Rodgers-Melnick *et al.*, 2015), and *A. thaliana* (Wijnker *et al.*, 2013; Choi *et al.*, 2018). To check whether the enrichment in the 5'-UTR is an artefact of skewed SNP density causing the CO regions to extend from the promoter to the 5'-UTR, we focused on a subset COs, overlapping either a promoter or 5'-UTR and excluded COs spanning both regions. The additional permutation test on this subset confirmed the enrichment of COs in the promoter (z-score = 2.29, p-value = 0.0126) and 5'-UTR (z-score = 2.04, p-value = 0.0202) (Supplementary Figure 2.5). Moreover, using the MEME suite (Bailey *et al.*, 2009), we identified significantly overrepresented sequence motifs in 283 high resolution (below 1kb) CO regions. The motifs were composed of poly-AG, poly-T and poly(AT) sequences, occurring in 282 (E = 6.2 x 10^{-24}), 266 (E = 2.1 x 10^{-20}) and 280 (E = 9.0 x 10^{-10}) CO regions, respectively (Figure 2.3B). Poly-T and poly(AT) signatures are typically found at non-coding and promoter regions, and often at transcription factor (TF) binding sites (O'Malley *et al.*, 2016; Cherenkov *et al.*, 2018) and have been found associated with crossovers for example in *Arabidopsis* (Choi *et al.*, 2013; Wijnker *et al.*, 2013; Shilo *et al.*, 2015) and tomato (Demirci *et al.*, 2017; Demirci *et al.*, 2018). Previously, we also found a weak correlation with AG dinucleotides (Demirci *et al.*, 2017; Demirci *et al.*, 2018), further in line our observation of high CO incidence in gene promoter regions.

In addition, the observed distributions of recombinant sites along chromosomes suggested 'hot and cold regions'. Using CO density estimates, such regions with CO frequencies significantly deviating from the average were observed (Figure 2.4). We identified 23 hot regions in euchromatin portions except for chromosomes 3 and 6. Furthermore, using a *bedtools fisher* test a significant number (p-value=4.77 x 10^{-05}, Fisher exact test) of 15 out of 91 cold regions with a minimum length of 100kb significantly overlapped with previously reported cold regions in F6 RILs (Demirci *et al.*, 2017), further suggesting that profiles in pollen gametes are representative for CO profiles observed in plants.

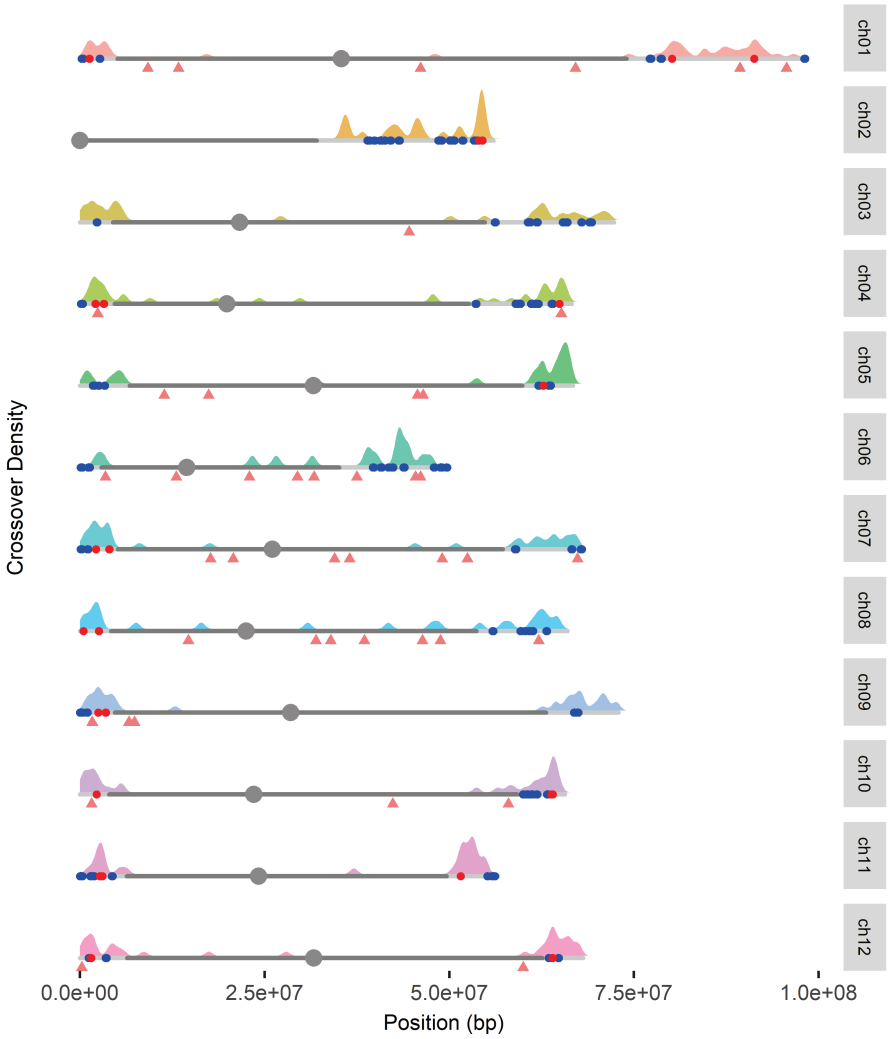


Figure 2.4. Distribution of recombination sites in F1 tomato hybrid chromosomes. Heterochromatin, euchromatin, and centromeres are depicted as dark and light grey horizontal bars, and dark grey circles respectively. CO density graphs are superimposed on each chromosome. Crossover-dense (hot) and crossover-depleted (cold) regions in F1 hybrid pollen are shown as red and blue dots along each chromosome, respectively. Chromosome positions for NCO-GCs are indicated by pink triangles.

2.3.4 Gene Conversion designation

Evidence has been presented for NCO-GC and CO-GCs in the animal and plant kingdom (Parniske *et al.*, 1997; Xu *et al.*, 2008; Duret and Galtier, 2009; Wijnker *et al.*, 2013; Trombetta *et al.*, 2016). When polymorphisms between alleles are involved, GCs in offspring alleles will be marked by SNPs, denoting the parental origin. We expected GCs also to occur during gametogenesis in our F1 tomato hybrid, though as far as we know a comprehensive genome wide analysis of GCs in interspecific tomato hybrids has not been presented until now. Both CO-GC and NCO-GC would consist of a small segment, containing a few SNPs marking a phase shift. Indeed, GCs observed in other plants have been reported in the 25-400bp range (Xu *et al.*, 2008; Wijnker *et al.*, 2013). However, without knowing the recombinant pairs of chromosomes, our screening method currently can only detect NCO-GCs, occurring in small phase blocks less than 400 bp (Supplementary Figure 2.6). After selecting candidate GCs < 400 bp and subsequent manual validation of supporting linked reads, we indeed detected 44 gene conversions within tracts between 35 to 369 bp. These 44 NCO-GCs do not share overlapping regions and thus were likely generated by independent recombination events. We prioritized on accuracy rather than sensitivity, by screening for small domains containing exactly one GC and at least 2 SNPs whose identity could be traced to the homologous donor allele. It is likely that the full chromosome complement is subject to GC and we indeed found footprints in all chromosomes. However, our stringent filtering did not leave GCs for chromosomes 2 and 11. We found read mappings at 3 GC sites showing high quality SNPs (Phred-score > 30) in addition to the expected polymorphisms that were detected between the parental homologous tomato and *pimpinellifolium* alleles. As these additional SNPs, occurring in a single read, could not be traced back to the parental alleles, we refer to them as 'non-parental' SNPs. The other linked reads generated from the same molecule flanking these non-parental SNPs mapped unambiguously to their homologous position. Using BLAST, we found for example a read with non-parental SNPs at a GC in chromosome 1, consisting of a segment specific for the target allele and another segment mapping to a nearby paralogous region in a reversed orientation approximately 250 bp downstream of the conversion (Supplementary Figure 2.7A,B). Considering that such clustered non-parental SNPs occurred in small segments identical to a parental paralogous allele, ectopic GC is the most parsimonious explanation. In addition, a separate set of 25 conversions that were reported as candidate GCs included insertions or deletions in a homopolymeric region. These indels, however, were not detected in the homologous or a paralogous allele. Currently, we cannot explain these indels by GC. We speculate they may have arisen as a result of illegitimate recombination such as non-homologous end-joining, or by polymerase slippage during replication, although sequencing errors of the homopolymeric region cannot be excluded.

2.4 Discussion

Here we present an algorithm that faithfully detects phase shifts marking both meiotic crossovers and non-crossover associated allelic and ectopic gene conversions in haploid male gametes of a *S. lycopersicum* x *S. pimpinellifolium* hybrid, based on 10X Genomics linked read sequencing data. The algorithm identifies the parental origin of homozygous SNPs from linked reads of pollen gametes that have been mapped to the SL3.0 Heinz reference genome. The previously established high quality reference genomes of the *S. lycopersicum* cv. Heinz and *S. pimpinellifolium* parental lines (Aflitos *et al.*, 2014) facilitated stringent filtering of false positives resulting from base calling or read mapping errors, and allowed for accurate CO and GC calls from phase shifts as presented here. Although, we couldn't establish the precise number of COs, because pairs of overlapping recombinant molecules do not have the same read profile and size, obscuring the identity of recombinant pairs in linked read data, we related non-overlapping recombinant molecules to a minimum number of CO events that generated them. Our analyses showed CO enrichment in promoter regions and the occurrence of specific sequence motifs consistent with previous results (Demirci *et al.*, 2017; Demirci *et al.*, 2018). In addition, we now also show an elevated CO frequency in the 5'-UTR regions. Homologous meiotic recombination has been shown to occur around the transcription start sites and nucleosome-depleted regions, suggesting that open chromatin is accessible to diverse cellular machineries (Pan *et al.*, 2011; Choi *et al.*, 2013; Wijnker *et al.*, 2013; Shilo *et al.*, 2015; Demirci *et al.*, 2017; Demirci *et al.*, 2018). A diversity analysis of 529 *Arabidopsis* TF-binding motifs showed that 25% have at least 4 consecutive A/T stretches (O'Malley *et al.*, 2016). These A/T rich motifs have been found overrepresented for example in TF-binding sites related to auxin response in *Arabidopsis* (Cherenkov *et al.*, 2018). A relationship between transcribed genes and the proximity to recombination hotspots was reported for yeast (Mancera *et al.*, 2008). Furthermore, AT-rich domains appear to disfavour nucleosome occupancy and recently have been shown to be involved in the formation of DSB and hotspots of recombination in *Arabidopsis* (Choi *et al.*, 2018). The enhanced recombination in both promoter and 5'–UTR regions might thus be explained by the local structural conformation of the DNA that is accessible for replication, recombination, and DNA transcription complexes as previously suggested (Gottipati and Helleday, 2009; Demirci *et al.*, 2017; Demirci *et al.*, 2018). The CO preference at promoter regions and CO hot and cold spots significantly overlapping with the regions of elevated and repressed CO in F6 RIL plants, further suggests that we faithfully detected COs and GCs in pollen gametes. It is possible that we have missed COs, especially in regions devoid of SNPs or with low coverage. However, since the observed recombination frequency in haploid spores was in the same range as previously observed for tomato hybrids (Demirci *et al.*, 2017), we speculate that the number of missed COs is relatively low. Nevertheless, approximately 5 $x10^6$ heterozygous SNPs and 4.7 x 10^5 indels have not been taken into consideration as potential CO markers yet, and may serve to further improve the resolution and sensitivity of CO and GC detection from linked read data. Additional adjustments, such as increased read coverage, and

allowing for multiple phase shifts per molecule, also might improve the sensitivity. Although, we currently do not consider recombination events from multiple phase shifts, molecules with complex phase patterns in ranges between 30 to 100kb were detected. We do not rule out complex GC in tomato, rather these observations hint at a tomato recombination machinery that is more complex than currently anticipated. Indeed, high-resolution mapping of meiotic crossovers and non-crossovers for example in yeast showed conversion tracts exhibiting complex patterns of genotype change, probably resulting from repair of multiple patches of heteroduplex, or alternatively, mismatch repair switching between conversion and restoration (Mancera *et al.*, 2008; Lesecque *et al.*, 2013). Furthermore, the involvement of tomato transposable elements as drivers of meiotic recombination cannot be ruled out, since we observed putative GC in recombinant molecules overlapping with repeats and transposon copies. Previously, meiotic recombination and GC induced by the *Mutator* (*Mu*) transposon in maize, resulted in deletions affecting gene function of the *knotted 1* gene (Mathern and Hake, 1997). A study on the trans-acting regulatory *MuDR* transposon revealed increasing GC and intragenic meiotic recombination in the vicinity of a nonautonomous *Mu* transposon (Yandeau-Nelson *et al.*, 2005). Furthermore, maize centromeres consisting of an abundant class *Ty3/Gypsy* like transposable elements, known as centomeric retroelements, apparently exhibited widespread GC (Shi *et al.*, 2010). We nevertheless currently prioritize on accuracy rather than sensitivity, excluding molecules overlapping with repeat and transposon copies to avoid aberrant recombination events called from misaligned recombinant molecules. This limitation could be overcome by lowering the molecule per GEM ratio and increasing the average length per molecule, sustaining faithful molecule mappings at a higher SNP levels, which then could shed more light on a possible role of transposons in meiotic CO recombination and CO associated and NCO associated GC in tomato.

Recombination profiling in plants has seen many techniques and approaches (Lambing and Heckmann, 2018). Among them are profiling strategies for male gametophytes which have relied on amplification of randomly amplified polymorphic DNA and single sequence repeat markers (Chen *et al.*, 2008), tetrad analysis of *Arabidopsis qrt* mutants (Copenhaver *et al.*, 2000; Wijnker *et al.*, 2013), whole genome amplification (WGA) of DNA of male spores (Aziz and Sauve, 2008) in combination with FACS sorting and KASP genotyping (Dreissig *et al.*, 2015), and recently also linked read sequencing of *Arabidopsis Col* X *Ler* F1 gametes (Sun *et al.*, 2019) and gametes from mouse and stickleback (Dreau *et al.*, 2019). While linked-read sequencing of pollen gametes and sperm cells enabled high-throughput profiling of COs, NCO related recombination events were not reported (Dreau *et al.*, 2019; Sun *et al.*, 2019). The genome coverages in these linked-read based profiling studies were considerably higher (~170X) compared to our genome coverage of approximately 100X. We could not retrieve information on the map-based coverages in the aforementioned studies, which prevented further comparison to our map-based coverage of 68X.

The strategy described in this study has both fundamental and applied research implications. Its high throughput enables profiling of pooled recombination events, without the need to screen an offspring population. Our screening requires a pollen sample from a single F1 offspring plant, which makes the technology applicable for screening many parental breeding line combinations. Such screenings could for example provide guidance for selection of compatible donor and recipient genomes to overcome linkage drag problems. Since our method requires a single generation, it also becomes practical to profile recombination in crops with relatively long generation periods. Crossover recombination can now be assessed relatively fast and cheap with unsurpassed precision down to a resolution of 2 bp as we have shown here. It thus becomes possible to assess the effects of genetic background, genome structure, and environmental conditions on recombination in greater detail and at a higher response level. This technology can thus boost breeding of advanced crops aimed at maintaining food security and safety, which is urgently needed in the face of the socio-economic and environmental challenges that traditional plant breeding currently faces. Furthermore, recombination profiles from genetically diverse panels that are tested under variable environmental conditions can be used in a machine learning approach to improve our current CO prediction models (Demirci *et al.*, 2018), providing further insight into gross similarities between species and species-specific aspects of recombination.

The current understanding of the recombination machinery in tomato is still incomplete as suggested by the complex patterns of exchanges that we detected. For example, besides 41 allelic GCs observed in both euchromatin and heterochromatin portions of the chromosomes, we detected 2 ectopic GCs in chromosome 1 and one in chromosome 3 in a reversed configuration relative to a nearby paralogous allele. We currently cannot explain the reversed configuration, but speculate that small inversions, as shown in mouse and *Solanum* hybrids, may cause local inversion loops in synapsed chromosomes (Borodin *et al.*, 1991; Sherman and Stack, 1995; Anderson *et al.*, 2010). A DSB in such inversion loops may perhaps be resolved into a gene conversion with an inverted orientation relative to the donor allele. Interestingly, a number of studies report on intrachromosomal DSB repairs in plants which, depending on a direct-repeat or inverted-repeat configuration of the involved locus, were thought to have been processed via the single strand annealing (SSA) or synthesis-dependent strand-annealing (SDSA) pathway, resulting either in a deletion or a gene conversion respectively (reviewed in Puchta, , 2005; Morrical, 2015). Furthermore, we observed 25 conversion tracts that cannot be explained by homologous CO or GC only. Careful breakpoint analysis showed that these recombinants contain indels that were not observed in the parental homologous or paralogous alleles. Our current analyses remain indecisive to the origin of these additional indels. These polymorphisms could have arisen for example by polymerase slippage during sequencing. Alternatively, these polymorphisms may have been introduced via alternative repair pathways such as non-homologous end joining (NHEJ) (Puchta, 2005). Indeed, micro-homology studies in human, have pointed at NHEJ and template-switching mechanisms generating small additional

indels at repaired breaks (Kidd *et al.*, 2010; Abyzov *et al.*, 2015). Interestingly, interrupted GC was shown to generate truncated *Mu* elements in maize. The interruption was proposed to result from hybridizing direct repeats, upon which the gap resolution would lead to a deletion (Mathern and Hake, 1997). In our data we occasionally detected GCs associated with small deletions in recombinant molecules overlapping transposon related repeats. Whether transposon related repeats around these putative recombination events in tomato mediated deletions at repaired breaks, currently remains speculative though. Furthermore, NHEJ has been reported in yeast, mammals, and *Drosophila* to repair double stranded breaks by direct re-joining without extensive processing, frequently introducing unequal sequence exchanges such as indels (Caridi *et al.*, 2017). The recruitment of the HR or NHEJ pathway seems to be dependent on the chromatin compaction as has been observed in human where DSB in heterochromatin appears more prone to NHEJ than HR mediated repair (Caridi *et al.*, 2017; Nisa *et al.*, 2019). NHEJ might partly explain the conversions that we observed in the pericentromeric heterochromatin of F1 tomato hybrid chromosomes. Moreover, combinations of HR and NHEJ mediated repairs have been reported to contribute to a complex mosaic of exchanges at repaired breaks (Puchta, 2005). Recently, it was demonstrated that DSB in repetitive 45S rDNA loci from *Arabidopsis*, during meiosis are shielded from HR and preferentially repaired via NHEJ (Sims *et al.*, 2019). Whether the occurrence of such additional indel markers can be attributed to aforementioned NHEJ mediated repair mechanism during meiotic recombination in tomato, currently remain speculative and will be subject to further study.

2.5 Acknowledgements

The work presented here is supported by the MEICOM Marie Sklodowska-Curie Innovative training Network (ITN), H2020-MSCA-ITN-2017 Horizon 2020 Grant agreement number 765212.

2.6 References

Abyzov, A., Li, S., Kim, D. R., Mohiyuddin, M., Stütz, A. M., Parrish, N. F., Mu, X. J., Clark, W., Chen, K., Hurles, M.*, et al.* 2015. Analysis of deletion breakpoints from 1,092 humans reveals details of mutation mechanisms. *Nature communications,* 6**,** 7256-7256.

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L.*, et al.* 2014. Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing. *Plant J,* 80**,** 136-48.

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

Aziz, A. N. & Sauve, R. J. 2008. Genetic mapping of *Echinacea purpurea* via individual pollen DNA fingerprinting. *Molecular Breeding,* 21**,** 227-232.

Bailey, T. L., Boden, M., Buske, F. A., Frith, M., Grant, C. E., Clementi, L., Ren, J., Li, W. W. & Noble, W. S. 2009. MEME SUITE : tools for motif discovery and searching. *Nucleic Acids Research,* 37**,** 202-208.

Baurens, F. C., Martin, G., Hervouet, C., Salmon, F., Yohome, D., Ricci, S., Rouard, M., Habas, R., Lemainque, A., Yahiaoui, N.*, et al.* 2019. Recombination and Large Structural Variations Shape Interspecific Edible Bananas Genomes. *Mol Biol Evol,* 36**,** 97-111.

Bernacchi, D. & Tanksley, S. D. 1997. An interspecific backcross of *Lycopersicon esculentum* x *L. hirsutum*: linkage analysis and a QTL study of sexual compatibility factors and floral traits. *Genetics,* 147**,** 861-77.

Borodin, P. M., Gorlov, I. P., Agulnik, A. I., Agulnik, S. I. & Ruvinsky, A. O. 1991. Chromosome pairing and recombination in mice heterozygous for different translocations in chromosomes 16 and 17. *Chromosoma,* 101**,** 252-8.

Borts, R. H., Chambers, S. R. & Abdullah, M. F. 2000. The many faces of mismatch repair in meiosis. *Mutat Res,* 451**,** 129-50.

Buske, F. A., Boden, M., Bauer, D. C. & Bailey, T. L. 2010. Assigning roles to DNA regulatory motifs using comparative genomics. *Bioinformatics,* 26**,** 860-6.

Canady, M. A., Ji, Y. & Chetelat, R. T. 2006. Homeologous recombination in *Solanum lycopersicoides* introgression lines of cultivated tomato. *Genetics,* 174**,** 1775-88.

Caridi, P. C., Delabaere, L., Zapotoczny, G. & Chiolo, I. 2017. And yet, it moves: nuclear and chromatin dynamics of a heterochromatic double-strand break. *Philos Trans R Soc Lond B Biol Sci,* 372.

Chen, P. H., Pan, Y. B. & Chen, R. K. 2008. High-throughput procedure for single pollen grain collection and polymerase chain reaction in plants. *J Integr Plant Biol,* 50**,** 375-83.

Cherenkov, P., Novikova, D., Omelyanchuk, N., Levitsky, V., Grosse, I., Weijers, D. & Mironova, V. 2018. Diversity of cis-regulatory elements associated with auxin response in Arabidopsis thaliana. *J Exp Bot,* 69**,** 329-339.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C.*, 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., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J.*, et al.* 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in *Arabidopsis thaliana* transposons and gene regulatory regions. *Genome Research,* 28**,** 532-546.

Copenhaver, G. P., Keith, K. C. & Preuss, D. 2000. Tetrad analysis in higher plants. A budding technology. *Plant Physiol,* 124**,** 7-16.

de Vicente, M. C. & Tanksley, S. D. 1991. Genome-wide reduction in recombination of backcross progeny derived from male versus female gametes in an interspecific cross of tomato. *Theor Appl Genet,* 83**,** 173-8.

Demirci, S., Peters, S. A., de Ridder, D. & van Dijk, A. D. J. 2018. DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J.*

Demirci, S., van Dijk, A. J., Sanchez Perez, G., Aflitos, S. A., de Ridder, D. & 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 Journal,* 89**,** 554-564.

Dreau, A., Venu, V., Avdievich, E., Gaspar, L. & Jones, F. C. 2019. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun,* 10**,** 4309.
Dreissig, S., Fuchs, J., Capal, P., Kettles, N., Byrne, E. & Houben, A. 2015. Measuring Meiotic Crossovers via Multi-Locus Genotyping of Single Pollen Grains in Barley. *PLoS One,* 10**,** e0137677.
Drouaud, J., Mercier, R., Chelysheva, L., Berard, A., Falque, M., Martin, O., Zanni, V., Brunel, D. & Mezard, C. 2007. Sex-specific crossover distributions and variations in interference level along *Arabidopsis thaliana* chromosome 4. *PLoS Genet,* 3**,** e106.
Duret, L. & Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet,* 10**,** 285-311.
Filler Hayut, S., Melamed Bessudo, C. & Levy, A. A. 2017. Targeted recombination between homologous chromosomes for precise breeding in tomato. *Nat Commun,* 8**,** 15605.
Ganal, M. W., Durstewitz, G., Polley, A., Berard, A., Buckler, E. S., Charcosset, A., Clarke, J. D., Graner, E. M., Hansen, M., Joets, J.*, et al.* 2011. A large maize (*Zea mays L.*) SNP genotyping array: development and germplasm genotyping, and genetic mapping to compare with the B73 reference genome. *PLoS One,* 6**,** e28334.
Gao, Y. B., Wang, C. L., Wu, J. Y., Wu, J., Zhou, H. S. & Zhang, S. L. 2015. A method to isolate male gametic nuclei from pear pollen tubes. *The Journal of Horticultural Science and Biotechnology,* 88**,** 313-319.
Gottipati, P. & Helleday, T. 2009. Transcription-associated recombination in eukaryotes: link between transcription, replication and recombination. *Mutagenesis,* 24**,** 203-10.
Havekes, F. W., de Jong, J. H., Heyting, C. & Ramanna, M. S. 1994. Synapsis and chiasma formation in four meiotic mutants of tomato (*Lycopersicon esculentum*). *Chromosome Res,* 2**,** 315-25.
Kidd, J. M., Graves, T., Newman, T. L., Fulton, R., Hayden, H. S., Malig, M., Kallicki, J., Kaul, R., Wilson, R. K. & Eichler, E. E. 2010. A human genome structural variation sequencing resource reveals insights into mutational mechanisms. *Cell,* 143**,** 837-47.
Lambing, C. & Heckmann, S. 2018. Tackling Plant Meiosis: From Model Research to Crop Improvement. *Front Plant Sci,* 9**,** 829.
Lander, E. S. & Waterman, M. S. 1988. Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics,* 2**,** 231-9.
Lesecque, Y., Mouchiroud, D. & Duret, L. 2013. GC-biased gene conversion in yeast is specifically associated with crossovers: molecular mechanisms and evolutionary significance. *Mol Biol Evol,* 30**,** 1409-19.
Lu, Y., Wei, L. & Wang, T. 2015. Methods to isolate a large amount of generative cells, sperm cells and vegetative nuclei from tomato pollen for "omics" analysis. *Front Plant Sci,* 6**,** 391.
Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. 2008. High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature,* 454**,** 479-85.
Mathern, J. & Hake, S. 1997. Mu element-generated gene conversions in maize attenuate the dominant knotted phenotype. *Genetics,* 147**,** 305-14.
McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M.*, et al.* 2010. The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res,* 20**,** 1297-303.
Mercier, R., Mezard, C., Jenczewski, E., Macaisne, N. & Grelon, M. 2015. The molecular biology of meiosis in plants. *Annu Rev Plant Biol,* 66**,** 297-327.
Moens, P. B. 1964. A new interpretation of meiotic prophase in *Lycopersicon esculentum* (tomato). *Chromosoma,* 15**,** 231-242.
Morrical, S. W. 2015. DNA-pairing and annealing processes in homologous recombination and homology-directed repair. *Cold Spring Harb Perspect Biol,* 7**,** a016444.
Nisa, M. U., Huang, Y., Benhamed, M. & Raynaud, C. 2019. The Plant DNA Damage Response: Signaling Pathways Leading to Growth Inhibition and Putative Role in Response to Stress Conditions. *Front Plant Sci,* 10**,** 653.
O'Malley, R. C., Huang, S. C., Song, L., Lewsey, M. G., Bartlett, A., Nery, J. R., Galli, M., Gallavotti, A. & Ecker, J. R. 2016. Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell,* 165**,** 1280-1292.

Pan, J., Sasaki, M., Kniewel, R., Murakami, H., Blitzblau, H. G., Tischfield, S. E., Zhu, X., Neale, M. J., Jasin, M., Socci, N. D.*, et al.* 2011. A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell,* 144**,** 719-31.

Paran, I., van der Voort, J. R., Lefebvre, V., Jahn, M., Landry, L., van Schriek, M., Tanyolac, B., Caranta, C., Chaim, A. B., Livingstone, K.*, et al.* 2004. An integrated genetic linkage map of pepper (*Capsicum spp.*). *Molecular Breeding,* 13**,** 251-261.

Parniske, M., Hammond-Kosack, K. E., Golstein, C., Thomas, C. M., Jones, D. A., Harrison, K., Wulff, B. B. & Jones, J. D. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. *Cell,* 91**,** 821-32.

Peters, S. A., Bargsten, J. W., Szinay, D., van de Belt, J., Visser, R. G., Bai, Y. & de Jong, H. 2012. Structural homology in the *Solanaceae*: analysis of genomic regions in support of synteny studies in tomato, potato and pepper. *Plant J,* 71**,** 602-14.

Peters, S. A., Datema, E., Szinay, D., van Staveren, M. J., Schijlen, E. G., van Haarst, J. C., Hesselink, T., Abma-Henkens, M. H., Bai, Y., de Jong, H.*, et al.* 2009. *Solanum lycopersicum cv.* Heinz 1706 chromosome 6: distribution and abundance of genes and retrotransposable elements. *Plant J,* 58**,** 857-69.

Phillips, D., Jenkins, G., Macaulay, M., Nibau, C., Wnetrzak, J., Fallding, D., Colas, I., Oakey, H., Waugh, R. & Ramsay, L. 2015. The effect of temperature on the male and female recombination landscape of barley. *New Phytol,* 208**,** 421-9.

Puchta, H. 2005. The repair of double-strand breaks in plants: mechanisms and consequences for genome evolution. *J Exp Bot,* 56**,** 1-14.

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G. & Mesirov, J. P. 2011. Integrative genomics viewer. *Nat Biotechnol,* 29**,** 24-6.

Rodgers-Melnick, E., Bradbury, P. J., Elshire, R. J., Glaubitz, J. C., Acharya, C. B., Mitchell, S. E., Li, C., Li, Y. & 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.

Sherman, J. D. & 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.

Shi, J., Wolf, S. E., Burke, J. M., Presting, G. G., Ross-Ibarra, J. & Dawe, R. K. 2010. Widespread gene conversion in centromere cores. *PLoS Biol,* 8**,** e1000327.

Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. & 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., Zhang, X., Zhang, Y., Zhang, Y., Tian, D., Wang, C., Yang, Y. & Yang, S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytol,* 206**,** 1491-502.

Sims, J., Copenhaver, G. P. & Schlogelhofer, P. 2019. Meiotic DNA Repair in the Nucleolus Employs a Nonhomologous End-Joining Mechanism. *Plant Cell,* 31**,** 2259-2275.

Sun, H., Rowan, B. A., Flood, P. J., Brandt, R., Fuss, J., Hancock, A. M., Michelmore, R. W., Huettel, B. & Schneeberger, K. 2019. Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nat Commun,* 10**,** 4310.

Tomato Genome, C. 2012. The tomato genome sequence provides insights into fleshy fruit evolution. *Nature,* 485**,** 635-41.

Trombetta, B., Fantini, G., D'Atanasio, E., Sellitto, D. & Cruciani, F. 2016. Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. *Sci Rep,* 6**,** 28710.

van Os, H., Andrzejewski, S., Bakker, E., Barrena, I., Bryan, G. J., Caromel, B., Ghareeb, B., Isidore, E., de Jong, W., van Koert, P.*, et al.* 2006. Construction of a 10,000-marker ultradense genetic recombination map of potato: providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics,* 173**,** 1075-87.

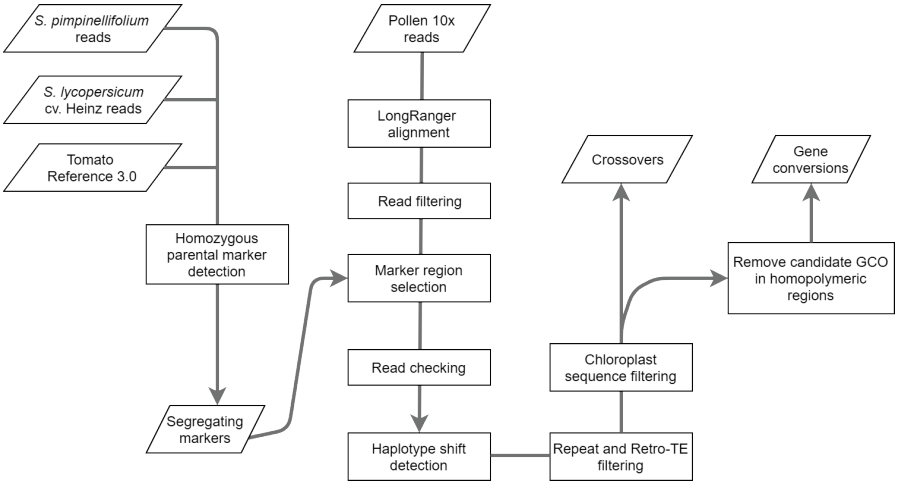
Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L.*, et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife,* 2**,** e01426.

Wu, J., Mizuno, H., Hayashi-Tsugane, M., Ito, Y., Chiden, Y., Fujisawa, M., Katagiri, S., Saji, S., Yoshiki, S., Karasawa, W.*, et al.* 2003. Physical maps and recombination frequency of six rice chromosomes. *Plant J,* 36**,** 720-30.

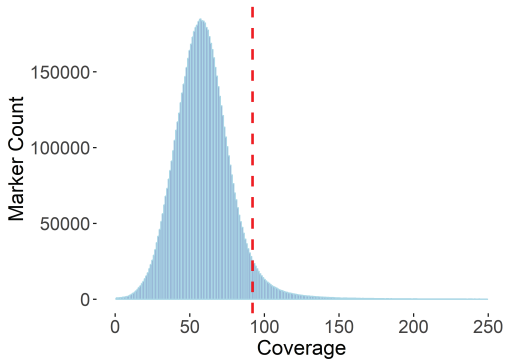
Xu, S., Clark, T., Zheng, H., Vang, S., Li, R., Wong, G. K., Wang, J. & Zheng, X. 2008. Gene conversion in the rice genome. *BMC Genomics,* 9**,** 93.

Yandeau-Nelson, M. D., Zhou, Q., Yao, H., Xu, X., Nikolau, B. J. & Schnable, P. S. 2005. MuDR transposase increases the frequency of meiotic crossovers in the vicinity of a Mu insertion in the maize a1 gene. *Genetics,* 169**,** 917-29.

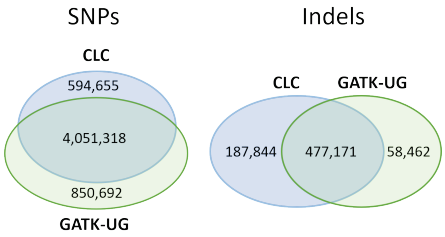
2.7 Supplementary figures



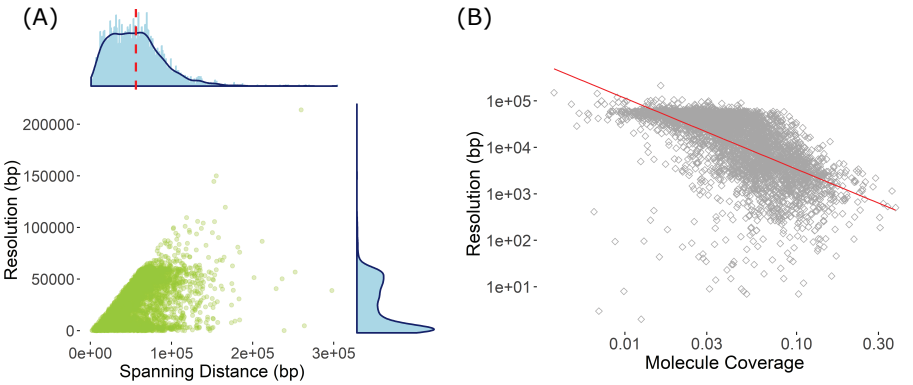
Supplementary Figure 2.1. Schematic representation of the crossover and gene conversion discovery pipeline. The pipeline includes the detection of segregating markers and the haplotyping of the 10x Genomics barcoded reads. Data sets used in this study are indicated in parallelograms, processing steps in rectangles and the data flow is indicated by connecting arrows. Annotation of repeat regions and sequences of retrotransposons and chloroplast are used to improve the detection accuracy. The final output consists of genomics coordinates of crossovers and gene conversions from the pooled pollen gametes.



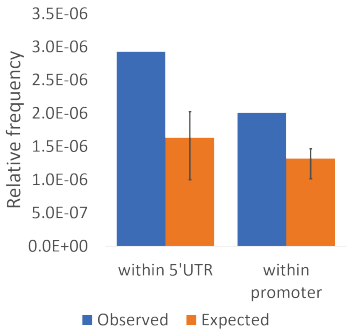
Supplementary Figure 2.2. Distribution of marker coverage and counts. Markers coverage was computed using SAMTOOLS as described. A coverage threshold is indicated by a dashed vertical line and used to remove markers in regions with read coverage higher than 95%.



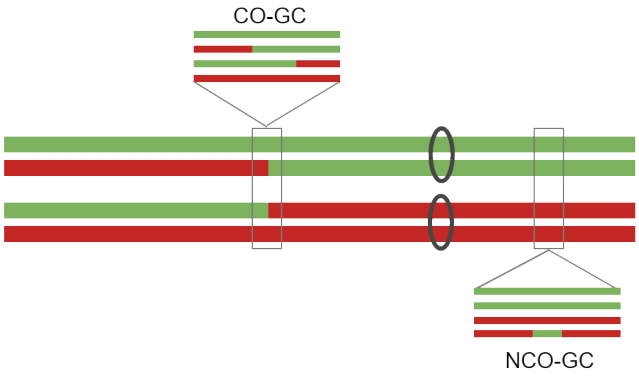
Supplementary Figure 2.3. Variant classes of *S. pimpinellifolium* compared to *S. lycopersicum* parental genomes. Both single nucleotide polymorphisms and indels were detected but only the number of homozygous variant calls are reported in the graph. The intersection indicates the number of polymorphisms shared between CLC and GATK-UG detected variants.



Supplementary Figure 2.4. (A) Spanning distance versus crossover resolution with CO represented by a green dot. The top graph shows distribution of spanning distance (blue) as defined in Figure 1. The dashed line represents the mean spanning distance of 55.9 kb. The right graph shows distribution of CO resolutions (orange). (B) Correlation of molecule coverage and resolution. The coverage of reads overlapping a marker is plotted against the resolution of each recombinant molecule (Pearson correlation coefficient r= -0.597; *p*-value < 2.2e-10).

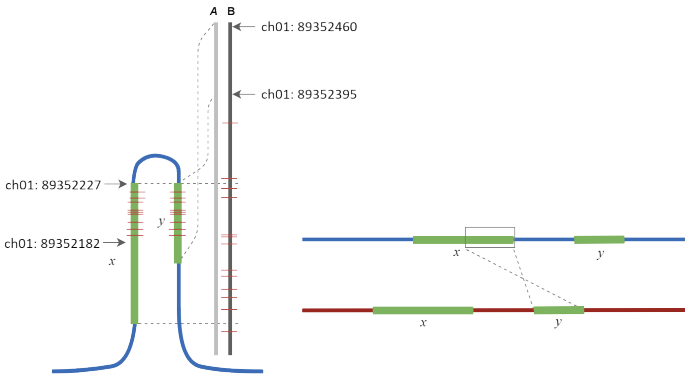


Supplementary Figure 2.5. Enrichment of COs in promoter and 5'-UTR regions.



Supplementary Figure 2.6. Schematic representation of recombinant chromosomes with COs and GCs. The green and red colors indicate the parental origin of genome segments while the oval shows the centromere. This figure shows recombinant chromosomes and both crossover-associated (CO-GC) and non-crossover-associated (NCO-GC) gene conversions in the enlarged sections. Phase shifts are positioned at block edges with alternating red and green colours. NCO-GCs can be detected in genomic regions smaller than 400bp flanked by phase shifts. The CO-GC is merged and positioned at the edge of a CO. The CO-GC and NCO-GC show a 3:1 segregation, while the flanking sections show a 2:2 segregation. Note that the segregation currently cannot be detected from linked read data. Currently, CO positions between pairs of recombinant chromosomes cannot be distinguished, preventing the detection of CO-GCs in linked reads.

(A)



(B)

```
> SL3.0ch01, Length=98455869
Score=230 bits (254), Expect=5e-59, Identities=141/150 (94%), Gaps=0/150 (0%) Strand=Plus/Plus

Query      1         GACAAATGACTTTTGCAAGTCATATTTCAATGAGTATTTTCAATTAAAATCAATTTTCTA  60
                     ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
SL3.0ch01  89352090  GACAAATGACTTTTGCAAGTCATATTTCAATGAGTATTTTCAATTAAAATCAATTTTCTA  89352149

Query      61        ATTAGTAAAAAAAATAGGAATAATGTCCAACTATCCCTCAACCTATGTCCGAAATCGCAGA  120
                     ||||||||||||||||||||||||||||||||||| ||||| ||||||| |||||   |||
SL3.0ch01  89352150  ATTAGTAAAAAAAATAGGAATAATGTCCAACTATCTCTCAATCTATGTCTGAAATATTAGA  89352209

Query      121       GACACACTTATAGTATACTAAGGTCCTATT  150
                     || ||| ||||| ||||||||||||||||
SL3.0ch01  89352210  GATACATTTATATTATACTAAGGTCCTATT  89352239


Score=102 bits (112), Expect=2e-20, Identities=63/66 (95%), Gaps=1/66 (2%), Strand=Plus/Minus

Query      86        CCAACTA-TCCCTCAACCTATGTCCGAAATCGCAGAGACACACTTATAGTATACTAAGGT  144
                     |||| || ||||||||||||||||||||||| ||||||||||||||||||||||||||||
SL3.0ch01  89352460  CCAAGTATTCCCTCAACCTATGTCCGAAATCCCAGAGACACACTTATAGTATACTAAGGT  89352401
                                                  *

Query      145       CCTATT  150
                     ||||||
SL3.0ch01  89352400  CCTATT  89352395
```

Supplementary Figure 2.7. (A) Possible configurations involved in an ectopic gene conversion in chromosome 1. SNPs within a sections ch01:89352182-89352227 and ch01:89352395-89352460 from *S. lycopersicum* cv Heinz and *S. pimpinellifolium* are represented by red horizontal lines. The *cis* configuration (left side graph) shows a hairpin section with an inverted paralogous gene approximately 200 bp downstream in the same chromosome. The *trans* configuration shows inverted paralogous genes on homologous chromosomes (right side graph). The transfer of the paralogous section from the same chromosome (cis) or between homologous chromosomes (trans) would lead to an inverted segment in the ectopic GC. Currently we cannot determine whether a *cis* or *trans* configuration was used by the repair mechanism due to insufficient coverage of flanking markers. (B) Sequence alignment of a linked read to repetitive alleles marking an ectopic GC in chromosome 1. In the upper part the linked read consists of domain 1 identical to a homologous tomato allele (highlighted in yellow), and domain 2 containing SNPs (highlighted in light blue). In the lower part domain 2 of the linked read matches to a downstream paralogous allele. The asterisk at position 89352429 marks a heterozygous SNP (C**C**C/C**G**C) in the tomato parental genome. Note that the alignment for domain 2 to the paralogous allele is inverted.

Chapter 3

Phasing HiFi reads to detect high-resolution crossovers in a pollen gamete pool

Roven Rommel Fuentes, Aalt D.J. van Dijk, Dick de Ridder, Sander A. Peters

Abstract

Meiotic recombination has been widely studied in plant breeding, as it generates genetic diversity and enables the generation of new haplotypes with desired phenotypes. Computational methods have been developed to profile recombination events and the resulting crossovers (COs) and gene conversions from offspring populations. We previously developed a pipeline to detect COs and gene conversions in linked reads generated from a pool of gametes and applied this to study recombination in an F1 interspecific tomato hybrid. Building on this, we now report the use of PacBio high-fidelity (HiFi) long reads as an alternative sequencing technology. We sequenced pollen gametes from a cross between *S. lycopersicum* and *S. pimpinellifolium* and found a significant increase in the resolution of COs compared to the previous version using linked reads, allowing a more precise characterization of recombination regions. However, sequencing errors in homopolymeric and microsatellite tracts produce false haplotype shift signals, limiting the performance for detecting gene conversions. Overall, we present an alternative solution to identify recombination events from F1 gametes.

3.1 Introduction

Ongoing advances in plant breeding techniques to address issues in food security and climate change have been accompanied by an increasing interest in understanding and manipulating meiotic recombination, a biological process that facilitates the reciprocal exchange of genetic material between homologous non-sister chromatids (Mercier *et al.*, 2015). During meiosis, double strand breaks (DSBs) are introduced in the genome which may be repaired to crossovers (COs), or to non-crossovers (NCO) such as gene conversions (GCs) depending on the repair template used (Mercier *et al.*, 2015). This process generates diversity through new combinations of parental alleles in progenies, potentially conferring novel phenotypes. To examine genomic recombination, earlier studies have used cytogenetic techniques such as counting bivalents or chiasmata (Stack *et al.*, 1989) and have produced genetic maps for different species (Anderson *et al.*, 2003; Francis *et al.*, 2007). With the advent of next generation sequencing, finer-scale CO events were profiled from offspring populations, allowing more extensive exploration of the factors influencing recombination (Huang *et al.*, 2009; Qi *et al.*, 2009; Wijnker *et al.*, 2013; Demirci *et al.*, 2017). However, such methods are laborious and costly. Recently, we and others implemented new methods to detect crossovers in linked reads generated from pools of gametes, uncovering high-resolution patterns of recombination in various organisms with less cost and labor (Dreau *et al.*, 2019; Sun *et al.*, 2019; Campoy *et al.*, 2020; Fuentes *et al.*, 2020). Linked read technologies produce groups of short reads that are sequenced from a long DNA fragment and tagged with a unique barcode. By comparing the alleles in the linked reads to the parental genomes, recombinant haplotypes can be detected. Cataloging recombination events in gamete pools in this way is cost-effective and efficient, as these contain numerous genomes with hundreds of DSBs that may have produced COs or GCs.

The method we developed to detect COs from pollen allowed us to analyze the landscape of recombination in tomato and identify CO-associated genome features (Fuentes *et al.*, 2020). The high throughput and accuracy of this method enable the characterization of local CO pattern with resolution comparable to the sequencing of large offspring population. Unfortunately, the 10X Genomics linked read technology that was used is no longer commercially available, requiring an alternative sequencing technology to phase long DNA molecules. In this study, we explore the use of PacBio high-fidelity (HiFi) long reads for the detection of COs and GCs in a pool of pollen gametes from an interspecific tomato hybrid. We improved our data analysis pipeline, allowing detection of COs associated with transposable elements and other repeats. Using this new pipeline, we investigate whether use of long reads produces better detection of COs and GCs than linked reads. We also analyze sequencing errors of HiFi reads in homopolymeric and microsatellite tracts that cause detection issues for both COs and GCs.

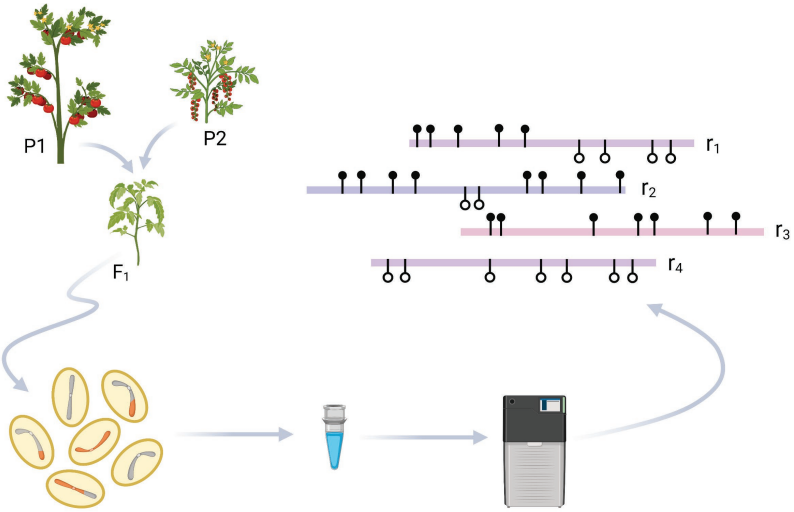


Figure 3.1. CO detection using HiFi reads. An F_1 plant was produced from a cross between tomato (P1) and its wild relative (P2). From the F_1 plant, a pool of pollen was collected and germinated to extract high-molecular weight DNA. The DNA was then sequenced using PacBio HiFi. Long reads were phased based on a set of segregating markers (black and white circles) between the parental genomes. Reads r_1 and r_2 illustrate CO and GC patterns, respectively, while r_3 and r_4 shows no recombination. This figure was created with BioRender.com.

3.2 Results

3.2.1 HiFi reads yield high SNP density

As proposed in Fuentes *et al.* (2020), polymorphisms between parental genomes can be used as markers to detect meiotic recombination events in a pool of pollen gametes (Figure 3.1). That study used 10X Genomics linked reads; here, we explore PacBio HiFi (Wenger *et al.*, 2019) as an alternative sequencing technology for the detection of meiotic recombination events. We sequenced the parental genomes *S. lycopersicum* (SLL) and *S. pimpinellifolium* (SP) using HiFi and aligned the resulting reads against the SL4.0 (*S. lycopersicum*) reference genome (Hosmani *et al.*, 2019). This yielded a coverage of 37x and 32x, respectively, and allowed to detect 3.2 million single nucleotide polymorphisms (SNPs) segregating the parental genomes, using the filtering criteria as previously reported (Fuentes *et al.*, 2020). In comparison with the 10X-based SNP calls, we found fewer SNPs in repeat and transposable element (TE) regions, suggesting that HiFi reads may yield fewer false

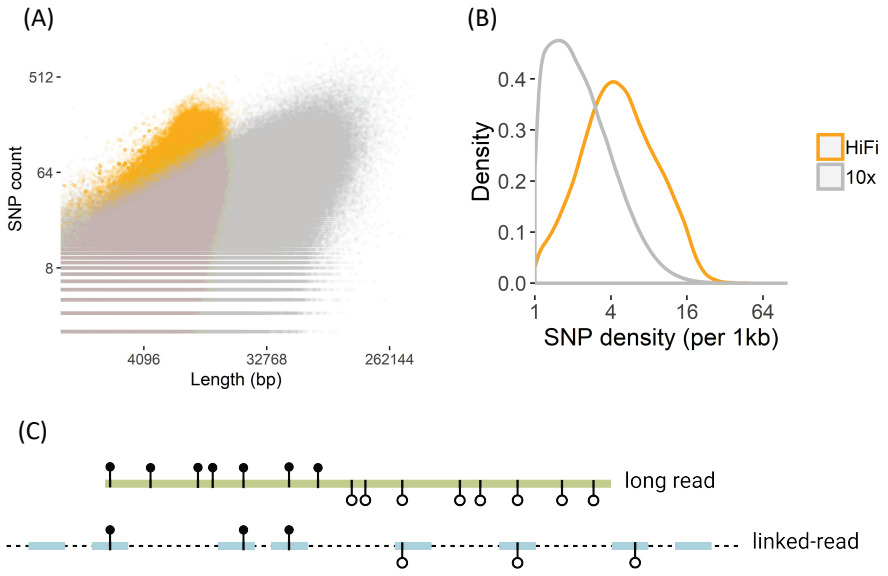


Figure 3.2. HiFi and 10X Genomics linked reads. a) Length and SNP counts. b) SNP density of HiFi and linked reads. c) Phasing of reads using the SNP markers. The black and white circles represent the parental origin. The broken line indicates the DNA molecule from which the short linked reads are sampled.

positive calls than short Illumina reads. The lower number of false SNPs may have resulted from less ambiguous mappings of long reads.

We then sequenced a pool of pollen gametes from an interspecific F1 cross between SLL and SP with HiFi and aligned the resulting reads against the reference genome. We retrieved reads that overlap markers and inferred the parental origin of the markers in each read. To examine whether markers for CO detection differ between the sequencing technologies, we compared the lengths and segregating SNP counts of HiFi reads and 10X linked reads. As shown in Figure 3.2A, HiFi reads have an average length of 7.5kb, shorter than linked read molecules. Furthermore, HiFi reads have a median density of 4.3 SNPs/kb compared to 1.5 SNPs/kb in the far more sparsely sequenced linked reads (Figure 3.2B, 3.2C). Both the density of markers between the parental genomes and of SNPs overlapped by reads directly influence the resolution of crossovers detected by our method. We defined resolution as the inverse of the distance between the SNPs immediately flanking a CO region. By using longer reads that can fully cover a region, resolution is maximized and only limited by the distance between the available markers flanking the crossover (CO) site.

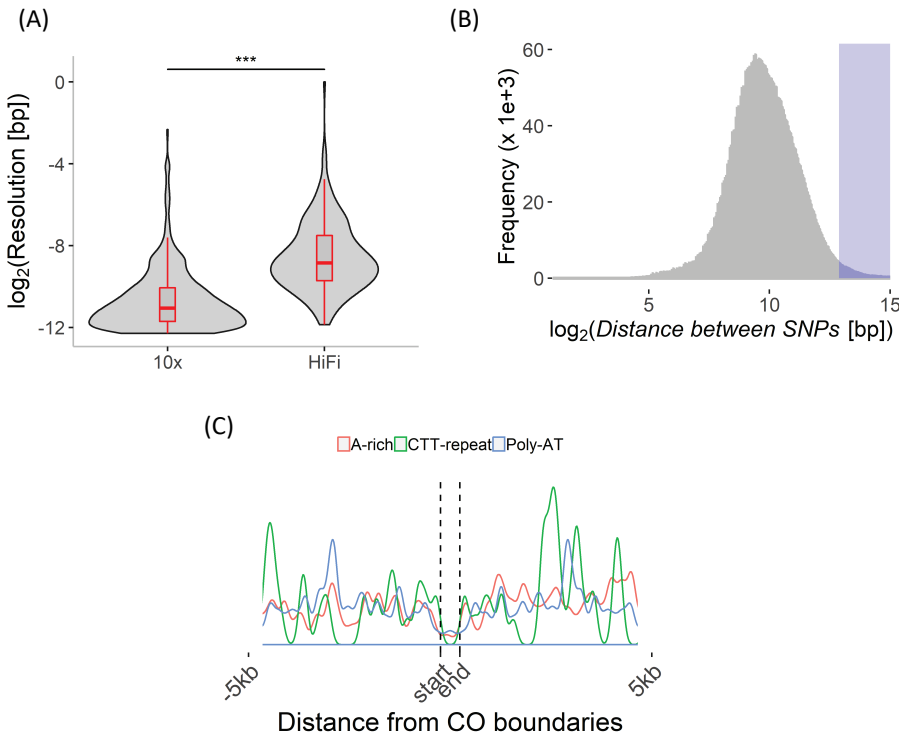


Figure 3.3. Higher crossover resolution when using HiFi reads. a) Resolution of COs detected using 10X and HiFi reads. The asterisk indicates the significant difference in resolution ($P < 2.2 \times 10^{-16}$) b) Distance between neighboring SNPs across the whole genome. The blue region marks the distances larger than the average read length in our HiFi data. c) Distances of known CO-associated motifs from COs detected using HiFi reads.

We modified our linked read-based pipeline to allow long read input and we subsequently phased the HiFi reads based on the parental origins of the marker alleles (Figure 3.1). Afterwards, we checked the reads with a mix of alleles from the two parents to determine whether recombination took place. Depending on the haplotype shifts, we discard artifacts due to misalignments and random mutations and subsequently differentiate COs from GCs. One crucial characteristic of HiFi reads that severely complicates this is the higher number of sequencing errors w.r.t. linked reads (Illumina sequencing), mostly single-base insertions (Wenger *et al.*, 2019). These are mostly located in homopolymers and microsatellites and result in misleading phasing signals. Since our method relies on single, individual HiFi reads to infer a haplotype shift, such signals cannot be corrected by taking multiple reads into account and thus may lead to inaccurate results in certain regions. To address this, we masked genomic regions with homopolymers and short repeats before

finding reads with CO signals. However, this exclusion of regions reduces the number of markers, consequently affecting the sensitivity of detecting COs, especially in low-complexity regions.

3.2.2 Crossovers can be detected at high resolution

After phasing the reads, we found 229 COs in the pollen data, demonstrating that our method can be adapted to new sequencing technologies. Similar to previous observations (de Haas *et al.*, 2017; Demirci *et al.*, 2017; Fuentes *et al.*, 2020), the majority of COs are located near or within genes (Fisher's exact test; $P = 9.5 \times 10^{-42}$). COs detected based on HiFi reads also show significant overlap with COs detected with 10X linked reads (Fisher's exact test; $P = 6.0 \times 10^{-3}$), COs found using Illumina short reads sequenced from recombinant inbred lines (RILs) ($P = 2.5 \times 10^{-7}$), and with historical recombination hotspots ($P = 1.9 \times 10^{-2}$), implying that PacBio HiFi sequencing provides a viable alternative means of pollen-based profiling of meiotic recombination (Demirci *et al.*, 2017; Fuentes *et al.*, 2020). The median resolution of COs detected using HiFi is 2.0×10^{-3}, which is a 4-fold improvement from 4.8×10^{-4} for COs detected using linked reads (Figure 3.3A). However, the shorter length of HiFi reads (<10kb) and the high sequencing error compared to 10X linked reads reduces sensitivity. We indeed find genomic regions that may contain CO but are not discoverable in our approach, because the distances between markers are larger than the average HiFi read length, which is limited due to the need for amplification in library preparation (Figure 3.3B).

Despite the effect of limited read lengths, the improved resolution offered by HiFi reads allows for more accurate comparison of CO occurrences with occurrences of genomic elements linked with CO localization such as genes, TEs, motifs etc. The locations of specific sequence motifs relative to CO regions (Figure 3.3C) are consistent with our earlier findings (Demirci *et al.*, 2017). We observe peaks of CTT-repeats flanking the CO sites; while this motif was previously linked with meiotic recombination (Choi *et al.*, 2013; Wijnker *et al.*, 2013; Shilo *et al.*, 2015; Demirci *et al.*, 2017), we here find it located near, but not within the CO region. This illustrates the benefit of high resolution CO data in precise examination of recombination sites.

3.2.3 Gene conversions are hard to detect

We next explored the use of HiFi reads to detect gene conversion (GC) events. GCs are difficult to detect in sparse linked read data of pooled gametes. Similar to CO detection, we examined reads for a short span of markers with different parental origin compared to the flanking regions (Fuentes *et al.*, 2020). While we detected some GC events (Figure 3.4), we also found many false positives caused by sequencing errors in HiFi reads (Supplementary Figure 3.1). Although masking homopolymers helps to reduce the issue, some regions are worsened by the misalignment of false insertion to the regions flanking the homopolymeric tracts, further extending the problematic regions. Excluding these error-prone regions reduces the number of SNPs available to detect GCs, which mostly span only a few

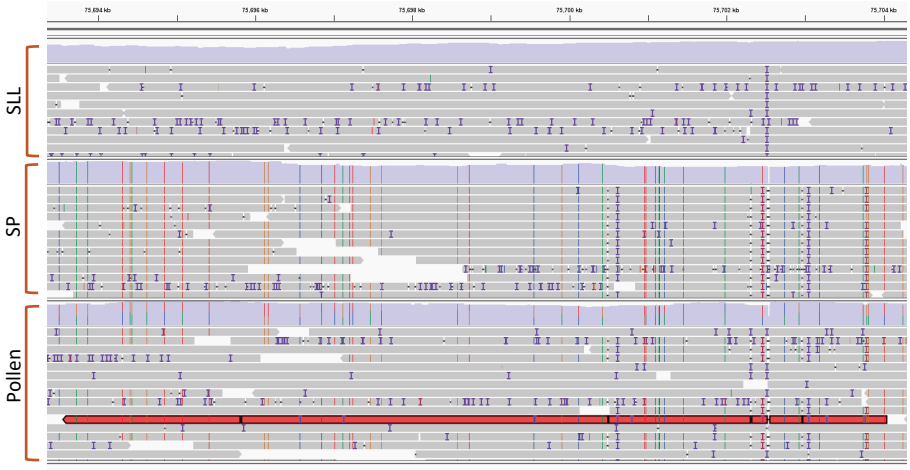


Figure 3.4. Example gene conversion in HiFi read. The top two track layers show the HiFi reads from the *S. lycopersicum* and *S. pimpinellifolium* parents, the third layer show reads from the pool of pollen from the an interspecific hybrid plant. The purple tracks show the mapping coverage and the colored markers indicate polymorphisms relative to the SL4.0 reference genome. The GC is detected in a 10kb read with orange color.

consecutive markers. In conclusion, given the relatively high sequencing error rate (compared to Illumina), HiFi reads are not well suited for GC detection.

3.3 Discussion

We demonstrated that we are able to detect COs with high resolution using HiFi reads derived from a pool of pollen gametes. The significant improvement of resolution shows that PacBio HiFi can serve as an alternative technology for gamete-based CO profiling and may even provide better data for analyses such as characterization of recombination regions. Both the sequencing protocol and the analysis algorithm can still be optimized further to increase sensitivity. First, we should avoid the amplification step that limits the size of DNA molecules to 10kb. Longer molecules potentially will increase the number of markers supporting the phase shifts, allowing detection of longer CO regions and increasing sensitivity. Second, optimizing the number of input genome copies or pollen gametes can also increase the coverage per genome, and thus the chance of sequencing the CO regions in each gamete. Furthermore, our current DNA isolation method requires the germination of pollen, which relies on the efficiency of germination and misses recombination events in non-germinating gametes, introducing a bias. Developing a protocol for DNA isolation that does not rely on germination may help us study

recombination events that reduce fertility or increase detrimental rearrangements and copy losses.

Another application that needs further research is the detection of gene conversions (GCs). Sequencing errors hamper accurate identification of GC signals in individual reads. Although we found a few GCs by manual inspection of results, we did not manage to accurately filter out false positives automatically without losing real GCs, i.e. introducing false negatives. If sequencing error rates do not decrease, resolving the errors in homopolymeric regions and microsatellites will require novel computational solutions that address the complications both in alignment and variant calling.

Despite these challenges, the high accuracy and improved resolution of COs detected using HiFi reads support it as an alternative sequencing platform to detect COs from pooled pollen gametes. This allows us to continue exploring meiotic recombination without the need to generate, screen and sequence an offspring population, saving time and cost.

3.4 Methods

3.4.1 Sequencing HiFi reads

Pacbio HiFi libraries were produced from ~5ng of gDNA isolated from germinated pollen as described in Fuentes *et al.* (2020). The pollen were collected from an F1 hybrid cross between *S. lycopersicum* and *S. pimpinellifolium*. We also isolated DNA from the inbreds of the parental genomes. Sample amplification of the target gDNA with a shear size of 10kb was carried out with the SMRTbell gDNA amplification kit followed by library construction using the SMRTbell Express TPK2.0 according to the manufacturers ultra-low DNA input protocol (Ultra-Low DNA Input Library Preparation Using SMRTbell Express Template Prep Kit 2.0, https://pacb.com/). HiFi reads of up to 10kb were generated by Circular Consensus Sequencing, using SMRT cells. Subsequent consensus calling was done using the pbccs v5.0.0 command line utility. HiFi reads were defined as CCS reads having at least 3 passes and a mean read quality score >20.

3.4.2 Crossover detection

We modified the method we previously developed to detect COs in 10X Genomics linked reads (Fuentes *et al.*, 2020) to enable the use of HiFi sequences as alternative input. The first step in the pipeline is to select genomic markers that can help distinguish the parental origin of reads, allowing the inference of a possible reciprocal recombination between homologous chromosomes. We identified such segregating markers between the parental genomes by aligning the *S.lycopersicum* (SLL) and *S.pimpinellifolium* (SP) HiFi reads against the SL4.0 (*S.lycopersicum*) reference genome (Hosmani *et al.*, 2019) using minimap2 (Li, 2018), followed by SNP calling using GATK HaplotypeCaller and hard-filtering (Poplin *et al.*, 2017). Only homozygous SNPs in SP that do not overlap with a SNP in SLL were selected, avoiding

inaccuracies in assigning parental origin when the SNP is heterozygous or when the SLL parent shows differences against the reference (which is of the same accession). In contrast to linked reads, which rely on Illumina sequencing, HiFi reads contain sequencing errors more frequently in homopolymers and microsatellites, causing false recombination signals due to incorrect alignment around some markers. Similarly, indels occurring in either of the parental genomes may lead to the detection of incorrect SNPs. To resolve these issues, we excluded SNPs in problematic regions, i.e. inside or within 15bp distance of homopolymers, microsatellites or indels.

Next, we used the selected segregating markers to detect changes in haplotype from one parent to another in a single HiFi read, potentially indicating a CO or GC event. For each marker, the identity of the corresponding base in the reads should match the expected alleles from either of the parents and their Phred quality score should be higher than 31, to avoid spurious haplotype shift detection caused by random mutations, sequencing errors and Mendelian errors. After assigning the parental origin of the markers in a read, we recovered the reads showing patterns of either a CO or GC. Compared to the previous version of the pipeline (Fuentes *et al.*, 2020), we deprecated the filtering of recombinant molecules overlapping transposable elements, to allow detection of COs enriched in specific TE superfamilies. We also computed for each putative recombinant molecule the CO resolution, which is the inverse of the distance between the SNPs immediately flanking a CO region. Only COs with a resolution above 2×10^{-4} , similar to the limit we previously used (Fuentes *et al.*, 2020), were reported.

3.4.3 Validation

Using *bedtools* (Quinlan and Hall, 2010), we compared the COs detected using HiFi reads against those detected from 10X linked reads (Fuentes *et al.*, 2020), offspring populations (Demirci *et al.*, 2017), and historical recombination (Chapter 4) and determined whether overlap is higher than expected by chance (Fisher's exact test). We used the same test for comparing CO and genic regions.

3.5 References

Anderson, L. K., Doyle, G. G., Brigham, B., Carter, J., Hooker, K. D., Lai, A., Rice, M. & Stack, S. M. 2003. High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics,* 165**,** 849-65.

Campoy, J. A., Sun, H., Goel, M., Jiao, W. B., Folz-Donahue, K., Wang, N., Rubio, M., Liu, C., Kukat, C., Ruiz, D.*, et al.* 2020. Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol,* 21**,** 306.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C.*, et al.* 2013. Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nat Genet,* 45**,** 1327-36.

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

Demirci, S., van Dijk, A. J., Sanchez Perez, G., Aflitos, S. A., de Ridder, D. & 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 Journal,* 89**,** 554-564.

Dreau, A., Venu, V., Avdievich, E., Gaspar, L. & Jones, F. C. 2019. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun,* 10**,** 4309.

Francis, K. E., Lam, S. Y., Harrison, B. D., Bey, A. L., Berchowitz, L. E. & Copenhaver, G. P. 2007. Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proc Natl Acad Sci U S A,* 104**,** 3913-8.

Fuentes, R., Hesselink, T., Nieuwenhuis, R., Bakker, L., Schijlen, E., van Dooijeweert, W., Diaz Trivino, S., de Haan, J. R., Sanchez Perez, G., Zhang, X.*, et al.* 2020. Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. *Plant J,* 102**,** 480-492.

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S.*, et al.* 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv*.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T.*, et al.* 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res,* 19**,** 1068-76.

Li, H. 2018. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics,* 34**,** 3094-3100.

Mercier, R., Mezard, C., Jenczewski, E., Macaisne, N. & Grelon, M. 2015. The molecular biology of meiosis in plants. *Annu Rev Plant Biol,* 66**,** 297-327.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D.*, et al.* 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

Qi, J., Wijeratne, A. J., Tomsho, L. P., Hu, Y., Schuster, S. C. & Ma, H. 2009. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC Genomics,* 10**,** 475.

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

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

Stack, S. M., Anderson, L. K. & Sherman, J. D. 1989. Chiasmata and recombination nodules in *Lilium longiflorum*. *Genome,* 32**,** 486-498.

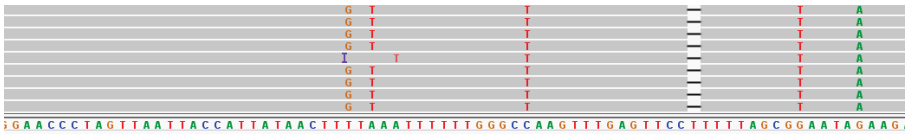
Sun, H., Rowan, B. A., Flood, P. J., Brandt, R., Fuss, J., Hancock, A. M., Michelmore, R. W., Huettel, B. & Schneeberger, K. 2019. Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nat Commun,* 10**,** 4310.

Wenger, A. M., Peluso, P., Rowell, W. J., Chang, P. C., Hall, R. J., Concepcion, G. T., Ebler, J., Fungtammasan, A., Kolesnikov, A., Olson, N. D.*, et al.* 2019. Accurate circular consensus long-

read sequencing improves variant detection and assembly of a human genome. *Nat Biotechnol,* 37**,** 1155-1162.

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L.*, et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife,* 2**,** e01426.

3.6 Supplementary figure



Supplementary Figure 3.1. An example of a false GC due to sequencing errors. The purple marker indicates the presence of a 1bp insertion in a homopolymeric region, causing a false GC signal. The extra 'T' insertion was aligned as a SNP while the expected marker 'G' was reported as an insertion.

Chapter 4

Domestication shapes recombination patterns in tomato

Published as Roven Rommel Fuentes, Dick de Ridder, Aalt D.J. van Dijk, Sander A. Peters. *Molecular Biology and Evolution* 39(1), 2022.

Abstract

Meiotic recombination is a biological process of key importance in breeding, to generate genetic diversity and develop novel or agronomically relevant haplotypes. In crop tomato, recombination is curtailed as manifested by linkage disequilibrium decay over a longer distance and reduced diversity compared to wild relatives. Here we compare domesticated and wild populations of tomato and find an overall conserved recombination landscape, with local changes in effective recombination rate in specific genomic regions. We also study the dynamics of recombination hotspots resulting from domestication and found that loss of such hotspots is associated with selective sweeps, most notably in the pericentromeric heterochromatin. We found footprints of genetic changes and structural variants, among them associated with transposable elements, linked with hotspot divergence during domestication, likely causing fine-scale alterations to recombination patterns and resulting in linkage drag.

4.1 Introduction

Generation of genetic diversity through meiotic recombination has been an integral subject of breeding and genome research following Mendel's work on patterns of inheritance and Morgan's theory of gene linkage and crossing-over (Hunter, 2015). Studies on meiotic recombination have indicated that one crossover (CO) per chromosome is obligatory and essential for proper chromosome segregation during prophase I (Jones and Franklin, 2006). COs are non-uniformly distributed across plant genomes, mostly located in distal chromosome regions and clustering in hotspots (Mercier *et al.*, 2015; Lambing *et al.*, 2017; Wang and Copenhaver, 2018). Characterizing the pattern of recombination has both practical and fundamental relevance, as it enables identification of informative markers, building of genetic maps and genome assemblies, reconstruction of evolutionary histories, association studies on important alleles and profiling linkage drags (Cardon and Abecasis, 2003; Wang *et al.*, 2010; Yang *et al.*, 2011; Fransz *et al.*, 2016; Dutta *et al.*, 2017; Marand *et al.*, 2019).

Different methods have been developed to detect recombination events between specific parents, such as chiasmata and recombination nodule counting (Stack *et al.*, 1989; Anderson *et al.*, 2003), pollen genotyping (Drouaud *et al.*, 2013), SNP-array-based profiling or sequencing of recombinant inbred line (RIL) populations (Huang *et al.*, 2009; Qi *et al.*, 2009; Wijnker *et al.*, 2013; Demirci *et al.*, 2017), and image analysis of meiotic tetrads (Francis *et al.*, 2007; Lim *et al.*, 2020). These approaches revealed useful information about the recombination landscape of gametes and offspring populations. Still, detection of recombination between all members of a population is impractical and often infeasible, especially for long-generation species. However, analysis of resequencing data of natural populations allows detection of historical recombination rates and associated genomic features, revealing their evolutionary histories. Here, historical recombination refers to the reciprocal exchange of chromosomal segments that has successively occurred between ancestral individuals over multiple generations, in various environments and under changing selective pressures and genetic backgrounds. Recombination rate landscapes and historical crossover hotspots in populations of different plants (Choi *et al.*, 2013; Dreissig *et al.*, 2019; Marand *et al.*, 2019; Schwarzkopf *et al.*, 2020), fungi (Stukenbrock and Dutheil, 2018), insects (Chan *et al.*, 2012) and mammals (Brunschwig *et al.*, 2012; Stevison *et al.*, 2016; Guo *et al.*, 2018) have previously been subjected to coalescent-based analysis of genetic variation.

Studying historical recombination can help to explain the changes that domestication enforced on recombination patterns. Domestication is a process of human-imposed evolution by selection of favorable phenotypes, resulting in genetic modification of wild progenitors to create new forms that meet human needs (Doebley *et al.*, 2006; Yang *et al.*, 2019). In this process, an initial stage of cultivating wild species with desirable traits is followed by a second stage of improvement, further targeting specific traits through selective breeding (Gross and Olsen, 2010; Meyer and Purugganan, 2013). Strong selection during domestication is

accompanied by increased recombination rate in many species (Ross-Ibarra, 2004; Moyers *et al.*, 2018). For example, the domesticated population of cacao was reported to have a higher recombination rate compared to wild populations (Schwarzkopf *et al.*, 2020). In addition, the recombination landscape in barley is highly conserved throughout domestication, with fine-scale changes in recombination rate that have been linked to different environmental conditions and with defense response genes (Dreissig *et al.*, 2019). However, currently, for most plants information on how domestication shaped the recombination landscape is still scarce.

Given that several genomic features like promoter regions, repeat motifs, nucleosome occupancy, chromatin accessibility, structural variations (SVs), transposable elements (TEs) and nucleotide diversity have been found linked to crossover incidence, changes to their patterns due to evolution of a species may have influenced the recombination profiles (Petes, 2001; Choi and Henderson, 2015; Termolino *et al.*, 2016; Lambing *et al.*, 2017; Dluzewska *et al.*, 2018). Population-level patterns of recombination in different species revealed substantial divergence of hotspots across populations of the same species. Distinct crossovers hotspots, associated with lineage-specific variation in SV and TE profiles, have been implicated as major drivers in population dynamics (Marand *et al.*, 2019). In rice, potato and *Arabidopsis thaliana,* specific superfamilies of DNA transposons are abundantly located in recombination-prone regions, which may be explained by nucleosome depletion in DNA transposons (Marand *et al.*, 2017; Choi *et al.*, 2018; Marand *et al.*, 2019).

Domesticated in South America, wild tomato species *S. pimpinellifolium* (SP) gave rise to the cherry tomato (*S. lycopersicum* var. *cerasiforme;* SLC), which was later improved into the big-fruited tomato (*S. lycopersicum* var. *lycopersicum;* SLL) in Mesoamerica (Blanca *et al.*, 2015; Razifard *et al.*, 2020; Lin *et al.*, 2014). Due to domestication and continued selection, current tomato cultivars have lost 95% of the genetic diversity of their wild relatives, pushing breeders to introgress alleles from compatible wild relatives underlying disease resistance, stress tolerance, adaptation to diverse environments, higher yield and fruit quality (Bai and Lindhout, 2007). However, there are some reproductive barriers such as SVs (Soyk *et al.*, 2019) that limit the applications of introgressive hybridization breeding. It was observed that linkage disequilibrium decays over a longer distance in domesticated tomato compared to its wild relatives, implying changes into the recombination patterns (Lin *et al.*, 2014; Zhu *et al.*, 2018). In this study, we address how domestication shaped the recombination patterns specifically for tomato and its related wild species. We investigate crossover profiles in tomato populations and its wild relative *S. pimpinellifolium,* using resequencing data assigned into three taxonomic groups: wild tomato (SP), early domesticated types (SLC), and vintage or heirloom cultivars (SLL). We generated recombination landscapes, identified recombination hotspots and analyzed genomic features that are associated with the differing recombination patterns between the populations. This provided insights into the factors that

contributed to or are associated with the changes in local recombination patterns during tomato domestication, revealing how domestication has severely constrained the ability of recombination to generate diversity, both for inbred and hybrid crosses. This new data may help the selection of targets for inducing meiotic recombination, cross checking hotspots in hybrids, identifying tightly-linked genes, and defining recombination barriers in hybridization.

4.2 Results

4.2.1 Conserved recombination landscape between wild and vintage tomato

After the domestication of tomato from its wild progenitor, its genetic diversity has dramatically reduced and linkage disequilibrium decay over a longer distance, indicating changes in recombination patterns (Aflitos *et al.*, 2014; Razifard *et al.*, 2020; Tieman *et al.*, 2017; Lin *et al.*, 2014). Furthermore, there are structural rearrangements between wild and domesticated plants that hamper recombination, manifested by the phenomenon of co-segregation of specific alleles linked to a desired trait (linkage drag). To get more insight into how domestication influences recombination patterns, we profiled the recombination landscape of tomato and wild relatives based on existing resequencing data of 75 accessions from each of the wild (SP), early-domesticated (SLC) and vintage (SLL) populations (Figure 4.1A; Supplementary Figure 4.1; Supplementary Table 1).

Consistent with available crossover data from recombinant-inbred lines (referred to from here on as COD1; Demirci *et al.*, 2017) and pollen gametes of an interspecific cross (COD2; Fuentes *et al.*, 2020) in tomato, and data from other species (Wijnker *et al.*, 2013; Kianian *et al.*, 2018), the majority of historical recombination in both wild and domesticated tomato occurred in the distal gene-rich euchromatic regions of the chromosomes. More in detail, the recombination landscape of each population correlates with both **COD1** (Spearman's rank correlation; euchromatin, $\rho = 0.32 - 0.44$; $P < 2.2 \times 10^{-11}$; heterochromatin, $\rho = 0.64 - 0.67$; $P < 2.2 \times 10^{-16}$) and **COD2** (Spearman's rank correlation; euchromatin, $\rho = 0.31 - 0.55$; $P < 2.2 \times 10^{-10}$; heterochromatin, $\rho = 0.38 - 0.51$; $P < 2.2 \times 10^{-16}$). To further verify the consistency of the recombination rates with available data, the population-scaled recombination rate computed using LDhat was converted from ρ/kb to cM/Mb, and compared against the EXPIM2012 genetic map generated from a cross between *S. lycopersicum* and *S. pimpinellifolium* (Sim *et al.*, 2012). The correlation between the genetic map and recombination rate estimates of each population is approximately 0.9 (Spearman's rank correlation; $P < 2.2 \times 10^{-16}$), supporting the concordance of population-scaled recombination rates with the genetic map.

To compare the recombination landscapes of the three populations in our study, we calculated multi-scale correlations, i.e. in varying window sizes (Supplementary Figure 4.2), and selected a 1-Mb window size. We found correlations in the range 0.6 - 0.7, indicating conservation of the genome-wide

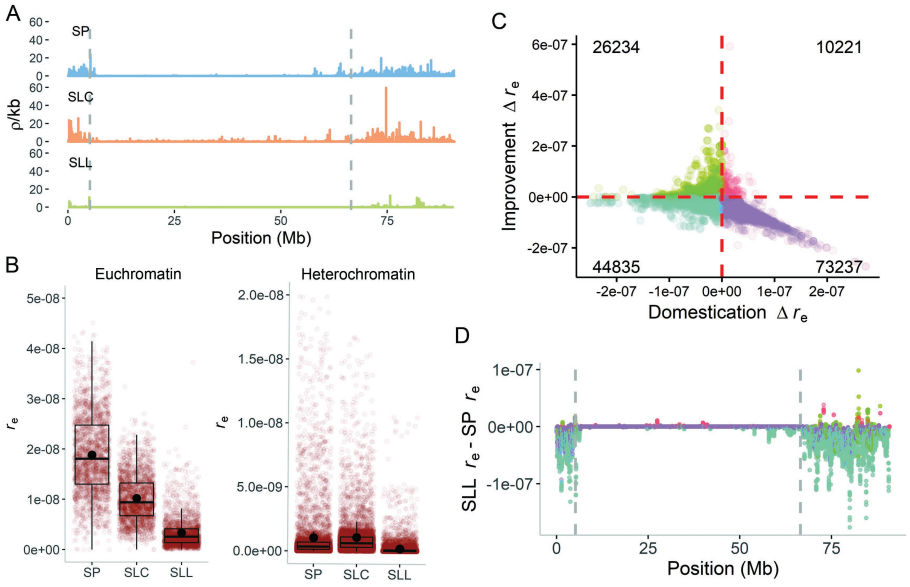


Figure 4.1. Recombination landscape and transformation from wild to domesticated tomato. (**A)** Recombination landscape in chromosome 1 of wild (SP), early-domesticated (SLC), and vintage (SLL) tomato. This *ρ*/kb landscape is intended to show overall landscape only; r_e is used to compare populations in other analyses. Gray vertical lines mark heterochromatin boundaries. (**B)** Effective recombination rate (r_e) in 1-Mb windows of both wild and domesticated tomato. (**C)** Change in effective recombination rate in 50-kb regions during domestication (SLC r_e – SP r_e) and improvement (SLL r_e – SLC r_e). (**D)** Resulting change in r_e for chromosome 1 after the domestication process or between the wild and vintage population. Gray vertical lines mark the heterochromatin boundaries and the colors correspond to the colors in **(C)**.

recombination landscape despite the differing selection and domestication processes (Spearman's rank correlation; $P < 2.2 \times 10^{-16}$, Supplementary Figure 4.3). The higher correlation coefficient between genetic distances and population-scaled recombination rate is influenced by the low marker density in the genetic map, limiting the comparison to the overall landscape. On the other hand, the use of shorter windows when comparing rates between populations or against COD1 and COD2 accounts for local changes in the recombination rates (Supplementary Figure 4.2).

4.2.2 Local increase of recombination rate in early-domesticated tomato

Although the general landscape of recombination is conserved, there are also clear local changes between populations (Supplementary Figure 4.1). Using the ρ value calculated by LDhat, we computed the effective recombination rate (r_e) to account for the difference in effective population size (N_e) of the three groups. We found that the median recombination rate in the SLC population ($r_e = 9.3 \times 10^{-10}$) is higher than the median in both SP ($r_e = 5.6 \times 10^{-10}$) and SLL ($r_e = 6.8 \times 10^{-11}$). Given that the heterochromatin regions have a low recombination rate, we separately analyzed r_e for euchromatic and heterochromatic regions. We determined the borders between euchromatin and heterochromatin by computing the euchromatin length (µm) from the average length of pachytene chromosome, multiplying it by the euchromatin DNA density (1.54 Mb/µm), and using the length of each euchromatin to identify the heterochromatin boundary. In euchromatin, SP has a significantly higher r_e of 1.8×10^{-8} compared to 9.4×10^{-9} for SLC and 2.6×10^{-9} for SLL. But in the heterochromatin, SLC has a higher median r_e of 6.3×10^{-10} compared to 3.5×10^{-10} and 4.8×10^{-11} of SP and SLL, respectively (Figure 4.1B; Supplementary Table 2).

We plotted the change in r_e during the domestication (from SP to SLC) and improvement (from SLC to SP) stages to detect localized changes of recombination rates (Figure 4.1C). In the majority of cases where there is an increase of r_e through domestication (SLC r_e – SP $r_e > 0$) across the whole genome, they are followed by a proportional decrease in r_e during improvement (SLL r_e – SLC $r_e < 0$). The reduction of r_e through domestication is confined mostly to the euchromatic region (Figure 4.1D; Supplementary Figure 4.4). Although 65% of euchromatin regions have reduced r_e during domestication, over the entire genome more regions (54%) have an increased rate. On the other hand, during the improvement stage the effective recombination rate in 76% of the genome was reduced, in both euchromatin and heterochromatin.

The local increase of r_e from SP to SLC is consistent with the fact that domestication increases the actual recombination rate in many species, as was demonstrated previously by counting chiasmata per bivalent (Ross-Ibarra, 2004; Moyers *et al.*, 2018). This increased recombination is favored during periods of rapid evolutionary change and specifically during domestication (Ross-Ibarra, 2004; Rees and Dale, 1974; Burt and Bell, 1987; Otto and Barton, 1997). On the other hand, the significant reduction of effective recombination rate during improvement may be explained by increased inbreeding and homozygosity in the vintage accessions (Moyers *et al.*, 2018); SP and SLC are known to have higher outcrossing rates than SLL (Rick *et al.*, 1978; Rick and Holle, 1990). As previously reported, inbreeding results in reduced heterozygosity and effective population size, and longer distance for linkage-disequilibrium decay (Allard, 1999; Morrell *et al.*, 2003; Kovach *et al.*, 2007). The estimated effective recombination rate may be reduced in inbred species if the homologous chromosomes are identical and no appreciable exchange of alleles is observed after recombination (Moyers *et al.*, 2018). The decreased

effective recombination rate has actually been observed in maize improved lines, with an estimated 82.3% reduction compared to the wild progenitors (Hufford *et al.*, 2012).

To check for the genetic diversity in each population, we first computed the number of single nucleotide polymorphisms (SNPs). The average SNP count per kilobase for SP, SLC and SLL accessions was 0.097, 0.075 and 0.021, respectively, which indicates a reduction of genetic diversity from wild to heirloom tomato (pairwise Wilcoxon rank sum test; $P < 8.5 \times 10^{-3}$; Supplementary Figure 4.5). At the individual sample level SP accessions have more SNPs than SLC accessions, but SLC has more unique SNP sites in the overall whole population. The mean nucleotide diversity measured for SLL (4.4×10^{-4}) and SLC (2.2×10^{-3}) populations is lower than for SP (3.8×10^{-3}), which is typically associated with domestication syndrome, the distinguishing characteristics between domesticated crop and wild ancestors (Sauvage *et al.*, 2017; Doebley *et al.*, 2006; Bai and Lindhout, 2007). Both increased homozygosity and inbreeding contributed to the reduced effectiveness of recombination in the vintage population.

4.2.3 Divergent hotspots between populations

Visual inspection of the landscape (Supplementary Figure 4.1, Figure 4.1A) reveals local peaks of recombination rates throughout the genome of each population. To further investigate this, we detected historical recombination hotspots for each population using sequenceLDhot, reporting 1,784, 2,899 and 667 hotspots for SP, SLC and SLL, respectively (Figure 4.2A; Supplementary Table 2 & 3) and a total of 5,082 unique hotspots with a median size of 2 kb. Pairs of the three populations have 4-10% of hotspots in common, significantly higher than expected by chance (pairwise Fisher's exact test; $P < 1.9 \times 10^{-5}$; Figure 4.2B). However, of 181 hotspots shared between SP and SLC, only 4 are retained in SLL. This low overlap in hotspots between populations of the same or closely related species was also reported in rice and cocoa (Marand *et al.*, 2019; Schwarzkopf *et al.*, 2020). SLC has 62.5% more unique hotspots than SP, in concordance with its increased recombination rate. Out of all the identified hotspots, 84 (1.6%) and 96 (1.8%) overlapped with empirical crossovers in **COD1** (Fisher's exact test; $P = 6.3 \times 10^{-24}$) and **COD2** (Fisher's exact test; $P = 1.2 \times 10^{-26}$;), respectively. Furthermore, 3 of the 23 reported hotspots in Fuentes *et al.* (2020) match hotspots in SP and SLC. The limited overlap between historical hotspots and crossovers found in these previous studies is in line with the relatively modest correlation between the population-scaled recombination rate and crossovers from **COD1** and **COD2** reported above. These observations may be explained by the inclusion of different genotypes in the different dataset, by the possibly differing crossover patterns between the RIL/pollen populations and the natural tomato population, or by the inability of experimental studies to exhaustively sample possible recombination sites. The hotspots mentioned in the succeeding sections refers to recombination hotspots, except for those that explicitly refer to hotspots from previous studies.

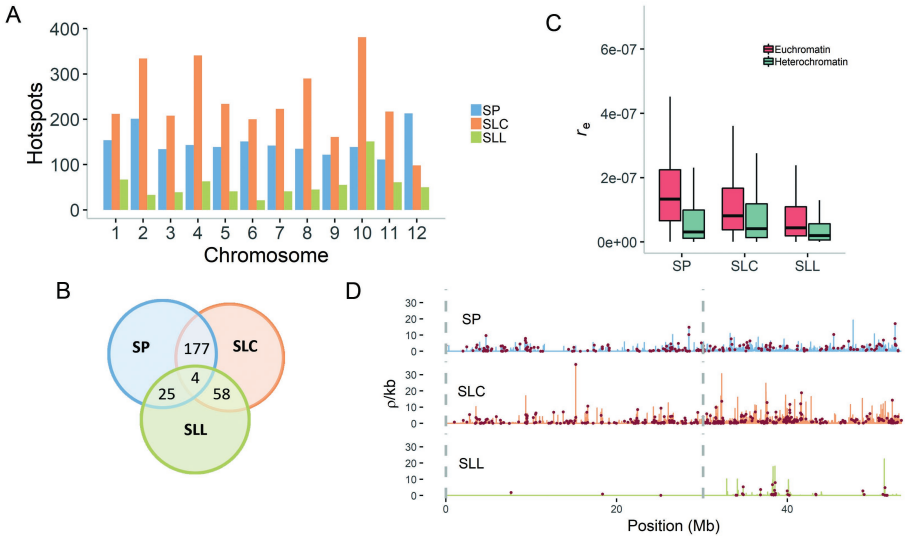


Figure 4.2. Historical recombination hotspots. (A) Number of hotspots in each chromosome of the wild and domesticated populations. (B) Small but significant numbers of hotspots are shared between populations. (C) Effective recombination rates of hotspots in euchromatic and heterochromatic regions. (D) Recombination hotspots in the upper heterochromatic arm of chromosome 2. Gray lines mark the heterochromatin boundaries.

4.2.4 Recombination in the pericentromeric heterochromatin

Aside from divergent hotspots, another distinct difference between the recombination landscapes of wild and domesticated tomato is the presence of hotspots in the pericentromeric regions. Previous studies in plants mostly report suppression of recombination in pericentromeric regions, which is largely heterochromatic, and high recombination rates in the distal chromosome regions. However, we observed that hotspots are not confined to the terminal chromosome ends, but also are scattered over the pericentrome. Compared with both SP (51.8%) and SLC (61.5%), only 32.7% of the genome-wide hotspots in vintage tomato (SLL) are located in the pericentromeric regions, which covers 75% of the genome. This rate is comparable to the 39.1% of hotspots distributed in the pericentromere of maize (Pan *et al.*, 2017). Sherman and Stack (1995) actually reported cases of recombination nodules in the pericentromeric heterochromatin of tomato, but these are 20-50x less frequent per unit length of the synaptonemal complex than in euchromatin. This frequency, though, may be different for historical hotspots. Additionally, hotspots in the heterochromatin have generally lower recombination rates than those located in the euchromatin (Wilcoxon rank sum test; SP, $P < 2 \times 10^{-}$

^{16}; SLC, $P < 2 \times 10^{-16}$; SLL, $P = 6.2 \times 10^{-12}$; Figure 4.2C). To determine whether the observed recombination rates in the heterochromatin are due to the use of a domesticated tomato as reference genome for wild accessions which contains genomic rearrangements, we recomputed recombination rates for the wild population using the *S. pimpinellifolium* genome assembly (Wang *et al.*, 2020) as reference. As shown in Supplementary Figure 4.6, the pericentromeres still exhibits presence of historical crossovers. It is also important to emphasize that SNPs located in the repeat or TE regions were excluded in the estimation of recombination rates to avoid issues due to misaligned reads or false positive SNPs.

To give an example of recombination occurring in the pericentromeric heterochromatin, we show the landscape for chromosome 2 in Figure 4.2D, where both SP and SLC show presence of historical recombination in the heterochromatin. However, the same region is almost devoid of recombination in SLL; hotspots are clustering mostly at the ends of the chromosome, consistent with population data, the genetic map (Sim *et al.*, 2012), RILs (Demirci *et al.*, 2017) and pollen gametes from interspecific crosses (Fuentes *et al.*, 2020). Around 55% and 42% of SP and SLC hotspots in chromosome 2, respectively, are located in the pericentromeric regions, in contrast to 9% of SLL, but the number of hotspots per megabase of euchromatin is still higher than in heterochromatin for both SP and SLC. We further examined these heterochromatic hotspots and found that they are close to or within genes (Fisher's exact test; $P = 4.9 \times 10^{-12}$), which suggests that these hotspots may be located in euchromatin islands or accessible regions in the heterochromatin.

4.2.5 Crossover hotspots in selective sweep genes

The results above confirm that historical crossovers are non-uniformly distributed over the genome and occur mostly in the distal part of the chromosome. This raises the question how the changing patterns of recombination hotspots may be linked to specific genomic features that evolve during domestication. To test the association of recombination hotspots with gene features, for each population, a permutation test of crossover hotspots with sizes below 5 kb was applied. This shows a significant enrichment in promoter regions, defined as 1-kb regions upstream of the transcriptional start sites (TSS), and in gene bodies. Moreover, hotspots are depleted in intergenic regions, further supporting previous reports of recombination mostly occurring near genes. To account for the significant difference of crossover distribution in the euchromatin and heterochromatin, we performed enrichment analysis for euchromatin regions only and still found an excess in promoters and gene bodies (Figure 4.3A). The promoter regions exhibit a 3 to 13-fold increase in crossovers over the background, which was computed based on a set of 10,000 permutations. Despite the reduction of hotspots in vintage tomato (Supplementary Table 2), the enrichment in both gene bodies and promoters persists. An excess of crossovers in promoter and UTR regions of tomato was previously reported in RILs and pollen data (de Haas *et al.*, 2017; Demirci *et al.*, 2017; Fuentes *et al.*, 2020) and is consistent with observations in other species

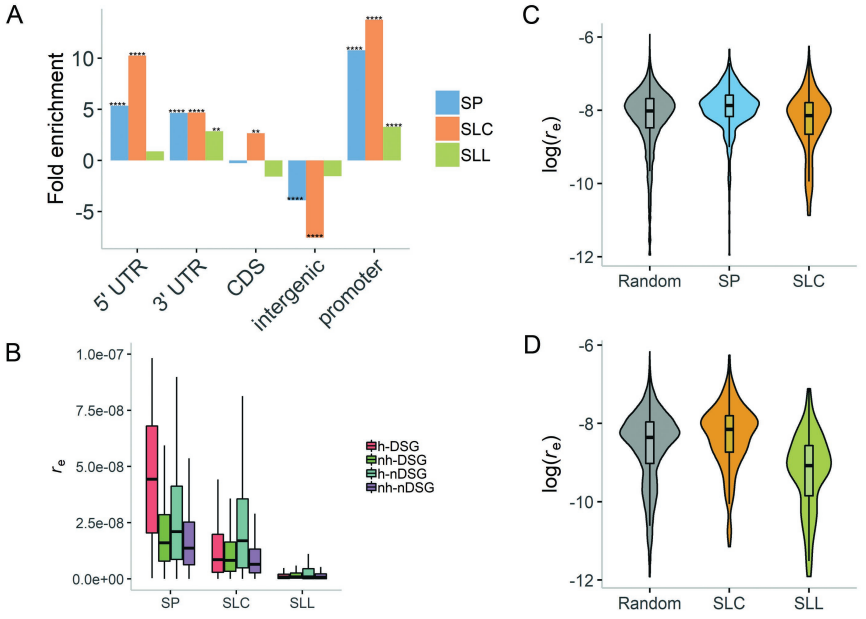


Figure 4.3. Recombination hotspots in genes. (A) Enrichment of euchromatin hotspots in UTRs and promoter regions (1-kb upstream of genes) in all three populations. (B) Recombination rates of domestication (DSG) and non-domestication (nDSG) sweep genes overlapping and not overlapping S. pimpinellifolium hotspots. **h** and **nh** mean hotspots and non-hotspots, respectively. (C-D) Recombination rate upstream (<1kb) of genes with excised promoters due to (C) domestication and (D) improvement.

(Wijnker *et al.*, 2013; Marand *et al.*, 2017; Choi *et al.*, 2018; Demirci *et al.*, 2018; Kianian *et al.*, 2018; Pan *et al.*, 2017). The overrepresentation in the 5' and 3' UTRs may be related to epigenetic modifications such as DNA and histone methylation, contrasting open chromatin that may be accessible for the recombination machinery (Pan *et al.*, 2017; Eichten *et al.*, 2011). With the preference of crossovers to occur near genes, the significant reduction of genes in domesticated tomatoes compared to wild relatives consequently limits the possible sites for recombination (Gao *et al.*, 2019).

Knowing that crossovers preferentially occur near genes and that specific genes are affected by domestication, we examined the relation between recombination hotspots and genes. In detecting association, we included the 2-kb regions flanking both sides of the genes. First, we compared hotspots against R genes, which are known to reside in recombination hotspots (Nieri *et al.*, 2017; Andolfo *et al.*, 2021), and observed a significant overlap (Fisher's exact test, $P = 3.38 \times 10^{-17}$). Afterwards,

we computed the overlap between hotspots and the selective sweeps or regions with reduced nucleotide diversity previously reported by Lin *et al.* (2014). We refer to genes in these selective sweeps as domestication (DSG) and improvement (ISG) sweep genes. For heterochromatic DSGs/ISGs, a significant enrichment in both SP and SLC hotspots is observed, while there is no association with SLL hotspots (**Supplementary Table 4**). Conversely, in the euchromatin, there is a difference between DSGs and ISGs. Euchromatic SP hotspots significantly overlap with both DSGs and ISGs, while euchromatic SLC hotspots only show enrichment in ISGs, which implies that many DSGs have lost hotspots after domestication. Lastly, SLL hotspots in the euchromatin overlapped DSGs and ISGs significantly less than expected by chance, indicating that most SLL hotspots are outside selective sweeps.

Many of the hotspots in SP and SLC sweep genes are lost in the SLL population. Nevertheless, it might be that sweep genes still undergo recombination even after hotspots are lost during domestication. To investigate this, we examined the changes in recombination rates of both sweep and non-sweep genes across these three populations. We computed the effective recombination rates in DSG and ISG genes across the different populations and compared them against the remaining genes (Figure 4.3B; Supplementary Figure 4.7). Genes with hotspots clearly exhibited an elevated effective recombination rate, with even higher rates for sweep genes (DSGs/ISGs) than non-sweep genes. However, the loss of these hotspots during the domestication or improvement process resulted in r_e being reduced to almost the same level as the non-hotspot genes. Interestingly, sweep genes show more reduction in r_e than non-sweep genes, which may reflect how sweep genes were directly affected by tomato domestication. Altogether, Supplementary Figure 4.7 underlines the severity of the decrease in r_e between wild and domesticated tomato genes, which resulted in reduced genetic diversity and forces breeders to introgress alleles from the wild relatives to recover desired traits.

Aside from increasing the frequency of alleles in specific genes, domestication of tomato also resulted in lost or negatively-selected promoters during both domestication and improvement stages (Gao *et al.*, 2019). These promoters are considered unfavorable because of their significantly lower frequency in SLC than SP or in SLL than SLC. Given that above it was demonstrated that recombination is associated with promoters, the question is how the loss of these promoters influence recombination. We found that the upstream region (<1kb) of the genes with promoters under selection during domestication have reduced effective recombination rate in SLC compared to SP (Wilcoxon rank sum test, $P < 2 \times 10^{-16}$) (Figure 4.3C). Similarly, the upstream region of the genes with promoters under selection during improvement have reduced r_e in SLL compared to SLC (Wilcoxon rank sum test, $P < 2 \times 10^{-16}$) (Figure 4.3D). This analysis reveals that loss of promoters due to domestication affected the recombination rate in the genomic regions from which these promoters were lost, highlighting a specific way in which domestication reduces recombination.

4.2.6 TE and SV-associated hotspots

Domestication and improvement not only influence gene content but have a broader effect on genomic variants (Gao *et al.*, 2019; Alonge *et al.*, 2020). Hence, we finally analyzed the association between hotspots and transposable elements (TEs) as well as structural variants (SVs). Certain transposable element families show strong association with hotspots (Figure 4.4A). Both *hAT-Tip100* and *Stowaway* show enrichment while *Tag1*, *L1*, *Copia* and *Gypsy* show strong depletion, based on a permutation test. Most class I TE or retrotransposons are under-represented in hotspots, except for *ERV1* in the SLL population and SINE in all three populations. Moreover, regions with low complexity and simple repeats have an excess of recombination hotspots. Similar to potato, rice and maize, *Stowaway* and SINE are significantly overrepresented, whereas *Gypsy* and *Copia* are depleted in hotspots (Marand *et al.*, 2017; Marand *et al.*, 2019; Pan *et al.*, 2017). In Arabidopsis, DSBs overlapped *Gypsy* and *Copia* elements significantly less than expected by chance (Choi *et al.*, 2018); however, in maize, most DSBs are formed in repetitive regions, predominantly *Gypsy* retrotransposons, but only genic DSBs contribute to crossover formation (He *et al.*, 2017). Schwarzkopf *et al.* (2020) reported that hotspots shared by both domesticated and wild cocoa populations appear to be associated with DNA transposons. Furthermore, Marand *et al.* (2019) found that the presence of *Stowaway* and *Harbinger* elements in crossover hotspots is associated with increased recombination rates, augmented chromatin accessibility and reduced DNA methylation. Our result is also consistent with the findings on differentially accessible chromatin regions between meiotic cells and somatic cells of tomato (Chouaref J, Tark-Dame M, Koes R, Fransz P, Stam M, unpublished data), corroborating that TE families with an excess of hotspots are accessible in meiotic cells while those with depletion of hotspots are inaccessible. TE families enriched with hotspots are known to be preferentially located in genic regions, but the retrotransposon LTR/ERV1 is particularly interesting because it was also reported to be one of the most transcriptionally active TE families due to its abundance in exonic regions (Mehra *et al.*, 2015). The insertion of retrotransposons in a promoter region or UTR can both regulate gene expression (Dominguez *et al.*, 2020; Alonge *et al.*, 2020) and negatively affect the chance of crossover incidence. Similarly, the accumulation of certain DNA transposons that are known to be accessible in meiotic cells may provide new sites for DSBs that can resolve to crossovers. Our results suggest that TE activities during domestication also influenced the landscape of meiotic recombination (He *et al.*, 2017; Choi *et al.*, 2018; Underwood and Choi, 2019; Kent *et al.*, 2017).

Alonge *et al.* (2020) reported that, compared to SLL and SLC, SP has significantly more SVs relative to the Heinz 1706 reference genome and that the majority of insertions and deletions are associated with *Gypsy* and *Copia* elements. To determine if SVs influence meiotic recombination, we compared the SVs between

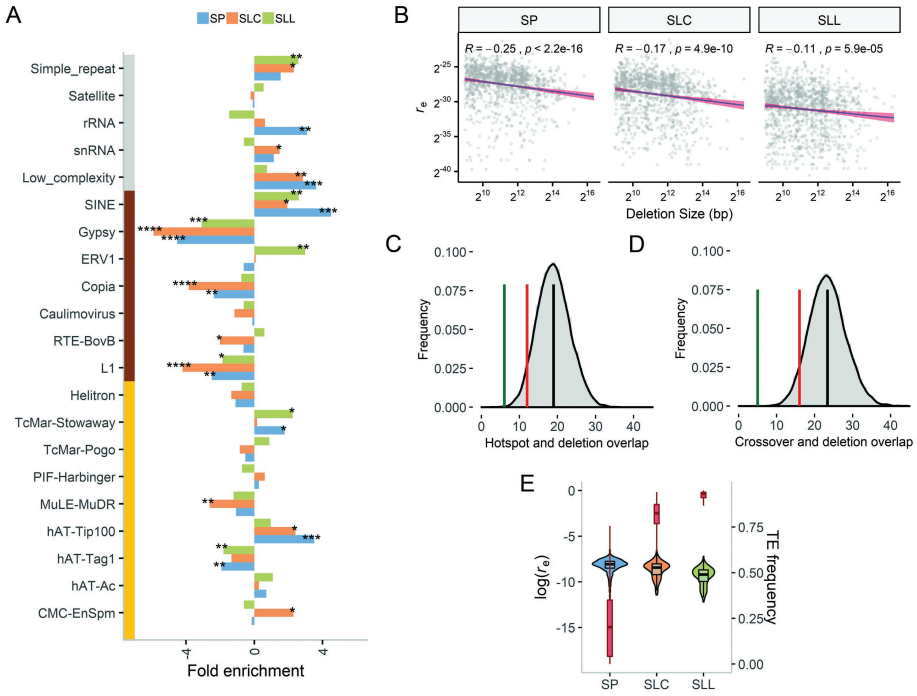


Figure 4.4. Recombination and genomic variants. Using permutation tests, we identified (A) specific TE families with an excess or depletion of recombination hotspots. TE families are grouped into repeat elements (gray), retrotransposons (brown), and DNA transposons (yellow). (B) Scatter plots of effective recombination rate and deletion size (n=1255) per population. (C-D) Significance of overlap between (C) hotspots and deletions in SP and (D) empirical crossovers and deletions segregating between SP and SLL. The black and red vertical lines indicate the average number of overlaps found in 10,000 permutation sets and the number of overlaps at P = 0.05, respectively. The green vertical line indicates the observed number of overlaps. (E) Recombination rates (violin) and allele frequencies (red boxplot) of *Gypsy*, *Copia* and *L1* elements.

S. pimpinellifolium and *S. lycopersicum* reported in Wang *et al.* (2020) against the effective recombination rate and hotspots. We found that the length of deletions correlates with recombination rate, specifically deletions longer than 500 bp (Figure 4.4B), and that recombination hotspots are suppressed in long deletions in the wild population (Fisher's exact test, $P = 2.1 \times 10^{-6}$). Hotspot suppression is less prominent in deletions with lower allele frequency (Supplementary Figure 4.8). As a specific example, we found a 4-fold reduction of SP hotspots in large deletions (> 500 bp) with allele frequency above 0.5 (Figure 4.4C). Interestingly, the same set of large deletions have low allele frequency and lack hotspot suppression in both SLC and

SLL populations, which may be explained by the increased inbreeding and lower heterozygosity of SV regions in domesticated varieties compared to the wild relatives. Knowing the wild ancestral alleles, we can instead look at these regions of low frequency deletions (relative to Heinz reference) as insertion sites for alleles that fixated in the vintage accessions. Lye and Purugganan (2019) reported that SVs associated with domestication traits have an increased allele frequency in the population due to selection or have become fixated in the population. This suggests that in natural populations, suppression of recombination hotspots in SVs occurs more in outcrossing populations with a certain level of heterozygosity of SVs between compatible individuals (Wang and Copenhaver, 2018; Dluzewska *et al.*, 2018). To put this result in a broader perspective, we analyzed the crossovers detected from the interspecific hybrid between *S.pimpinellifolium* and *S.lycopersium* (**COD2**) and found crossover suppression in the large deletions segregating between SP and SLL populations (Fisher's exact test, $P = 1.4 \times 10^{-5}$; Figure 4.4D). This suppression of crossovers in deletions segregating between the parents of hybrid crosses is consistent with results obtained from the natural populations with high levels of outcrossing.

Interestingly, the shorter deletions (< 500 bp) overlap recombination hotspots in each population significantly more than expected by chance (Fisher's exact test, SP $P = 1.1 \times 10^{-6}$; SLC $P = 4.9 \times 10^{-12}$; SLL $P = 6.4 \times 10^{-13}$; Supplementary Figure 4.9). These regions with short deletions and hotspots have high homology and are enriched with simple repeats (z-score > 11.1) and hAT-Tip100 (z-score > 3.1) elements. These small deletions have significant allele frequency differences between the wild and domesticated populations based on Wang *et al.* (2020), mostly have low frequency in the vintage population. Moreover, we found that the fixated SLL alleles in these deletion sites are causal mutations in domestication syndrome genes, like OVATE (Solyc02g085500), Lin5 (Solyc09g010080) and YABBY (Solyc11g071810) and other genes controlling horticulture traits in tomato (Wang *et al.*, 2020).

As we reported above, we identified some TE families that associate with hotspots and we speculated that TE activities during domestication affects recombination patterns. Comparing TEs to deletions, we selected TE elements that show a significant difference in frequency between the wild and domesticated populations, mostly *Gypsy*, *Copia* and *L1* elements. As shown in Figure 4.4E, the frequencies of these TE elements increased during domestication (fewer deletion alleles), while recombination rates in these elements significantly declined. This is consistent with the suppression of hotspots in these specific TE families that became fixed in the vintage population (Figure 4.4A). The result indicates that despite the lower deletion frequency or lower heterozygosity in a region, genomic content such as presence of specific TEs can influence whether recombination is suppressed or not.

4.3 Discussion

Throughout domestication, the general recombination landscape of both wild and domesticated tomato remains conserved, consistent with the known phenomenon in tomato and other plant species that recombination rates are significantly higher in distal parts of chromosomes (Mercier *et al.*, 2015; Lambing *et al.*, 2017; Wang and Copenhaver, 2018). Effective recombination rates are in agreement with empirical crossover data derived from RILs and pollen gametes, and with genetic distances from the EXPIM2012 linkage map. Despite the conservation of the recombination landscape, we observed local increases in recombination rates across the chromosomes of early-domesticated tomato (SLC), which reflects the expected increase of recombination due to domestication (Ross-Ibarra, 2004; Moyers *et al.*, 2018). Concomitantly, some regions of the gene-rich euchromatin showed a reduction of recombination. After the domestication stage, further reduction of effective recombination rates is observed in vintage (SLL) tomato. This might either be due to low diversity in highly homozygous regions limiting the detection of actual recombination, or because the recombination has actually decreased during the improvement stage (Otto and Barton, 1997; Moyers *et al.*, 2018). Nevertheless, the effective recombination rates reported here, allowed us to identify genomic regions that still recombine and generate diversity in vintage tomato.

We also identified recombination hotspots and found a small but significant overlap between the three populations, representing conserved recombination sites apparently not affected by domestication. The number of hotspots in vintage tomato is four times less than in its progenitor, revealing genomic sites in domesticated tomato preferentially maintained with lower if not completely suppressed recombination. Previous studies have reported that the pericentromeric heterochromatin of tomato displays extremely low recombination compared to the distal chromosome regions (Demirci *et al.*, 2017; Fuentes *et al.*, 2020), but we identified historical recombination hotspots near or within genes located in the heterochromatin, hinting on their accessibility as possible euchromatin islands within heterochromatin. These heterochromatin hotspots have a lower recombination rate than those located in euchromatin, which could explain why it can be hard to observe them in previous studies. Furthermore, compared to SP and SLC, the majority of SLL hotspots are located in the euchromatin. Our results on the divergence of hotspots between tomato populations suggests highly dynamic recombination patterns, likely as a result of the intense selection in domestication (Otto and Barton, 1997; Schwarzkopf *et al.*, 2020). Despite the use of interspecific tomato crosses in several studies, genetic shuffling through meiosis still remains strongly confined to the ends of the chromosomes (Demirci *et al.*, 2017; Fuentes *et al.*, 2020). Interestingly, Dreissig *et al.* (2019) reported an increase in pericentromeric crossovers, although they have been observed for crosses of domesticated and wild barley accessions. The low recombination rate in 75% of the tomato genome limits possibilities of generating diversity or breaking fixed haplotype blocks in either or both wild and domesticated tomato. However, despite

the very limited sites for recombination in the domesticated tomato, detecting recombination hotspots in the wild genomes may help reveal factors that led to their divergence and ways to restore hotspots and genetic diversity in tomato.

The recombination hotspots found in wild and domesticated tomatoes are enriched in gene bodies and promoter regions and depleted in intergenic regions, which agrees with previous reports on tomato crossovers (Demirci *et al.*, 2017; Fuentes *et al.*, 2020). This implies that the reduction in the number of genes and promoters from wild to vintage tomato limits possible hotspot locations. Gao *et al.* (2019) reported hundreds of genes and promoters excised as a result of domestication and improvement, reducing recombination rate in specific genomic regions of vintage tomato and increasing the spans of haplotype blocks. The profile of recombination hotspots in the wild population reveals genomic regions that provide candidate targets for inducing meiotic recombination or loss-of-function mutations to domesticate wild plants or re-introduce desirable traits, or serve as guide for cross checking hotspots from experimental populations (Zsogon *et al.*, 2018; Li *et al.*, 2018).

Comparing with the known selective sweep genes, we discovered that they overlap with recombination hotspots in both the wild and early-domesticated tomato significantly more than expected by chance. The hotspot-associated selective sweep genes in the SLC population indicated selection for increased recombination rate during domestication, which was then followed by a loss of hotspots after subsequent improvement. Otto and Barton (1997) proposed that higher recombination rates during domestication were favored to elevate the fixation rate of adaptive or beneficial alleles. These manifested as reduced nucleotide diversity in sweep regions of the SLL population that historically comprised hotspots. The increased recombination rate and number of hotspots in SLC may also relate to adaptive evolution in response to changing environments during domestication. Similar observations have been made for rice genes that showed higher recombination in response to stresses (Si *et al.*, 2015). In addition, Razifard *et al.* (2020) identified an overrepresentation of defense-response genes in the selective sweeps of an SLC population, which agrees with our observation of excess recombination hotspots in R genes. Unlike SP and SLC hotspots which significantly overlap selective sweeps, SLL hotspots are mostly located outside the sweeps, showing the divergence of SLL hotspots from regions under selection during tomato domestication and improvement. The reduced genetic diversity in sweep regions of the highly inbred SLL population could have resulted from the loss of hotspots or reduced recombination incidence. Alternatively, the low diversity, possibly due to a selection bottleneck, may have led to the low effectivity of recombination in these sweep regions. Our results support an evolution of effective recombination related to the fixation of alleles in selective sweeps genes.

In all three populations, recombination hotspots were found to be positively (e.g. *Stowaway*, SINE) and negatively (e.g. *Gypsy*, *Copia*) associated with specific families of transposable elements. TE families with excess recombination hotspots

were preferentially located in genic regions (Mehra *et al.*, 2015). During meiosis, these specific TEs probably are accessible to recombination complexes, whereas most retrotransposons have closed chromatin status (Chouaref J, Tark-Dame M, Koes R, Fransz P, Stam M, unpublished data). Aside from TEs, we also identified association of hotspots with structural variations. We showed that large deletions suppress recombination in SP populations, but this depends on the frequency of the deletion variant in the population. Since most large deletions have lower frequency in both SLC and SLL populations, we do not observe hotspot suppression by deletions in these populations. Reduced recombination in SVs combined with geographic isolation can lead to the development of alleles that are incompatible with distantly-related haplotypes (Jiao and Schneeberger, 2020; Bomblies and Weigel, 2007). As an example, crossover suppression is observed in a hybrid cross between SP and SLL parental accessions, implying that SVs can cause linkage drag that constrains introgression of specific alleles from the wild relatives while excluding unwanted alleles (Taagen *et al.*, 2020). Rowan *et al.* (2019) provided several hypotheses as to why crossovers are suppressed in SV regions, such as absence of repair template, COs in SVs creating inviable gametes, DSB in SVs preferentially resolving to NCOs, constrained interaction with the central element and homologous chromosomes and lastly the DNA methylation in SVs.

On the other hand, we found enrichment of hotspots in short deletions across all populations, which we hypothesize is not due to mutations inducing hotspots but more likely due to mutations formed by non-allelic homologous recombination in the hotspot regions with tracts of homology (Balachandran and Beck, 2020; Escaramís *et al.*, 2015). These small causal mutations, that are located in many horticulture trait genes, may have originated from the wild progenitor and may confer adaptive domestication traits that promote their fixation in the vintage population (Lye and Purugganan, 2019). Previous studies already reported recurrent SVs in regions with high recombination rate (McVean, 2010; Liu *et al.*, 2012; Rowan *et al.*, 2019; Badet *et al.*, 2021), although there are other mechanisms that generate more SVs (Escaramís *et al.*, 2015). SVs are reported to have key roles in the evolution of domesticated species and are associated with post-domestication traits (Lye and Purugganan, 2019). Thus, the divergence of TEs and SVs, both associated with recombination rates and hotspots, also contributes to the differing patterns of recombination during domestication. Identifying these divergent sequences between populations can allow us to determine genomic regions that are unlikely to recombine, or alleles that remain tightly linked in the hybrid crosses. Our results further confirm that domestication influenced local genomic contents, which in turn affected the recombination patterns in tomato.

4.4 Methods

4.4.1 Computing recombination rates and detecting hotspots

We collected resequencing data from previous studies under accession numbers PRJEB5235 (Aflitos *et al.*, 2014), PRJNA454805 (Razifard *et al.*, 2020), PRJNA353161 (Tieman *et al.*, 2017) and PRJNA259308 (Lin *et al.*, 2014) and based on previous population analyses (Tieman *et al.*, 2017; Zhu *et al.*, 2018; Gao *et al.*, 2019; Alonge *et al.*, 2020; Razifard *et al.*, 2020) we selected accessions that were unanimously assigned to the *S. pimpinellifolium* (SP), *S. lycopersicum var. cerasiforme* (SLC) or *S. lycopersicum var. lycopersicum* (SLL) taxonomic groups. Most accessions were sequenced using Illumina HiSeq2000 except those from Razifard which utilized Illumina NextSeq. Modern processing cultivars and hybrids were excluded to avoid obscuring the recombination patterns in each of these three major groups. From the combined list of accessions, we selected 81 SPs, 140 SLCs, and 136 SLLs.

Reads were trimmed using Trimmomatic (Bolger *et al.*, 2014) and aligned against the tomato Heinz 1706 (SL4.0) reference genome (Hosmani *et al.*, 2019) using bwa mem (Li, 2013). Accessions with read and physical coverage below 5x and 70%, respectively, were discarded. Filtering left 75 SP, 122 SLC and 117 SLL samples. SNPs were identified per accession using GATK HaplotypeCaller (Poplin *et al.*, 2017) with default parameters and then we randomly selected 75 accessions per population for GATK joint-genotyping and hard-filtering. Using bedtools (Quinlan and Hall, 2010) and bcftools (Danecek *et al.*, 2011), we selected bi-allelic SNPs located outside repeat or transposable element regions, with a minimum allele frequency of 0.05, less than 10% missing data, and consisting of at most one heterozygous genotype. Heterozygous genotypes were then converted to missing data. All missing calls were imputed and phased using Beagle v. 5.1 (window = 5, overlap = 2, iterations = 30, err = 0.001, burnin = 10) (Browning *et al.*, 2018).

A widely used method to infer historical recombination rate is LDhat (Auton and McVean, 2007), which implements coalescent resampling with a Bayesian reversible-jump Markov Chain Monte Carlo (rjMCMC) algorithm to estimate population-scaled recombination rates ($\rho = 4N_e r_e$) from pairs of SNPs. We estimated recombination rate using the LDhat v2.2 *interval* program per window of 5000 SNPs, overlapping by 500 SNPs. LDhat was run with 20 million iterations, sampling every 2000 iterations, using a mutation rate of 0.001 and the first 2000 samples for burn-in. We subsequently detected recombination hotspots using sequenceLDhot (Fearnhead, 2006) with non-overlapping 1-kb windows, a 500-bp step size, and background recombination rate set as the median rate in a 50-kb window centered at each 1-kb window. Hotspots with rates greater than 10 times and less than 200 times the background and with likelihood ratio (LR) above the 95th percentile are reported. We merged hotspots within 500 bp of each other and used the highest LR and recombination rate for merged intervals. Only hotspots with at least a 10 times increase in recombination rate, based on the ratio of LDhat-computed ρ and the background ρ, were reported as the final set of crossover hotspots.

To compare recombination rates between populations and account for potential differences in effective population sizes (N_e), we first computed the nucleotide diversity (θ_w, Watterson's theta), using LDhat *convert* and used the neutral mutation rate (μ) of 1 x 10^{-8} (Baer *et al.*, 2007; Lin *et al.*, 2014; Moyers *et al.*, 2018) to estimate the $4N_e$ ($4N_e = \theta_w/\mu$). The estimated $4N_e$ was then used to calculate the effective recombination rate per generation ($r_e = \rho/4N_e$) for each population.

4.4.2 Comparison with genetic maps

For validating the recombination map, we compared median recombination rates against the crossover frequencies from tomato RILs (Demirci *et al.*, 2017) and pollen gametes (Fuentes *et al.*, 2020) using a 50-kb sliding window and performed Spearman's rank correlation test. Further correlation testing was done by comparing against EXPIM2012 genetic map (Sim *et al.*, 2012), using the method from Choi *et al.* (2013). Between every pair of adjacent SNPs in the genetic map, the population-scaled recombination rate (ρ/kb) was converted to cM/Mb. We also compared the recombination landscapes between populations using multi-scale correlations by calculating the average recombination rates in varying window sizes, by randomly sampling these windows 10,000 times, and subsequently calculating Spearman's correlation coefficients.

4.4.3 Compute euchromatin regions

The border between euchromatin and heterochromatin in each chromosome was computed according to Stack *et al.* (2009) and Demirci *et al.* (2017). Using Table 1 of Sherman and Stack (1992), we calculated the average length and heterochromatin length (µm) of each pachytene chromosome. Euchromatin length was calculated by subtracting heterochromatin length from each arm length and multiplying by the euchromatin DNA density (1.54 Mb/µm). Then, we determined the euchromatin-heterochromatin boundary based on the length of each euchromatic region. We also computed the heterochromatic region boundaries per chromosome relative to the Heinz 1706 (SL4.0) reference genome.

4.5 Supplementary material

Supplemental data and figures are available at *Molecular Biology and Evolution* online (https://doi.org/10.1093/molbev/msab287). Scripts are available at https://github.com/rrfuentes/histo-recom.git.

4.6 Acknowledgments

This work was supported by the MEICOM Marie Sklodowska-Curie Innovative Training Network (ITN), H2020-MSCA- ITN-2017 Horizon 2020 Grant agreement number 765212. We would like to acknowledge Paul Fransz, Jihed Chouaref and Maike Stam for the helpful discussions on accessible regions in tomato meiotic cells and Alexandre Marand for his insights on the analysis of historical recombination.

4.7 References

Aflitos, S., Schijlen, E., de Jong, H., de Ridder, D., Smit, S., Finkers, R., Wang, J., Zhang, G., Li, N., Mao, L.*, et al.* 2014. Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing. *Plant J,* 80**,** 136-48.

Allard, R. W. 1999. History of plant population genetics. *Annu Rev Genet,* 33**,** 1-27.

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D.*, 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., Doyle, G. G., Brigham, B., Carter, J., Hooker, K. D., Lai, A., Rice, M. & Stack, S. M. 2003. High-resolution crossover maps for each bivalent of *Zea mays* using recombination nodules. *Genetics,* 165**,** 849-65.

Andolfo, G., D'Agostino, N., Frusciante, L. & Ercolano, M. R. 2021. The Tomato Interspecific NB-LRR Gene Arsenal and Its Impact on Breeding Strategies. *Genes (Basel),* 12.

Auton, A. & McVean, G. 2007. Recombination rate estimation in the presence of hotspots. *Genome Res,* 17**,** 1219-27.

Badet, T., Fouche, S., Hartmann, F. E., Zala, M. & Croll, D. 2021. Machine-learning predicts genomic determinants of meiosis-driven structural variation in a eukaryotic pathogen. *Nat Commun,* 12**,** 3551.

Baer, C. F., Miyamoto, M. M. & Denver, D. R. 2007. Mutation rate variation in multicellular eukaryotes: causes and consequences. *Nat Rev Genet,* 8**,** 619-31.

Bai, Y. & Lindhout, P. 2007. Domestication and breeding of tomatoes: what have we gained and what can we gain in the future? *Ann Bot,* 100**,** 1085-94.

Balachandran, P. & Beck, C. R. 2020. Structural variant identification and characterization. *Chromosome Res,* 28**,** 31-47.

Blanca, J., Montero-Pau, J., Sauvage, C., Bauchet, G., Illa, E., Diez, M. J., Francis, D., Causse, M., van der Knaap, E. & Canizares, J. 2015. Genomic variation in tomato, from wild ancestors to contemporary breeding accessions. *BMC Genomics,* 16**,** 257.

Bolger, A. M., Lohse, M. & Usadel, B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

Bomblies, K. & Weigel, D. 2007. Hybrid necrosis: autoimmunity as a potential gene-flow barrier in plant species. *Nat Rev Genet,* 8**,** 382-93.

Browning, B. L., Zhou, Y. & Browning, S. R. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet,* 103**,** 338-348.

Brunschwig, H., Levi, L., Ben-David, E., Williams, R. W., Yakir, B. & Shifman, S. 2012. Fine-scale maps of recombination rates and hotspots in the mouse genome. *Genetics,* 191**,** 757-64.

Burt, A. & Bell, G. 1987. Mammalian chiasma frequencies as a test of two theories of recombination. *Nature,* 326**,** 803-5.

Cardon, L. R. & Abecasis, G. R. 2003. Using haplotype blocks to map human complex trait loci. *Trends in Genetics,* 19**,** 135-140.

Chan, A. H., Jenkins, P. A. & Song, Y. S. 2012. Genome-wide fine-scale recombination rate variation in *Drosophila melanogaster*. *PLoS Genet,* 8**,** e1003090.

Choi, K. & Henderson, I. R. 2015. Meiotic recombination hotspots - a comparative view. *Plant J,* 83**,** 52-61.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C.*, 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., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J.*, et al.* 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in *Arabidopsis thaliana* transposons and gene regulatory regions. *Genome Research,* 28**,** 532-546.

Danecek, P., Auton, A., Abecasis, G., Albers, C. a., Banks, E., DePristo, M. a., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T.*, et al.* 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England),* 27**,** 2156-8.

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

Demirci, S., Peters, S. A., de Ridder, D. & van Dijk, A. D. J. 2018. DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J*.

Demirci, S., van Dijk, A. J., Sanchez Perez, G., Aflitos, S. A., de Ridder, D. & 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 Journal,* 89**,** 554-564.

Dluzewska, J., Szymanska, M. & Ziolkowski, P. A. 2018. Where to Cross Over? Defining Crossover Sites in Plants. *Front Genet,* 9**,** 609.

Doebley, J. F., Gaut, B. S. & Smith, B. D. 2006. The molecular genetics of crop domestication. *Cell,* 127**,** 1309-21.

Dominguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jimenez-Gomez, J. M., Colot, V. & Quadrana, L. 2020. The impact of transposable elements on tomato diversity. *Nat Commun,* 11**,** 4058.

Dreissig, S., Mascher, M. & Heckmann, S. 2019. Variation in Recombination Rate Is Shaped by Domestication and Environmental Conditions in Barley. *Mol Biol Evol,* 36**,** 2029-2039.

Drouaud, J., Khademian, H., Giraut, L., Zanni, V., Bellalou, S., Henderson, I. R., Falque, M. & Mezard, C. 2013. Contrasted patterns of crossover and non-crossover at *Arabidopsis thaliana* meiotic recombination hotspots. *PLoS Genet,* 9**,** e1003922.

Dutta, R., Mainsah, J., Yatskiv, Y., Chakrabortty, S., Brennan, P., Khuder, B., Qiu, S., Fedorova, L. & Fedorov, A. 2017. Intricacies in arrangement of SNP haplotypes suggest "Great Admixture" that created modern humans. *BMC Genomics,* 18**,** 433.

Eichten, S. R., Swanson-Wagner, R. A., Schnable, J. C., Waters, A. J., Hermanson, P. J., Liu, S., Yeh, C. T., Jia, Y., Gendler, K., Freeling, M.*, et al.* 2011. Heritable epigenetic variation among maize inbreds. *PLoS Genet,* 7**,** e1002372.

Escaramís, G., Docampo, E. & Rabionet, R. 2015. A decade of structural variants: Description, history and methods to detect structural variation. *Briefings in Functional Genomics,* 14**,** 305-314.

Fearnhead, P. 2006. SequenceLDhot: detecting recombination hotspots. *Bioinformatics,* 22**,** 3061-6.

Francis, K. E., Lam, S. Y., Harrison, B. D., Bey, A. L., Berchowitz, L. E. & Copenhaver, G. P. 2007. Pollen tetrad-based visual assay for meiotic recombination in Arabidopsis. *Proc Natl Acad Sci U S A,* 104**,** 3913-8.

Fransz, P., Linc, G., Lee, C. R., Aflitos, S. A., Lasky, J. R., Toomajian, C., Ali, H., Peters, J., van Dam, P., Ji, X.*, et al.* 2016. Molecular, genetic and evolutionary analysis of a paracentric inversion in *Arabidopsis thaliana*. *Plant J,* 88**,** 159-178.

Fuentes, R., Hesselink, T., Nieuwenhuis, R., Bakker, L., Schijlen, E., van Dooijeweert, W., Diaz Trivino, S., de Haan, J. R., Sanchez Perez, G., Zhang, X.*, et al.* 2020. Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. *Plant J,* 102**,** 480-492.

Gao, L., Gonda, I., Sun, H., Ma, Q., Bao, K., Tieman, D. M., Burzynski-Chang, E. A., Fish, T. L., Stromberg, K. A., Sacks, G. L.*, et al.* 2019. The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. *Nat Genet,* 51**,** 1044-1051.

Gross, B. L. & Olsen, K. M. 2010. Genetic perspectives on crop domestication. *Trends Plant Sci,* 15**,** 529-37.

Guo, J., Chen, H., Yang, P., Lee, Y. T., Wu, M., Przytycka, T. M., Kwoh, C. K. & Zheng, J. 2018. LDSplitDB: a database for studies of meiotic recombination hotspots in MHC using human genomic data. *BMC Med Genomics,* 11**,** 27.

He, Y., Wang, M., Dukowic-Schulze, S., Zhou, A., Tiang, C. L., Shilo, S., Sidhu, G. K., Eichten, S., Bradbury, P., Springer, N. M.*, et al.* 2017. Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proc Natl Acad Sci U S A,* 114**,** 12231-12236.

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S.*, et al.* 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv*.

Huang, X., Feng, Q., Qian, Q., Zhao, Q., Wang, L., Wang, A., Guan, J., Fan, D., Weng, Q., Huang, T.*, et al.* 2009. High-throughput genotyping by whole-genome resequencing. *Genome Res,* 19**,** 1068-76.

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhajarvi, T., Chia, J. M., Cartwright, R. A., Elshire, R. J., Glaubitz, J. C., Guill, K. E., Kaeppler, S. M.*, et al.* 2012. Comparative population genomics of maize domestication and improvement. *Nat Genet,* 44**,** 808-11.

Hunter, N. 2015. Meiotic Recombination: The Essence of Heredity. *Cold Spring Harb Perspect Biol,* 7.

Jiao, W. B. & Schneeberger, K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun,* 11**,** 989.

Jones, G. H. & Franklin, F. C. 2006. Meiotic crossing-over: obligation and interference. *Cell,* 126**,** 246-8.

Kent, T. V., Uzunovic, J. & Wright, S. I. 2017. Coevolution between transposable elements and recombination. *Philos Trans R Soc Lond B Biol Sci,* 372.

Kianian, P. M. A., Wang, M., Simons, K., Ghavami, F., He, Y., Dukowic-Schulze, S., Sundararajan, A., Sun, Q., Pillardy, J., Mudge, J.*, et al.* 2018. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat Commun,* 9**,** 2370.

Kovach, M. J., Sweeney, M. T. & McCouch, S. R. 2007. New insights into the history of rice domestication. *Trends Genet,* 23**,** 578-87.

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

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 00**,** 3-3.

Li, T., Yang, X., Yu, Y., Si, X., Zhai, X., Zhang, H., Dong, W., Gao, C. & Xu, C. 2018. Domestication of wild tomato is accelerated by genome editing. *Nat Biotechnol*.

Lim, E. C., Kim, J., Park, J., Kim, E. J., Kim, J., Park, Y. M., Cho, H. S., Byun, D., Henderson, I. R., Copenhaver, G. P.*, et al.* 2020. DeepTetrad: high-throughput image analysis of meiotic tetrads by deep learning in *Arabidopsis thaliana*. *Plant J,* 101**,** 473-483.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X.*, et al.* 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet,* 46**,** 1220-6.

Liu, P., Carvalho, C. M., Hastings, P. J. & Lupski, J. R. 2012. Mechanisms for recurrent and complex human genomic rearrangements. *Curr Opin Genet Dev,* 22**,** 211-20.

Lye, Z. N. & Purugganan, M. D. 2019. Copy Number Variation in Domestication. *Trends Plant Sci,* 24**,** 352-365.

Marand, A. P., Jansky, S. H., Zhao, H., Leisner, C. P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A. J., Veilleux, R. E.*, 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. & Jiang, J. 2019. Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *Plant Cell,* 31**,** 645-662.

McVean, G. 2010. What drives recombination hotspots to repeat DNA in humans? *Philos Trans R Soc Lond B Biol Sci,* 365**,** 1213-8.

Mehra, M., Gangwar, I. & Shankar, R. 2015. A Deluge of Complex Repeats: The Solanum Genome. *PLoS One,* 10**,** e0133962.

Mercier, R., Mezard, C., Jenczewski, E., Macaisne, N. & Grelon, M. 2015. The molecular biology of meiosis in plants. *Annu Rev Plant Biol,* 66**,** 297-327.

Meyer, R. S. & Purugganan, M. D. 2013. Evolution of crop species: genetics of domestication and diversification. *Nat Rev Genet,* 14**,** 840-52.

Morrell, P. L., Lundy, K. E. & Clegg, M. T. 2003. Distinct geographic patterns of genetic diversity are maintained in wild barley (*Hordeum vulgare ssp. spontaneum*) despite migration. *Proc Natl Acad Sci U S A,* 100**,** 10812-7.

Moyers, B. T., Morrell, P. L. & McKay, J. K. 2018. Genetic Costs of Domestication and Improvement. *J Hered,* 109**,** 103-116.

Nieri, D., Di Donato, A. & Ercolano, M. R. 2017. Analysis of tomato meiotic recombination profile reveals preferential chromosome positions for NB-LRR genes. *Euphytica,* 213.

Otto, S. P. & Barton, N. H. 1997. The evolution of recombination: removing the limits to natural selection. *Genetics,* 147**,** 879-906.

Pan, Q., Deng, M., Yan, J. & Li, L. 2017. Complexity of genetic mechanisms conferring nonuniformity of recombination in maize. *Sci Rep,* 7**,** 1205.

Petes, T. D. 2001. Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics,* 2**,** 360-369.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D.*, et al.* 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

Qi, J., Wijeratne, A. J., Tomsho, L. P., Hu, Y., Schuster, S. C. & Ma, H. 2009. Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC Genomics,* 10**,** 475.

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

Razifard, H., Ramos, A., Della Valle, A. L., Bodary, C., Goetz, E., Manser, E. J., Li, X., Zhang, L., Visa, S., Tieman, D.*, et al.* 2020. Genomic Evidence for Complex Domestication History of the Cultivated Tomato in Latin America. *Mol Biol Evol,* 37**,** 1118-1132.

Rees, H. & Dale, P. J. 1974. Chiasmata and variability in Lolium and Festuca populations. *Chromosoma,* 47**,** 335-351.

Rick, C. M. & Holle, M. 1990. Andean lycopersicon esculentum var. cerasiforme: genetic variation and its evolutionary significance. *Economic Botany,* 44**,** 69-78.

Rick, C. M., Holle, M. & Thorp, R. W. 1978. Rates of cross-pollination inLycopersicon pimpinellifolium: Impact of genetic variation in floral characters. *Plant Systematics and Evolution,* 129**,** 31-44.

Ross-Ibarra, J. 2004. The evolution of recombination under domestication: a test of two hypotheses. *Am Nat,* 163**,** 105-12.

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R. & Weigel, D. 2019. An Ultra High-Density *Arabidopsis thaliana* Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics,* 213**,** 771-787.

Sauvage, C., Rau, A., Aichholz, C., Chadoeuf, J., Sarah, G., Ruiz, M., Santoni, S., Causse, M., David, J. & Glemin, S. 2017. Domestication rewired gene expression and nucleotide diversity patterns in tomato. *Plant J,* 91**,** 631-645.

Schwarzkopf, E. J., Motamayor, J. C. & Cornejo, O. E. 2020. Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. *BMC Genomics,* 21**,** 332.

Sherman, J. D. & Stack, S. M. 1992. Two-dimensional spreads of synaptonemal complexes from solanaceous plants. V. Tomato (Lycopersicon esculentum) karyotype and idiogram. *Genome,* 35**,** 354-359.

Sherman, J. D. & 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.

Si, W., Yuan, Y., Huang, J., Zhang, X., Zhang, Y., Zhang, Y., Tian, D., Wang, C., Yang, Y. & Yang, S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytol,* 206**,** 1491-502.

Sim, S. C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., Van Deynze, A., Hamilton, J. P., Buell, C. R., Causse, M., Wijeratne, S.*, et al.* 2012. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One,* 7**,** e40563.

Soyk, S., Lemmon, Z. H., Sedlazeck, F. J., Jimenez-Gomez, J. M., Alonge, M., Hutton, S. F., Van Eck, J., Schatz, M. C. & Lippman, Z. B. 2019. Duplication of a domestication locus neutralized a cryptic variant that caused a breeding barrier in tomato. *Nat Plants,* 5**,** 471-479.

Stack, S. M., Anderson, L. K. & Sherman, J. D. 1989. Chiasmata and recombination nodules in *Lilium longiflorum*. *Genome,* 32**,** 486-498.

Stack, S. M., Royer, S. M., Shearer, L. A., Chang, S. B., Giovannoni, J. J., Westfall, D. H., White, R. A. & Anderson, L. K. 2009. Role of fluorescence in situ hybridization in sequencing the tomato genome. *Cytogenet Genome Res,* 124**,** 339-50.

Stevison, L. S., Woerner, A. E., Kidd, J. M., Kelley, J. L., Veeramah, K. R., McManus, K. F., Great Ape Genome, P., Bustamante, C. D., Hammer, M. F. & Wall, J. D. 2016. The Time Scale of Recombination Rate Evolution in Great Apes. *Mol Biol Evol,* 33**,** 928-45.

Stukenbrock, E. H. & Dutheil, J. Y. 2018. Fine-Scale Recombination Maps of Fungal Plant Pathogens Reveal Dynamic Recombination Landscapes and Intragenic Hotspots. *Genetics,* 208**,** 1209-1229.

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

Termolino, P., Cremona, G., Consiglio, M. F. & Conicella, C. 2016. Insights into epigenetic landscape of recombination-free regions. *Chromosoma,* 125**,** 301-8.

Tieman, D., Zhu, G., Resende, M. F., Jr., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S., Taylor, M., Zhang, B.*, et al.* 2017. A chemical genetic roadmap to improved tomato flavor. *Science,* 355**,** 391-394.

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

Wang, J., de Villena, F. P., Moore, K. J., Wang, W., Zhang, Q. & McMillan, L. 2010. Genome-wide compatible SNP intervals and their properties. *ACM Int Conf Bioinform Comput Biol (2010),* 2010**,** 43-52.

Wang, X., Gao, L., Jiao, C., Stravoravdis, S., Hosmani, P. S., Saha, S., Zhang, J., Mainiero, S., Strickler, S. R., Catala, C.*, et al.* 2020. Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat Commun,* 11**,** 5817.

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

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L.*, et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife,* 2**,** e01426.

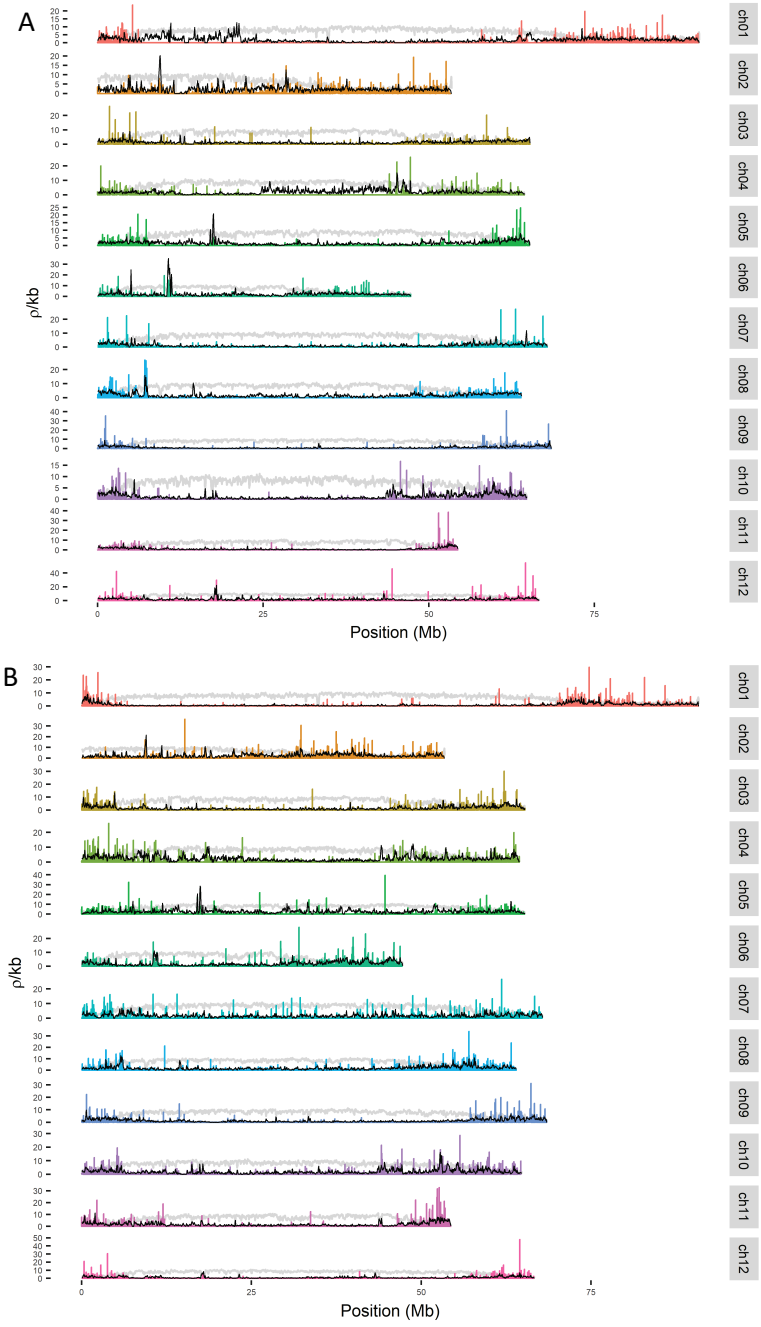
Yang, H., Wang, J. R., Didion, J. P., Buus, R. J., Bell, T. A., Welsh, C. E., Bonhomme, F., Yu, A. H., Nachman, M. W., Pialek, J.*, et al.* 2011. Subspecific origin and haplotype diversity in the laboratory mouse. *Nat Genet,* 43**,** 648-55.

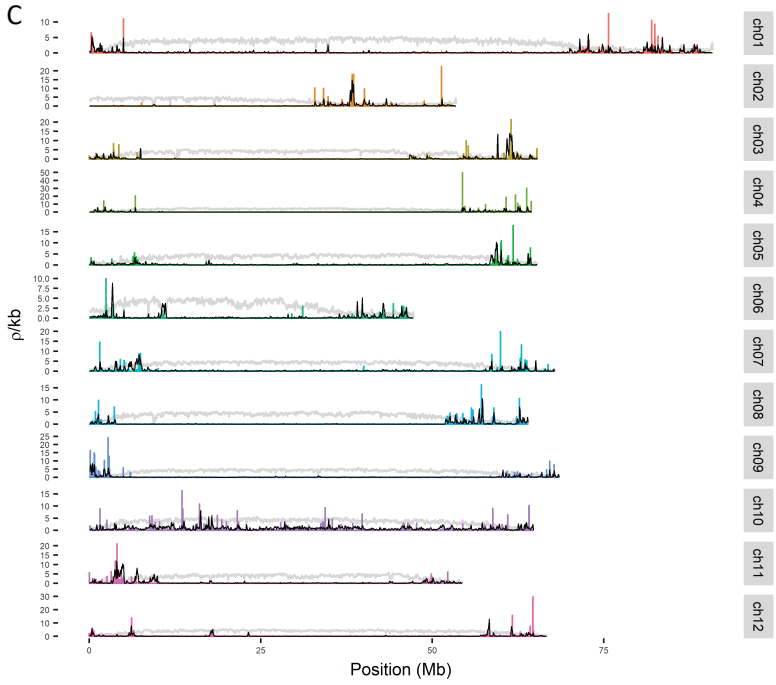
Yang, Z., Li, G., Tieman, D. & Zhu, G. 2019. Genomics Approaches to Domestication Studies of Horticultural Crops. *Horticultural Plant Journal,* 5**,** 240-246.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C.*, et al.* 2018. Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell,* 172**,** 249-261 e12.

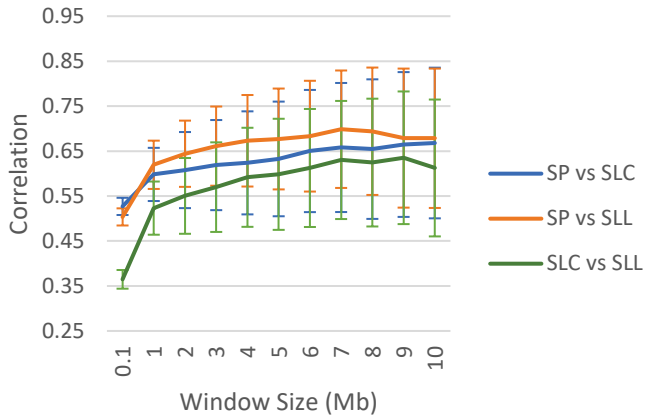
Zsogon, A., Cermak, T., Naves, E. R., Notini, M. M., Edel, K. H., Weinl, S., Freschi, L., Voytas, D. F., Kudla, J. & Peres, L. E. P. 2018. De novo domestication of wild tomato using genome editing. *Nat Biotechnol*.

4.8 Supplementary figures

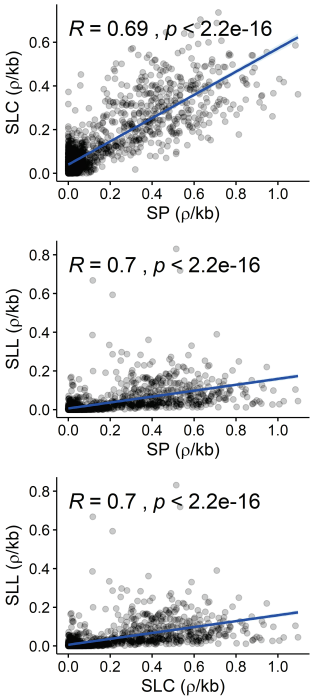




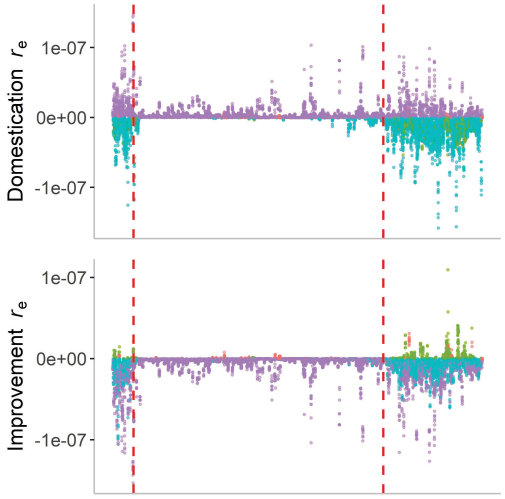
Supplementary Figure 4.1. Recombination landscape of different populations. (**A**) wild tomato (SP); (**B**) early-domesticated tomato (SLC); (**C**) vintage tomato (SLL). The colored, black and gray lines represent the recombination rate, SNP distribution and repeat content, respectively.



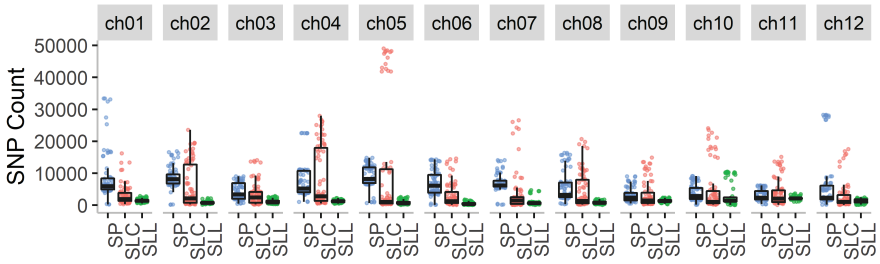
Supplementary Figure 4.2. Multi-scale correlation of recombination rates between different populations. Spearman's rank correlation of recombination rates between SP and SLC (blue), SP and SLL (orange) and, SLC and SLL (green) with varying window sizes. For larger windows, the mean correlation increases but the standard deviation among the 10,000 sets of randomly sampled windows increases as well.



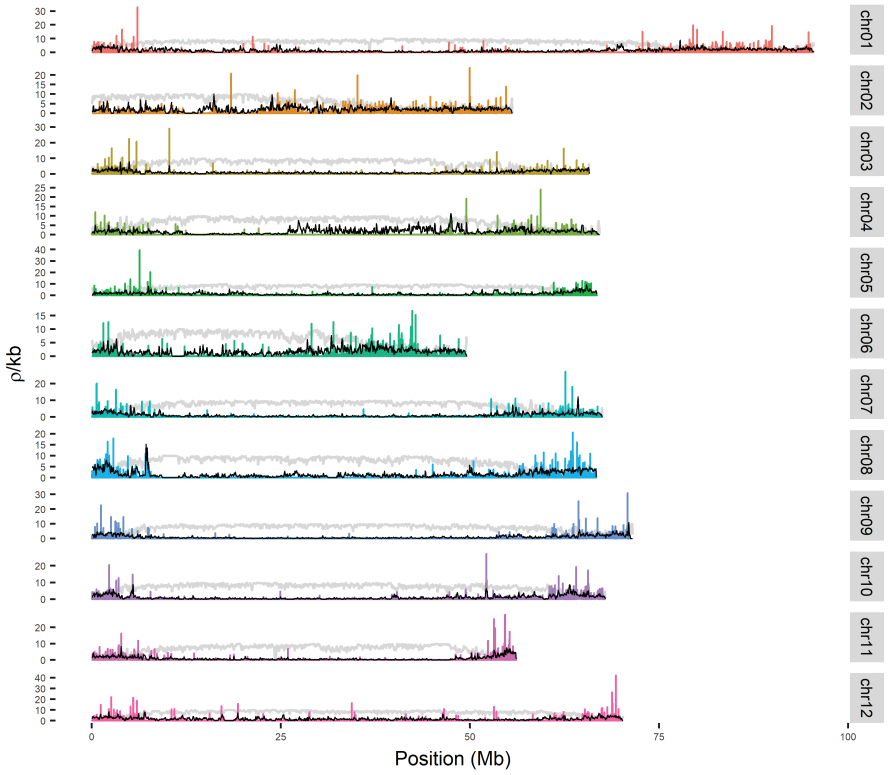
Supplementary Figure 4.3. Correlations of population-scaled recombination rates. The pairwise comparison of the rates in 1-Mb sliding window shows significant correlations between populations.



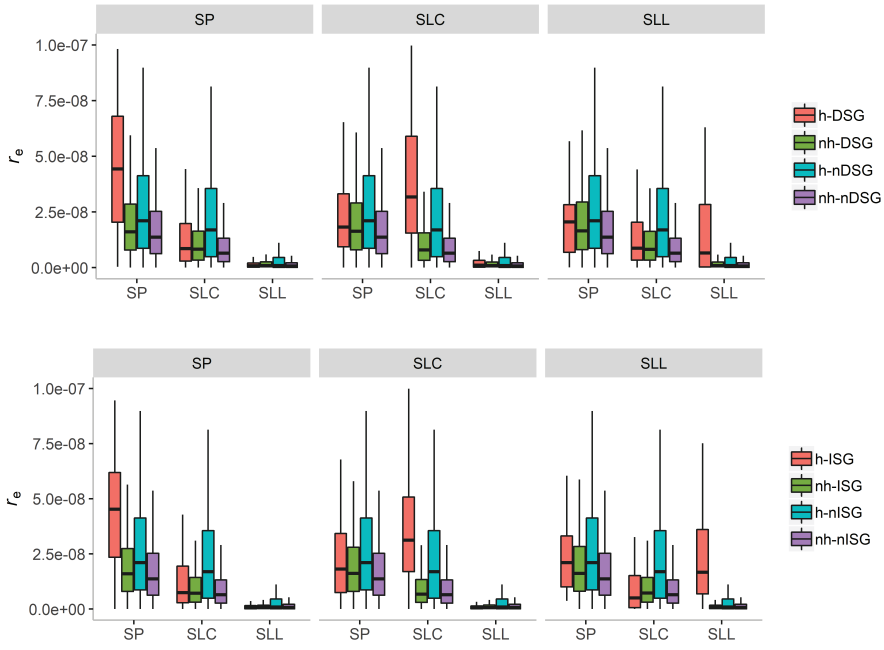
Supplementary Figure 4.4. Localized transformation of recombination rate in chromosome 1. Change in recombination rate in 50-kb regions during two-stage domestication of tomato where Domestication r_e = SLC r_e - SP r_e and Improvement r_e = SLL r_e - SLC r_e, respectively. The colors match the quadrants in Figure 1c. The blue dots represent mostly euchromatic regions with continuous r_e reduction in both stages while the violet dots are regions with increased r_e during domestication, followed by proportional reduction during improvement.



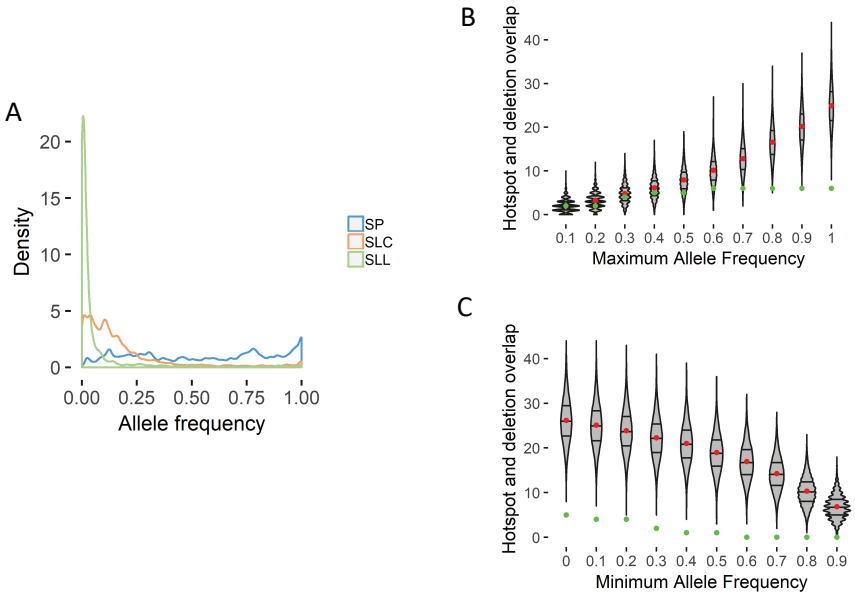
Supplementary Figure 4.5. SNPs in wild and domesticated tomato. Across all chromosomes, SP accessions generally have more SNPs than SLC and SLL accessions. In total, there are 454,326, 488,393, and 74,134 SNP sites for SP, SLC and SLL populations, respectively.



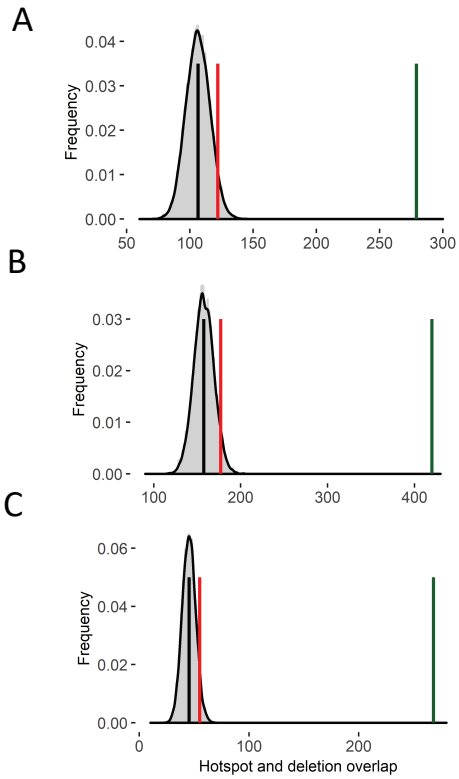
Supplementary Figure 4.6. Recombination landscape of SP population relative to the *S.pimpinellifolium* reference genome. The colored, black and gray lines represent the recombination rate, SNP distribution and repeat content, respectively.



Supplementary Figure 4.7. Recombination rate of selective sweep genes. Effective recombination rates per population of domestication (DSG) and non-domestication (nDSG) sweep genes overlapping and not overlapping hotspots. We also plotted recombination for the same plot but for improvement sweep genes (ISG). Each panel is the set of hotspots per population in which genes are compared against. ***h*** and ***nh*** mean hotspots and non-hotspots, respectively.



Supplementary Figure 4.8. Deletion frequency and hotspot suppression. (**A**) Allele frequency (AF) of the detected deletions in each population. The median AF of SP, SLC, and SLL are 0.57, 0.11 and 0.01, respectively. Data is taken from Wang *et al.*, 2020. A permutation test was performed to examine the overlap between *S. pimpinellifolium* hotspots and deletions (> 500bp) for various thresholds on allele frequencies. (**B**) At a maximum allele frequency of 0.1 to 0.5, the difference between the observed overlap (green dot) and the expected overlap (red dot) is not significant. (**C**) When limiting minimum allele frequency, we observed that all groups show a significant difference between the observed and expected overlap. However, at a frequency of 0.5, the observed overlap starts decreasing. Deletions with lower allele frequencies tend to be associated with less suppression of hotspots in outcrossing populations.



Supplementary Figure 4.9. Excess hotspots in small deletions. We computed the difference between the observed and expected overlap between deletions and historical recombination hotspots of (**A**) wild tomato (SP), (**B**) early-domesticated tomato (SLC) and (**C**) vintage tomato (SLL). The black and red vertical lines indicate the mean overlaps in 10,000 permutation sets and the number of overlaps at $P = 0.05$, respectively. The green vertical line indicates the observed number of overlaps.

Chapter 5

Aberrant patterns of recombination in interspecific tomato hybrids reveal breeding bottlenecks

Roven Rommel Fuentes, Ronald Nieuwenhuis, Jihed Chouaref, Thamara Hesselink, Willem van Dooijeweert, Hetty C. van den Broeck, Elio Schijlen, Henk J. Schouten, Yuling Bai, Paul Fransz, Maike Stam, Hans de Jong, Sara Diaz Trivino, Dick de Ridder, Aalt D.J. van Dijk, Sander A. Peters

Abstract

Increasing natural resistance and resilience in plants is key for ensuring food security within a changing climate. Breeders improve these traits by crossing cultivated tomatoes with their wild relatives and introgressing specific alleles through meiotic recombination. However, some genomic features seem to hamper recombination especially between divergent genomes. Here, we used pooled-pollen sequencing to reveal unprecedented details of recombination patterns in five interspecific tomato hybrids. We detected hybrid-specific recombination coldspots that underscore the influence of structural divergence in shaping recombination landscapes. Crossover regions and coldspots show strong association with specific TE superfamilies exhibiting differentially accessible chromatin between somatic and meiotic cells. We also found gene complexes associated with domestication syndrome traits and stress resistance, revealing undesired consequences of linkage drag. Finally, we demonstrate a way to overcome recombination barriers by finding alternative parental genomes. Overall, our results will allow breeders to make better informed decisions on generating new tomato varieties.

5.1 Introduction

Crop breeding relies on the availability of genetic diversity to generate novel allele combinations that are agronomically valuable. However, long term selection by inbreeding often causes loss of essential allelic information. To reintroduce lost genetic variation, breeders have introgressed alien chromatin by crossing crops with wild relatives, followed by repeated backcrossing and selection. Among the most desirable traits to be incorporated into the breeding material are abiotic stress tolerance and disease resistance, higher yield, and fruit quality (Bailey-Serres *et al.*, 2019). The success of introgression breeding largely depends on the process of recombination to introduce genetic material from the donor into the recipient crop. Meiotic recombination generates genetic diversity, but may also break apart co-adapted allele combinations, resulting in fitness reduction (Ortiz-Barrientos *et al.*, 2016). Moreover, low frequency or complete absence of recombination in a genomic region leads to linkage drag and limits the ability of breeders to develop novel allele combinations. Chromosome regions where recombination is suppressed are found in pericentromeres, including retrotransposons and other DNA-methylated regions (Mercier *et al.*, 2015; Underwood and Choi, 2019). Furthermore, heterozygous structural variants (SVs) have been reported to limit pairing and crossovers (COs) or lead to lethal gametes, suggesting that genomic rearrangements affect recombination patterns, especially in hybrids (Seah *et al.*, 2004; Rowan *et al.*, 2019; Crow *et al.*, 2020; Wang and Copenhaver, 2018).

Genomic rearrangements may exist between related species and different genotypes of the same species. Characterization of these rearrangements has revealed recombination coldspots, some of which are associated with resistance genes or adaptive traits (Szinay *et al.*, 2012; Jiao and Schneeberger, 2020). Due to absent or diminished COs in SV regions, clusters of tightly linked alleles known as supergenes are inherited together, contributing to local adaptation and reproductive isolation (Schwander *et al.*, 2014; Thompson and Jiggins, 2014; Kirkpatrick, 2010). Suppression or absence of recombination has been found essential in speciation and domestication by allowing the fixation of alleles such as those within selective sweeps (Moyers *et al.*, 2018; Fuentes *et al.*, 2022). One of the best studied rearrangements in plants is the 1.17Mb paracentric inversion in Arabidopsis, which shows complete lack of recombination in the rearranged genomic segment linked with fecundity under drought (Fransz *et al.*, 2016). It was reported that recombination is prevented by SVs in genomic regions causing self-incompatibility in Brassicaceae plants (Casselman *et al.*, 2000) and reproductive isolation in monkeyflower (Lowry and Willis, 2010). In the backcross descendants of a *Solanum habrochaites* introgression into cultivated tomato (*S. lycopersicum*), an inversion containing the *Ty-2* resistance genes and at least 35 more genes causes linkage drag, rendering selection of appropriate agronomic trait combinations in the offspring impossible (Yang *et al.*, 2014; Wolters *et al.*, 2015). Another example is the lack of CO in the inverted region of a *S. esculentum* x *S.*

peruvianum cross*,* containing the nematode-resistance gene, Mi-1, and other genes conferring resistance to different pathogens and insects (Seah *et al.*, 2004).

Although previous studies addressed the role of SVs as recombination barriers in a limited number of genomic regions, a genome-wide analysis of decreased or absent COs related to SVs in tomato and multiple hybrid crosses is currently lacking, due to the absence of cost-effective and high-resolution crossover detection methods and accurate SV prediction. The effect of structural differences on chromosome pairing during meiosis has been studied using electron microscopy (Anderson *et al.*, 2010), comparing spreads of synaptonemal complexes (SCs) from multiple F1 tomato hybrids. The synaptic configurations revealed mismatched kinetochores, inversion loops and translocation complexes, pointing to structural differences in the parental genomes that likely influence the recombination landscape. However, electron microscopic studies could not reveal the consequences of erratic recombination patterns between the parental partners. To analyze these patterns in higher resolution, Demirci *et al.* (2017) sequenced F_6 recombinant inbred lines obtained from a cross between tomato (*S. lycopersicum*) and its wild relative *S. pimpinellifolium*, and computationally detected recombination sites. We subsequently developed a less laborious and costly method, involving pollen profiling. Using a pool of pollen from *S. pimpinellifolium* x *S. lycopersicum* F1 hybrids, we generated a recombination landscape at nucleotide resolution level (Fuentes *et al.*, 2020), revealing significant reduction of COs in heterozygous deletions (Fuentes *et al.*, 2022).

To better understand occurrence and frequencies of CO events, we profiled here the recombination landscape in multiple crosses of tomato and wild relatives by sequencing pools of pollen gametes. We identified CO coldspots in each hybrid cross and examined recombination patterns and barriers. Our results suggest a major role for SVs and transposable elements in shaping the recombination landscape in hybrids, specifically in suppressing COs in gene complexes that relate to adaptation, speciation, and domestication. In addition, we present an example of syntenic and non-syntenic accessions for specific genomic regions, which may be considered in the selection of parental breeding lines as so called 'bridge accessions' to avoid or overcome introgression bottlenecks.

5.2 Results

5.2.1 Crossovers in multiple hybrid crosses

In this study, we have generated hybrid crosses of *S. lycopersicum Heinz1706* and its wild relatives *S. pimpinellifolium* (CGN14498; **PM**)*, S. neorickii* (LA0735; **NE**), *S. chmielewskii* (LA2663; **CH**)*, S. habrochaites* (LYC4; **HB**)*,* and *S. pennellii* (LA0716; **PN**). Hereafter, we use these abbreviations and the species name when referring to the hybrids and the parental genome, respectively. The pool of pollen from each hybrid was sequenced using 10X Genomics kits (Supplementary Table 1) based on the protocol described in Fuentes *et al.* (2020). To detect crossover events (COs),

Table 1. Crossovers detected in multiple interspecific (with *S. lycopersicum*) hybrid populations

Hybrid Cross	**Number of SNPs**	**Number of COs**	**Distance** (kb; 1/Resolution)	**Distal euchromatin genes** (*p*-val) *	**Pericentric heterochromatin genes** (*p*-val)*
***S. pimpinellifolium* (PM)**	4,742,049	1,040	2.3 ± 1.4	9.3×10^{-3}	8.2×10^{-5}
***S. neorickii* (NE)**	13,749,445	1,700	2.3 ± 1.5	2.1×10^{-107}	4.3×10^{-20}
***S. chmielewskii* (CH)**	13,770,207	1,618	2.2 ± 1.5	2.3×10^{-104}	4.4×10^{-16}
***S. habrochaites* (HB)**	14,909,955	832	1.9 ± 1.5	6.9×10^{-66}	1.5×10^{-9}
***S. pennellii* (PN)**	15,447,841	1,192	2.1 ± 1.6	1.2×10^{-86}	7.8×10^{-30}

*Enrichment of COs in genes based on permutation test

we first profiled single-nucleotide polymorphism (SNP) markers. We then filtered out regions prone to false positive COs, manifested by a high density of heterozygous SNPs and excessive sequence coverage (Supplementary Figure 5.1). *S. pennellii* and *S. pimpinellifolium* were the most distant and closest species to *S. lycopersicum* in this study, respectively, and have the highest and lowest number of SNPs with respect to the reference genome (*S. lycopersicum*; SL4.0), respectively (Table 1). Using the filtered SNPs, we were able to detect haplotype shifts, leading to identification of putative recombinant haplotypes. These were further screened as described in the Methods (Supplementary Figure 5.2). In each recombinant molecule, CO events or sites are reported based on the bounding SNP markers.

We detected a total of 6,382 COs in all hybrids, mostly located in distal segments of chromosomes, which is consistent with previous reports in tomato and other plant species (Mercier *et al.*, 2015; Wang and Copenhaver, 2018). For each hybrid, COs are confined to 6-12% of the distal euchromatin (DEU) and below 1% of the pericentric heterochromatin (PCH). In total, CO regions account for only 2% of the whole genome, consistent with other eukaryotic organisms where recombinations are concentrated in hotspots (Choi and Henderson, 2015; Mercier *et al.*, 2015; Lambing *et al.*, 2017; Wang and Copenhaver, 2018). Although tomato PCH regions are known to exhibit low recombination rates (Termolino *et al.*, 2016; Yelina *et al.*, 2015), we detected there a total of 710 COs (11.1%) in all hybrids, which most likely locate in euchromatin islands in the PCH. It was proposed that suppression of double-strand-breaks (DSBs), the precursor of COs, by heterochromatin on repetitive DNA helps safeguard against genome destabilization (Tock and Henderson, 2018; Yelina *et al.*, 2015).

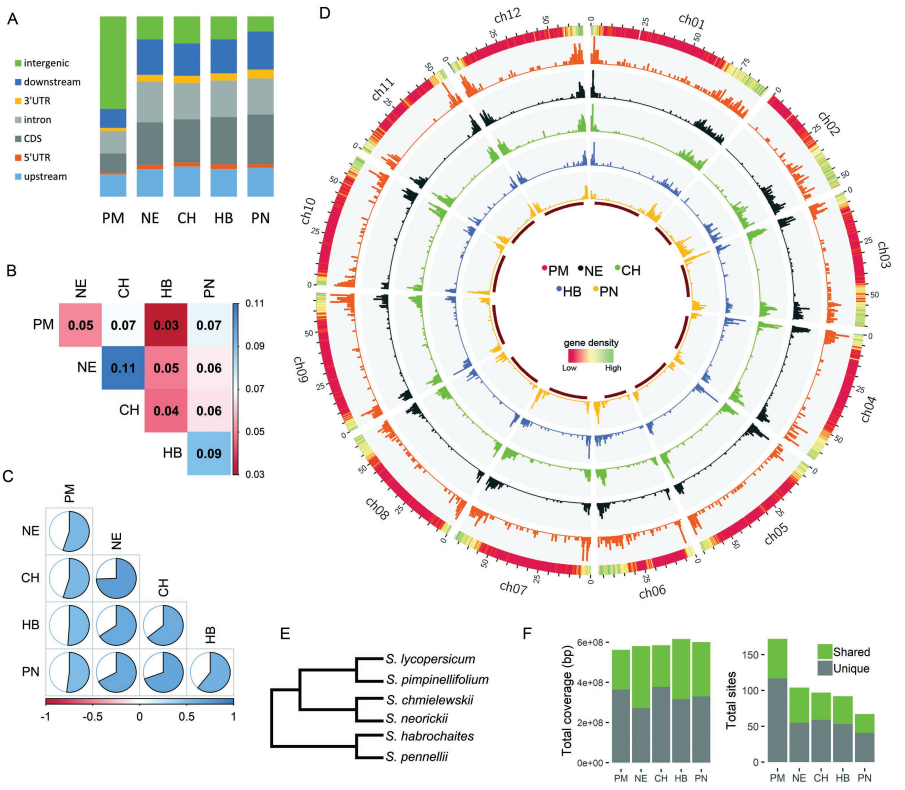


Figure 5.1. Recombination landscapes. A) Distribution of crossover regions over gene features. Upstream and downstream covers 1 kb from the transcription start and termination sites, respectively. B) Fraction of shared CO sites and C) Correlation of the genome-wide CO landscape between hybrids. D) Distribution of COs per hybrid. The outermost track indicates gene density while the red innermost track marks the pericentric heterochromatin regions. E) Phylogenetic tree of the parental species based on Moyle (2008). F) Coverage and number of recombination coldspots in different crosses.

We validated the resulting recombination profile of PM by comparison to existing CO data. The COs in the pollen gametes significantly overlap with COs previously detected in a RIL population of the same parental cross (Fisher's exact test; $P = 5.8 \times 10^{-18}$) (Demirci *et al.*, 2017). More in detail, the frequency of COs in sliding genomic windows in DEU or in PCH also revealed a significant correlation between the recombination landscapes generated from pollen and the RIL population sequence data (Spearman's rank correlation; distal euchromatin, $\rho = 0.33$; $P < 2.2 \times$

10^{-16}; pericentric heterochromatin, ρ = 0.22; P < 2.2 x 10^{-16}). Furthermore, comparison with historical recombination hotspots detected in natural populations of wild and domesticated tomato (Fuentes *et al.*, 2022) revealed that the COs in hybrids overlap with 294 (Fisher's exact test; P = 2.0 x 10^{-13}) and 36 (Fisher's exact test; P = 3.8 x 10^{-11}) historical hotspots in DEU and PCH, respectively. Previous observations and our results both confirm recombination sites in PCH, which thus far were rarely observed due to their low frequency and the limitations of other CO-detection methods.

The vast majority (5,150) of COs are located within genes and their 1kb flanking regions, while another 471 are positioned between 1kb and 3kb from genes (Table 1; Figure 5.1A; Supplementary Figure 5.3). PM and NE have the lowest CO resolution, defined as the inverse of the distance between the SNP markers bounding the CO site (resolution = 1/distance). We consider a detected CO as high resolution if the distance between the markers flanking the CO site is below 1kb, *i.e.* if the resolution is above 0.001. The number of COs near or within genes, in both DEU and PCH regions, is significantly higher than expected by chance (Table 1). Although genic regions account for only 15.5% of tomato genome, the majority of COs overlap gene features (Figure 5.1A). *S. pimpinellifolium* COs apparently overlap more with intergenic regions than COs in other hybrids. One possible confounding factor could be that *S. pimpinellifolium* has fewer SNPs than the other species (Table 1), which may lead to a lower resolution of detected COs. To determine if this contributed to the higher overlap between gene features and COs in PM, COs with similarly distributed resolution for all hybrids were separately analyzed (Supplementary Figure 5.4A). However, the result still shows the same higher intergenic overlap of crossover events in PM (Supplementary Figure 5.4B).

Aside from association with genes, sequence motifs are discovered at CO sites as well (Wijnker *et al.*, 2013; Demirci *et al.*, 2017; Shilo *et al.*, 2015). We found that CTT-like repeats, poly-AT and A-rich motifs are actually enriched in regions flanking rather than within high- resolution COs (Supplementary Figure 5.5). These CO motifs have been previously identified in other plant species but, due to low resolution, it was not possible to determine whether they were actually located within or just near CO sites. However, their close proximity to CO sites hints that they may have a role in recruiting recombination-promoting factors as previously proposed (Choi *et al.*, 2018) .

5.2.2 Unique recombination patterns between hybrids

All hybrids show similar recombination landscapes with COs mostly in distal, gene-rich chromosome regions. Yet there are also unique, local patterns of COs, shown in Figure 5.1D. Comparisons of recombination profiles from different hybrids are essential to learn about variability and genomic factors contributing to CO patterns. To examine similarities between hybrids, we first identified overlapping COs and found a significantly higher fraction than expected by chance (Figure

5.1B). The highest overlap of COs is observed between hybrids with wild parents that are evolutionarily closely related to each other (NE and CH or HB and PN). On the contrary, CO sites in PM have more overlap with PN than with other closely related species, which does not reflect their evolutionary distance (Figure 5.1E). A low but significant overlap has also been observed when comparing recombination hotspots in natural populations of wild and domesticated rice, cocoa and tomato (Marand *et al.*, 2019; Schwarzkopf *et al.*, 2020; Fuentes *et al.*, 2022). About 35% (1,161) of CO regions, containing a total of 3,996 CO events, are shared by at least two hybrids. CO regions per hybrid cover around 2% of the genome, whereas they cover 10% (77.6 Mbp) when combined, apparently not extensively overlapping, thus indicating divergent CO regions between the hybrids.

Given the low rate of CO region overlap between hybrids, we decided to investigate whether the overall recombination landscapes across the genome are significantly correlated. Figure 5.1C shows that NE and CH have the most similar landscape. The low CO overlap (4%; Figure 5.1B) between CH and HB does not translate to a low landscape correlation ($\rho = 0.64$); similarly, despite the high overlap between PM and PN COs (7%), the correlation between their landscapes is one of the lowest ($\rho = 0.52$), consistent with their evolutionary distance. Although the number of overlapping COs is significant, it is far less than the number of non-overlapping COs, that contribute more to shaping the overall recombination landscape. This result suggests that despite the similar overall landscape, the hybrids exhibit local differences in CO patterns.

The patterns of genomic regions without recombination in the hybrids differ as well. To analyze these patterns, we identified CO coldspots of more than 1Mb and found that they cover 72-79% of the genomes, with the highest coverage in HB and PN. Grouping by genomic position and size, we assigned coldspots into 325 *unique* and 101 *shared* clusters (Figure 5.1F), with 63.6% of the genome (6.4Mb euchromatic; 485Mb heterochromatic) lacking CO in all five hybrids, which we refer to here as *conserved* coldspots. PM has significantly shorter coldspots than the other hybrids (pairwise Wilcoxon rank-sum test; $P < 1.4 \times 10^{-2}$) and a large number of unique coldspot regions. These divergent patterns of CO regions and coldspots confirm that hybridization of tomato with different wild parents results in variable recombination bottlenecks, uncovering the additional complexity in breeding.

5.2.3 Absence of COs in structural variant heterozygosity

With the results above indicating clear variation in the occurrence of COs in the different hybrids, we speculated that large genomic rearrangements between species may underlie the varying patterns of recombination. To investigate this, we detected SVs between the parental species *S. lycopersicum* and the wild relatives. Furthermore, given that heterozygous SVs may exist in the wild species genomes, allowing the F1 hybrid to inherit an allele that is similar to the reference genome, we also genotyped SVs in the F1 hybrid pollen sequences and retained only the

heterozygous ones (Figure 5.2A). Combining all parental wild species genomes, we detected 59,265 SVs with size above 50bp. We found more deletions than inversions, which may be due to either the inherently low frequency of large inversions (Alonge *et al.*, 2020) or the difficulty of detecting inversions compared to deletions (Figure 5.2B). Among the wild genomes, HB and PN have the highest number of SVs, which are also significantly longer than the other parental genomes (Supplementary Figure 5.6). To check the accuracy of the filtered SV set, we manually verified SVs from *S. pennellii* using dot plots between *S. lycopercisum* and *S. pennellii* assemblies (Supplementary Figure 5.7). 88% of 50 randomly selected deletions are supported, while an additional 10% belong to more complex translocation events and the remaining 2% are false positives. For inversions, we found 76.7% true positives.

To further examine the relationship between SVs and recombination, we identified rearrangements and syntenic regions between *S. lycopersicum* and *S. pennellii* assemblies and compared them against PN COs. We found that 94% of PN COs are in syntenic segments in distal chromosomal regions (Fisher's exact test; P <0.001; Supplementary Figure 5.8), which corresponds to the essential role of synteny in synapsis and crossing-over of homeologous chromosomes during meiosis (Jiao and Schneeberger, 2020; Shen *et al.*, 2019). Using a permutation test, we indeed found strong reduction of recombination in SVs across all hybrids, specifically for SVs larger than 1kb (Figure 5.2H). Further analyses will only use SV larger than 1kb (Supplementary Figure 5.9). About 62-74% of SVs in the wild genomes overlap with coldspots, which may relate with the absence of recombination. Most SVs are located a few to tens of kilobases away from COs (Figure 5.2F), similar to findings in *A. thaliana* (Rowan *et al.*, 2019), but SV size is not correlated to distance from the CO site (Supplementary Figure 5.10).

Given that DEU and PCH in tomato have distinct genomic features, we examined their SV composition and found more SVs in DEU than in PCH regions, with an average ratio of 1.55 to 1. This agrees with previous observations that wild and domesticated tomato accessions have higher SV density in DEU than in PCH (Alonge *et al.*, 2020; Wang *et al.*, 2020). In addition, SVs in PCH are on average longer than those in DEU (Wilcoxon rank-sum test; $P < 5.8 \times 10^{-16}$; Supplementary Figure 5.11). We also observed that in PCH, higher genome coverage by SV regions comes with lower coverage by CO events (Figure 5.2C). PM has the largest total number of CO events in PCH, while PN has the lowest number. As these PM COs overlap with the SVs in the other wild genomes, it seems that the higher SV content in other wild genomes leaves less sites for recombination in hybrids. Given that there are other complex rearrangements and SV types that we cannot detect with our data, it is likely that more divergent CO sites are defined by the presence or absence of SVs.

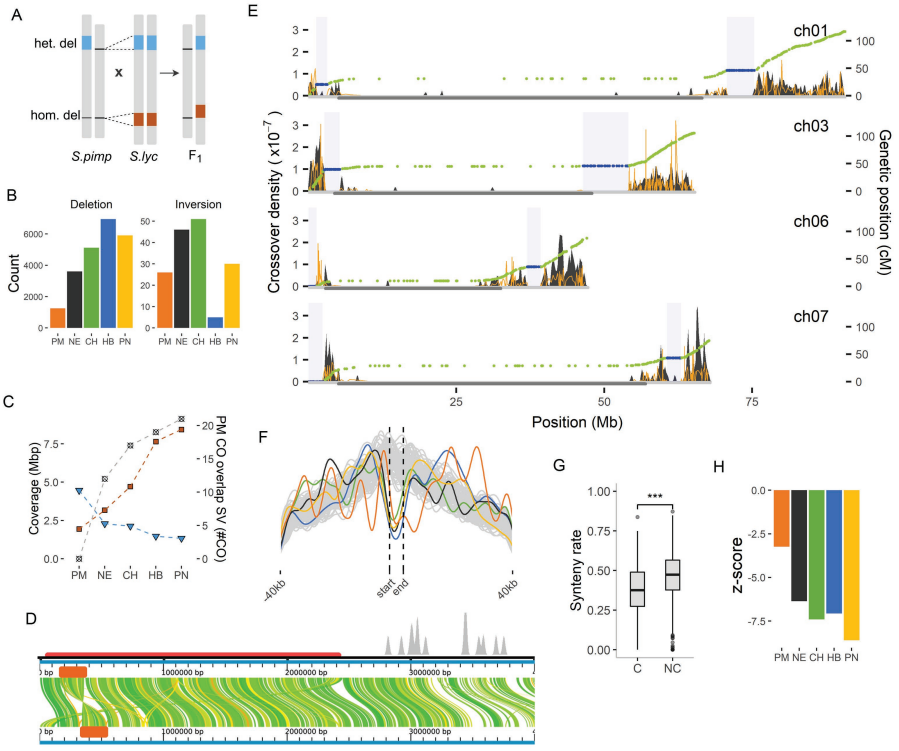


Figure 5.2. Lack of crossover in structural variations. A) Selection of parental SVs causing heterozygosity in the F1 pollen genomes. B) Frequency of SVs per wild relative. Inversions only include events > 30 kb. C) Genome coverage of COs (blue) and SVs (orange) in the PCH (left y-axis). The gray squares show the number of PM COs that overlap with SV regions in the wild genome (right y-axis). D) Large inversion (orange block) and rearrangements within the PN coldspot (horizonal purple segment) of chromosome 7, short arm. CO density is indicated in grey at the top. E) Crossover density of selected PN chromosomes (gray peaks) plotted together with Marey map (green dots) of EXPEN2012 (Sin et al., 2012). The blue dots are genetic markers within coldspot regions (blue box). The yellow distribution line indicates the recombination rate obtained by taking the derivative of the Marey map. The gray horizontal segment in the middle of the chromosome marks the PCH. F) Distance of COs to the nearest SV compared to the 10,000 permutation sets represented by gray lines. The vertical lines marks the boundaries of COs. G) Rate of synteny in coldspot (C) and non-coldspot (NC) regions of PN (Wilcoxon rank-sum test; $P < 2x10^{-16}$). H) Suppression of COs in SV regions based on permutation test. The negative z-score means the overlap of COs in SV regions is lower than expected by chance.

We identified large parts of the DEU in PN with prominent spots without CO. To validate whether these represent real coldspots, we compared them against the recombination coldspots in the EXPEN2012 linkage map (Sim *et al.*, 2012). First, we identified large DEU coldspot regions in the linkage map by mapping the EXPEN2012 markers against the tomato reference genome, retrieving the physical position and subsequently plotting against the genetic position (Figure 5.2E). Then, we compared the EXPEN2012 coldspot against the PN coldspots. Large coldspots are observed in some chromosomes, spanning 0.14 to 7.64 Mb, and they match the coldspots we found in PN, demonstrating the accuracy of our method. Unlike the coarse-grained genetic map, the fine-scale recombination profile we generated allows comparison with genome features, aiding the elucidation of factors influencing recombination landscapes.

Further inspection of these large PN coldspots revealed that they have significantly lower levels of synteny compared to non-coldspots (Figure 5.2G). These coldspots, however, may be specific to PN or may not fully overlap coldspots in other hybrids, as we have found 518 COs in the other hybrids. Among the PN coldspots, we found that at least two, specifically in the short arm of chromosomes 6 and 7, contain large inversions relative to the reference genome as previously validated using BAC-FISH (Szinay *et al.*, 2012). They may also correspond to the inversion loops found at the distal chromosome ends and in the euchromatin-heterochromatin borders (Anderson *et al.*, 2010). We were able to identify the exact location of an inversion in chromosome 7 (Figure 5.2D) by comparing genome assemblies and inspecting linked reads (Supplementary Figure 5.12). Aside from the inversion, this 2.4 Mbp coldspot region also contains other rearrangements, like translocations, that could inhibit proper synapsis and recombination. Upon examining the other large coldspots, we similarly found complex rearrangements and large insertions and deletions. Across all hybrids, our results suggest that SVs contributed significantly to shaping the recombination patterns by inhibiting COs, which may have been vital in the fixation of specific alleles during domestication (Lye and Purugganan, 2019; Fuentes *et al.*, 2022). Importantly, the SVs that have been implicated with domestication of tomato can cause heterozygosity during hybridization with wild relatives, which consequently suppresses recombination.

5.2.4 Widespread coldspots in TE regions

Aside from SVs, studies on other species also linked the presence of transposable elements (TEs) with CO incidence, specifically retrotransposons with COs suppression (Underwood and Choi, 2019). In tomato hybrids, most retrotransposons (Class I), except SINEs and RTE-BovBs, indeed show suppression of COs (Figure 5.3A). However, *Stowaway* and *Tip100* (Class II), simple repeats and low complexity regions are enriched with COs. TEs associated with CO suppression are densely distributed in the PCH, whereas *Stowaway* and *Tip100* are located mostly in the DEU (Figure 5.3C). Similarly, this association with TE superfamilies

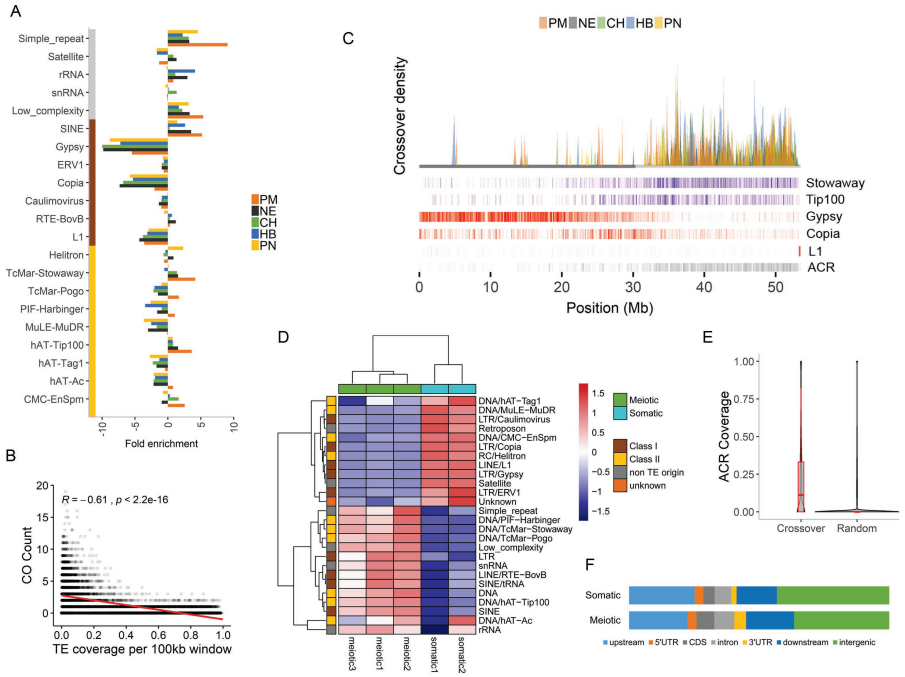


Figure 5.3. TE-associated crossovers. A) TE superfamilies and repeats showing enrichment of COs. Elements are clustered into DNA transposons (yellow), retrotransposons (brown) and other repeats (gray). B) Spearman's rank correlation of crossover count and retrotransposons (*Gypsy*, *Copia*, *L1*) coverage in a sliding genome window. Each dot indicates a window. The red line is the local regression fitting. C) Recombination landscape of acrocentric chromosome 2 from multiple hybrids (colored peaks) with layers of density heatmaps representing different features, including class I (red) and II (blue) TEs, and meiotic ACRs (gray). The horizontal grey line represents the PCH. D) Normalized enrichment of ATAC-seq read coverage over repetitive elements of meiotic and somatic cells. E) Total coverage of ACR per region. F) Total ACR coverage per genome feature. Upstream and downstream covers 1 kb from the transcription start and termination sites, respectively.

was reported in historical recombination hotspots of wild and domesticated populations of tomato (Fuentes *et al.*, 2022). As shown in Figure 5.3B, the presence of retrotransposons such as *Gypsy*, *Copia* and *L1* in a genomic region correlates with CO suppression, consistent with reports in many other species (He *et al.*, 2017; Kent *et al.*, 2017; Choi *et al.*, 2018; Marand *et al.*, 2017; Pan *et al.*, 2017; Marand *et al.*, 2019). In contrast, *Stowaway* and *Tip100* show positive correlation with CO incidence (Supplementary Figure 5.13).

The enrichment of COs correlates with lower nucleosome occupancy and reduced DNA methylation (Choi *et al.*, 2018). To investigate the chromatin state of

TE elements with and without COs, we performed an ATAC-seq analysis of *S. lycopersicum* meiotic and somatic cells and found 52,802 and 25,101 accessible chromatin regions (ACRs), respectively. These ACRs have an average size of 733bp and represent accessible chromatin in the *S. lycopersicum* parent. Read distributions over the genome were highly correlated between biological replicates (Supplementary Figure 5.14). Based on a permutation test, we found significant overlap between COs and meiotic ACRs (z-score = 87.2), confirming reports that COs occur in regions accessible to recombination machinery. Figure 5.3E shows that crossover regions have higher ACR coverage compared to random genomic regions. Upon comparing meiocyte ACRs with TEs, we found that TE superfamilies enriched with COs have an accessible chromatin segments, whereas retrotransposons like *Gypsy*, *Copia* and *L1* are not associated with accessible chromatin (Figure 5.3D). This is similar to reports in *A. thaliana* of DNA transposons showing nucleosome depletion and higher SPO11-1-oligo levels (Choi *et al.*, 2018). Moreover, retroelements like *Gypsy*, *Copia* and *L1* have very few SPO11-1-oligos, and high DNA methylation and nucleosome occupancy. This association of CO with specific class I and class II TEs is also observable in the landscape in Figure 5.3C**,** where the former are densely distributed in the PCH while the latter are predominantly found in DEU. Furthermore, the chromatin accessibility of TE superfamilies flips between somatic and meiotic cells, hinting at a preference to keep specific superfamilies inaccessible during meiosis (Figure 5.3D). The differential ACRs suggests that TE superfamilies may have different roles or activities in different tissue types and in relation to recombination. Our results emphasize the major role of chromatin structure in the suppression or enrichment of COs in TEs and the need to particularly analyze meiocytes to account for tissue-specific ACRs.

Similar to the association of COs with proximal promoter regions (Fuentes *et al.*, 2020), it was previously reported that ACRs are strongly associated with transcription start sites (TSSs) (Qiu *et al.*, 2016). To evaluate this, we examined the average ATAQ-seq signal in genes and their flanking regions, and found the highest coverage at the TSS in both meiotic and somatic cells (Supplementary Figure 5.15). We also found that the majority of ACRs are located near or within genes (Figure 5.3F; Fisher's exact test; $P < 0.05$), similar to COs (Figure 5.1A). Normalized by the total genome coverage of the feature, the promoter regions and the UTRs (untranslated regions) have the highest ACR density.

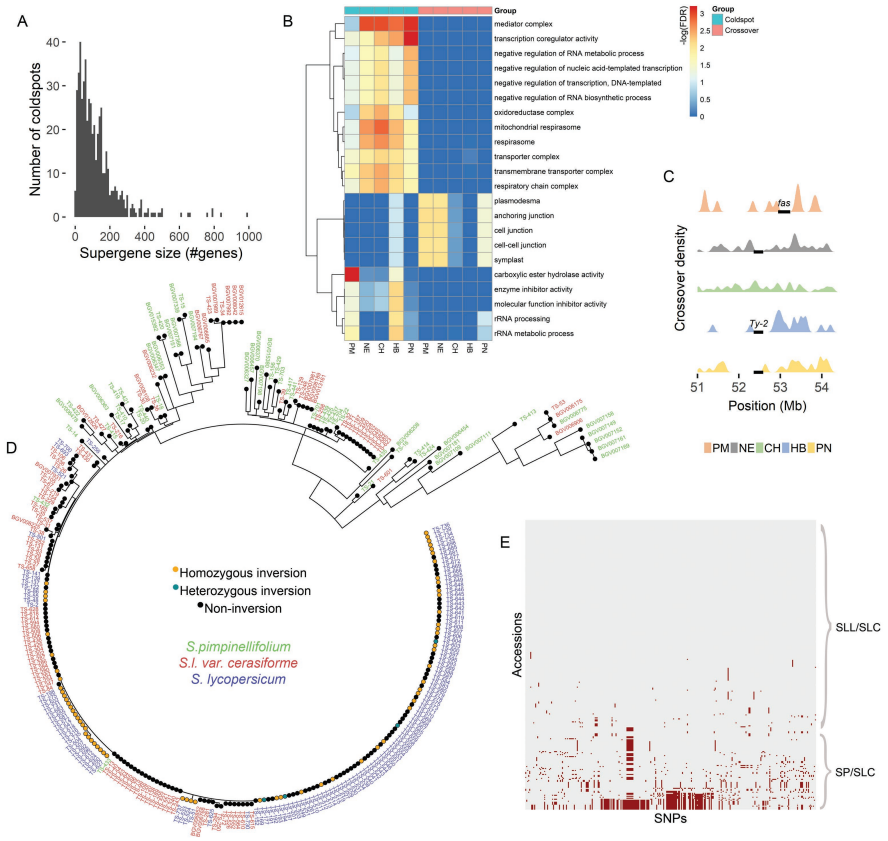


Figure 5.4. Supergenes and inversions. A) Sizes of supergene clusters in CO coldspots. B). Gene ontology (GO) terms enriched (at least 2x) in coldspot and crossover regions. C) Recombination coldspots in chromosome 11 long arm. D) Phylogenetic tree of genic SNPs in an inversion region of wild and domesticated tomato accessions. E) SNPs w.r.t S. lycopersicum within the inversion region.

Aside from the fact that many SVs are generated by *Gypsy* and *Copia* retrotransposons (Alonge *et al.*, 2020), 67% of coldspots we detected are covered by these TE elements for at least 50%. About 98.6% of the conserved coldspots are in PCH where retrotransposon presence is dense (Choi *et al.*, 2018). Furthermore, the retrotransposon families that are linked with CO suppression cover 450Mb (~52%) of the tomato genome, implying the wide span of suppression due to retrotransposons. This underscores the importance of transposable elements in shaping recombination patterns, both in hybrids and inbreeding materials and predominantly in regions with high retrotransposon density.

5.2.5 Supergenes and breeding bottlenecks

COs tend to occur near genes but certain genomic elements prohibit the recombination between loci, causing co-segregation of these loci to the offspring. About 62% of genes are in CO coldspots of one or more of the hybrids and 484 of these coldspots contain at least 20 genes (Figure 5.4A). Gene complexes within coldspots are supergenes in a specific hybrid, but the same supergenes are not necessarily found in other hybrid tomato crosses. Although many supergenes are located in the conserved coldspots in PCH, other supergenes are located in the 81 coldspots in gene-dense DEU. In a gene ontology (GO) enrichment analysis of crossover and coldspot regions (Figure 5.4B), we found that 45 biological processes, 14 molecular functions, and 48 cellular components are significantly enriched (false discovery rate < 0.05) and that overrepresented GO terms in coldspots are associated with basal housekeeping functions (e.g like transcription coregulator activity, transporter complex, rRNA processing, metabolic processes). A more detailed list of enriched GO terms is reported in Supplementary Figure 5.16. Interestingly, we identified multiple metabolic processes enriched in the coldspots (Supplementary Figure 5.16A), which may reflect the evolutionary divergence between tomato and the wild species. Many of the coldspot genes are directly related to modification in metabolism, which is considered a prominent manifestation of the domestication process (Tieman *et al.*, 2017; Alseekh *et al.*, 2021; Sauvage *et al.*, 2017; Zhu *et al.*, 2018). Supergenes, which are mostly generated by inversions or translocations, have actually been linked to metabolic pathways and alternative phenotypes in plants (Schwander *et al.*, 2014; Alseekh *et al.*, 2021).

To further investigate links between coldspots and phenotypes, we characterized the supergenes residing in the large coldspots in PN (Figure 5.2E). These coldspots contain 2,736 genes, 877 of which have been identified as domestication syndrome genes (Lin *et al.*, 2014). The coldspots in the short arms of chromosome 6 and 7, coinciding with the inversion as previously reported by Szinay *et al.* (2012), contain 130 and 295 genes, respectively, and are associated with responses to oxidative stress ($P = 1.27 \times 10^{-4}$) and specific catabolic and metabolic processes (9.66×10^{-8}) (Supplementary Figure 5.16). Similarly, the coldspots in both arms of chromosome 1 contain genes involved in metabolic processes of organic substances ($P = 2.26 \times 10^{-5}$) and transcription coregulator activity ($P = 6.06 \times 10^{-5}$). This result does not simply imply association between metabolic processes and coldspots, but it underlines the highly constrained metabolic profile in hybrids or in tomato introgressed with these coldspot regions.

Aside from the rewiring of the metabolome, we are interested whether domestication is associated with linkage drag in regions containing resistance (*R*) genes. Upon inspecting the coldspots in PN, we found that they are enriched with *R* genes (Fisher's exact test; $P = 5.1 \times 10^{-4}$) and at least 29 coldspots (23 clusters) contain *R*-gene hotspots (Supplementary Figure 5.17). The coldspot in

chromosome 7 contains 295 genes, including *R* genes and *chitinase* genes. In this region, we found an enrichment of genes related to the *chitin catabolic process* (FDR = 1.49 x 10^{-4}), *chitin binding* (FDR = 1.63 x 10^{-4}) and *chitinase activity* (FDR = 2.18 x 10^{-2}), which are involved in plant defense responses against pathogens (Zhu *et al.*, 1994; Shrestha *et al.*, 2007; Herget *et al.*, 1990). Our findings are consistent with observations in Arabidopsis, in which CO coldspots with many SVs contain clusters of *R* genes (Choi *et al.*, 2016). Some *R* gene hotspots can become CO hotspots to overcome new pathogens through rapid diversification. In contrast, *R* genes conferring resistance to pathogens with low genetic plasticity are located in CO coldspots, possibly maintained by structural heterozygosity (Nieri *et al.*, 2017; Hulbert *et al.*, 2001). The association between resistance genes and some unfavorable alleles due to genetic linkage limits the introgression of resistance haplotypes into breeding lines. A specific case of linkage drag involving the resistance to *Fusarium* wilt race 3, reduced fruit size and increased sensitivity to bacterial spot, was broken by reducing the size of the introgression (Chitwood-Brown *et al.*, 2021). However, this shrinking of the introgressed region is feasible only because it is not induced by an inversion or other CO-suppressing type of SV, unlike the *Ty-1* and *Ty-2* introgressions which both are located within inversions (Wolters *et al.*, 2015). Aside from *R* genes, the coldspot in chromosome 7 also contain the *SUN* locus, which is linked with variable fruit shape in the wild and cultivated tomato (van der Knaap *et al.*, 2004; Rodriguez *et al.*, 2011). The remaining coldspots in PN contain 17 genes with putative roles in fruit shape determination, further substantiating the association between coldspots and domestication syndrome traits (Huang *et al.*, 2013).

To break linkage drag in an SV region with no recombination, alternative crosses that do not result in coldspots are needed. For the coldspot in chromosome 7 of PN, the other wild relatives can serve as alternative parent as they all exhibit recombination in this region. Another set of examples are located in the chromosome 11 euchromatic long arm, where local patterns of recombination vary between hybrids (Figure 5.4C). One of the variable regions is colocalized with the known *Ty-2* inversion (Wolters *et al.*, 2015), showing complete absence of CO over the full length of the inversion in three hybrids. The other region overlaps with the known recombination-suppressed 294-kb inversion in the *fasciated* (fas) locus with breakpoints in the first intron of a YABBY transcription factor gene (*SlYABBY2b*) and 1 kb upstream of the *SlCV3* start codon (Cong *et al.*, 2008; Huang and van der Knaap, 2011; Xu *et al.*, 2015). This inversion contains 41 genes, including 4 disease resistance genes, and confers a large fruit phenotype to domesticated tomato. All parental lines do not have the inversion, but *S. pimpinellifolium* has several smaller rearrangements that could have caused CO suppression. Furthermore, this inversion is more common in the domesticated tomato accession, which may impose a bottleneck in breeding. Given the resequencing data for populations of wild and domesticated tomato, it might be possible to find “bridge accessions” or accessions without the allele causing

heterozygosity. We therefore screened 56 accessions of wild (*S. pimpinellifolium*; SP), 109 early-domesticated (*S. lycopersicum* var. *cerasiforme*; SLC) and 127 vintage tomato (*S. lycopersicum* var. *lycopersicum*; SLL) that were genotyped for the inversion, including the SNPs within the inversion. All SP and 96% (109) of SLC accessions have non-inversion genotypes while half (64) of the SLL group have at least one inversion allele (Figure 5.4D), which may suggest that the inversion could have occurred or have been selected for during tomato domestication. This is consistent with the drastic reduction of nucleotide diversity in this region when comparing the SLC/SLL and the SP population (Lin *et al.*, 2014). Upon inventorying the inversion and non-inversion accessions, we compared the SNP profile within the inversion region and found haplotypes in SP distinct from those in SLL accessions (Figure 5.4E). We subsequently identified at least 12 SLL accessions without the inversion and with larger fruit weight phenotype compared to SP (Lin *et al.*, 2014). These candidate bridge accessions may be crossed with SP accessions to overcome CO suppression in the inversion region, while maintaining genetic background that confers large fruit other than the *fas* inversion. Aside from this inversion, there are at least 236 additional non-overlapping SV sites in the population of SP and SLC/SLL that may be analyzed to predict recombination barriers, especially lineage-specific rearrangements. Most importantly, this SV profile may be used to select bridge accessions to introgress genetic diversity into genetically eroded domains of crop tomato.

As presented, CO-suppressed regions containing supergenes can limit breeding by linking agronomically beneficial alleles with deleterious or unfavorable alleles. We identified an example of undesirable linkage (Figure 5.5A) by examining

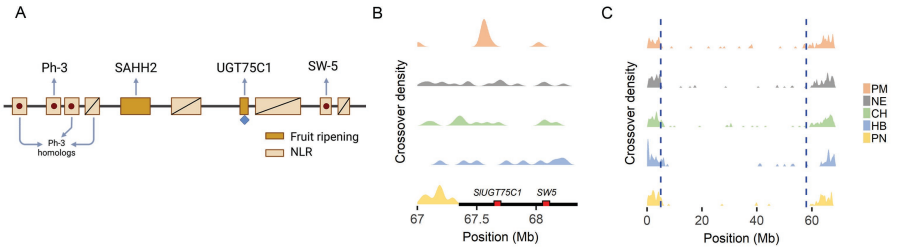


Figure 5.5. Unfavorable linkage. A) Coldspot in PN chromosome 9 containing genes associated with resistance and agronomic traits. Red dots and diagonal lines indicate missense mutations and frameshift mutations, respectively. Blue diamond indicates the differential expression between the parental genomes. B) Recombination landscape for the same coldspot region in all hybrids. The black horizontal line represents the coldspot in PN. C) Rare recombination events in the *Tm-2* locus, which span the region between the two blue lines, in chromosome 9.

genes located in coldspots. We first compiled a list of genes associated with agronomic and resistance traits and compared them against the coldspots in the hybrids. Shown in Figure 5.5B, a coldspot in chromosome 9 of PN spans genes associated with resistance and fruit ripening, linking *SlUGT75C1* and *SW5.* The significantly lower expression of *SlUGT75C1* in *S. lycopersicum* relative to *S. pennellii* may yield accelerated fruit ripening (Sun *et al.*, 2017; Koenig *et al.*, 2013). On the hand, *S. lycopersicum* has the marker ($Sw5a^S$) linked with tomato spotted wilt virus (TSWV) susceptibility (De Oliveira *et al.*, 2016) and lacks the resistance allele (missense mutation) to leaf curl New Delhi virus (ToLCNDV) *(Sharma et al., 2021)*, which we observed in *S. pennellii*. This suggests that in PN, we cannot generate gametes with both accelerated ripening and resistance to ToLCNDV. The tight linkage may be resolved by pairing *S. lycopersicum* with the other wild relatives, which may allow recombination between these genes and may introgress the resistance allele.

A remarkable case of introgression of disease resistance and associated linkage drag of three quarters of a wild chromosome is the Tomato Mosaic Virus (ToMV) resistance from the *S. peruvianum* accession PI 126926. The resistance from this wild species is conferred by the *Tm-2* gene or its allelic variant $Tm\text{-}2^2$, located on ch09 of tomato (Lanfermeijer *et al.*, 2005). Because of the serious damage caused by ToMV in cultivated tomatoes, breeders introgressed the Tm-2 resistance gene already in 1960s into commercial tomato varieties. However, the *Tm-2* mediated resistance was broken by the virus, triggering tomato breeders to introgress the $Tm\text{-}2^2$ allele from the *S. peruvianum* accession PI 18650, which provided more durable resistance (Lanfermeijer *et al.*, 2005). Nowadays more than 90% of the commercial tomato varieties carry the *Tm-2* locus (Schouten *et al.*, 2019). The introgression starts at ~5 Mbp and continues until ~58 Mbp of that chromosome, and exhibits low levels of COs in all evaluated hybrids (Figure 5.1D). However, some COs do occur in our pollen sequencing data (Figure 5.5C). A recently released assembly of a chromosome with this introgression revealed that it contains small and medium sized inversions, and series of relatively small translocations and duplications (van Rengs *et al.*, 2022). Although there are many SVs in this introgression, there are still sufficient syntenic regions that could allow COs. However due to the low frequency of CO in these regions, they may not be observed in limited size offspring populations previously used in tomato breeding. Nevertheless, our results demonstrate a way to identify recombination coldspots leading to undesirable linkage and possibly resolve them by finding alternative parental genomes.

5.3 Discussion

COs are mostly distributed in the gene-rich DEU regions of each chromosome, consistent with previous reports (Demirci *et al.*, 2017; Fuentes *et al.*, 2020). The recombination landscape in one of the hybrid accurately matched a genetic linkage map, underscoring the importance of our method which provide high resolution

CO with less cost and labor. Despite the similar overall CO landscape between hybrids, we discovered fine-scale differences in CO patterns and regions without recombination. CO coldspots have limited ability to reshuffle alleles between tomato and wild species, which hampers introgressive hybridization breeding and reduces efficiency of backcrossing. Although the majority of the coldspots are conserved between all hybrids, some coldspots are unique to a cross, which may serve as putative targets to break linkage drag or to study the underlying fitness advantage that necessitates the suppression of recombination. This is so far the most comprehensive profile of recombination in tomato hybrids and possibly even in plant species.

Across all hybrids, we found conspicuous absence of crossover events in SV regions, particularly in lineage-specific rearrangements. The varying patterns of recombination between the hybrids are associated with rearrangements between the wild parental genomes, implying that SV profiles in F1 progeny may help distinguish regions that may or may not allow crossovers. Knowing these regions enables breeders to fine-tune introgression plans by inspecting recombination patterns in the loci of interest prior to the elaborate hybridization and screening processes. Although multiple studies have already reported the negative association between SVs and COs, it is still not clear how SVs inhibit recombination. Rowan *et al.* (2019) proposed several possible explanations for the observed suppression of COs in heterozygous SVs, such as absence of a repair template, tendency to produce non-viable gametes, DNA methylation in the SV region, and blocking physical interaction in variant regions preventing proper synapsis. Furthermore, it has been reported that DSBs in inversion regions are preferentially resolved as noncrossover gene conversions and not as COs (Crown *et al.*, 2018; Korunes and Noor, 2019; Rowan *et al.*, 2019). Although we have already found an association between CO and SV patterns, further studies must be conducted to improve the detection of SVs, specifically of the insertion and translocation type, for better recognition of the underlying causes of suppression in each coldspot.

Structurally heterozygous regions in the genome, that cause lack of recombinant haplotypes, have been linked to adaptive phenotypes and plant domestication and speciation (Schwander *et al.*, 2014; Thompson and Jiggins, 2014; Lye and Purugganan, 2019; Fuentes *et al.*, 2022). An inversion, capturing two or more alleles adapted to an environment, prevents recombination and confers a selective advantage that subsequently promotes its spread in the population (Kirkpatrick and Barton, 2006). Based on visual examination of the synaptonemal complexes, it was suggested that SVs form interspecific reproductive barriers in the tomato clade (Anderson *et al.*, 2010). We confirmed this by our results on the absence of COs in SV regions of multiple interspecific crosses. Some of these recombination coldspots contain supergenes, which may confer alternative or differentiated phenotypes between the parental genomes (Schwander *et al.*, 2014). This information can help identify unfavorable gene complexes prior to

hybridization, exposing possible undesired consequences of introgression. In this study, we showed that some CO coldspots in interspecific hybrids overlap *R* gene hotspots, which not only accumulated nucleotide variations during the evolution of wild tomato relatives but underwent copy expansion and contraction, conferring varying resistance to pathogens (Seong *et al.*, 2020). However, CO suppression prevents traditional introgression methods from selecting favorable alleles and gene copies in these *R* gene hotspots (Andolfo *et al.*, 2021), curbing the efforts to develop disease-resistant tomato.

Some coldspots contain genes associated with metabolic processes and fruit traits, implying linkage between genes that may relate with the considerable change in chemical composition of tomato fruit due to fruit mass-targeted selection during domestication (Zhu *et al.*, 2018). These coldspots can serve as targets for metabolite engineering in *de novo* domestication of wild tomato relatives. The enrichment of genes in CO coldspots linked with resistance and metabolomes is partly brought about by plant evolutionary events involving SVs (Lye and Purugganan, 2019; Alseekh *et al.*, 2021). Further examination of recombination coldspots can help breeders to understand the genetic or epigenetic cause of CO suppression and determine divergent phenotypes resulting from the evolution of locally adapted alleles and from domestication.

Next to from the association between SVs and CO coldspots, we found specific superfamilies of TEs strongly associated with crossovers and accessible chromatin regions (ACRs). By checking the ACRs in meiocytes, we determined that the varying association between superfamilies may be influenced by their chromatin configuration, keeping elements like *Gypsy* and *Copia* inaccessible during meiosis, which consequently prohibits COs. We also discovered differential chromatin accessibility of TE elements in somatic and meiotic cells, necessitating further studies to explain whether this relate with different functions or the regulation to limit proliferation of specific TEs during meiosis (Thieme *et al.*, 2017). Although we found an association between TEs and COs, it is not clear whether TEs directly shape the recombination landscape, or that recombination and TE insertions simply collocate in ACRs and genic regions because of TE insertion bias (Kent *et al.*, 2017; Choi *et al.*, 2018). In tomato, *Stowaway* elements preferentially insert within or near genes, while *Gypsy* elements insert in pericentromeric regions (Dominguez *et al.*, 2020; Kuang *et al.*, 2009), agreeing with the correlation of COs and TEs. On the other hand, the consistent chromatin state per TE superfamily may indicate that, depending on the type, new TE insertions can either suppress or promote recombination (Kent *et al.*, 2017; Choi *et al.*, 2018). For example, the expansion of pericentromeric regions in *A. alpina* due to retrotransposon insertions resulted in more regions with suppressed recombination(Willing *et al.*, 2015). It would also be interesting to further examine how the activity of TEs, such as during stress exposure, can influence the recombination landscape (Ito *et al.*, 2011; Zervudacki *et al.*, 2018). Our findings make the value of both SV and TE profiles in the parental genomes for CO hotspots and coldspots prediction more apparent.

Previous studies have tried to increase CO frequencies but failed to do so homogeneously along the genome (Ziolkowski *et al.*, 2017; Mieulet *et al.*, 2018; Serra *et al.*, 2018), missing CO coldspots. Recently, it was demonstrated that recombination can be restored by inverting an inversion using genome editing (Schmidt *et al.*, 2020). However, current regulations may restrict the use of such solution to break linkage drag in direct breeding applications. As an alternative, we demonstrated that we can find bridge accessions that can solve the lack of recombination in regions with SVs while maintaining the desired genetic background. However, applicability dependents on whether such accessions exist in nature or not and on the comprehensiveness of the resequencing data. Recent work by Alonge *et al.* (2020) involves the profiling of SVs in 100 accessions that represent the diversity of over 800 tomato accessions, providing more data that can be used in finding compatible genomes. If we cannot find a bridge accession and the SV of interest is heterozygous in one of the parents, it is possible to screen for a homozygous genotype in an offspring population. Nevertheless, we emphasize the importance and advantage of performing compatibility or linkage drag checks, in a cost-effective way, as part of a breeding scheme. Future work can focus on profiling CO-associated features in resequencing data of tomato and wild relative populations and on predicting CO coldspots between a pair of accessions without developing a mapping population.

Linkage drag caused by undesired chromosomal fragments linked and co-introgressed with the desired allele from a wild relative is one of the main hurdles of introgression breeding. Linkage drag may considerably slow down the development of commercial plant varieties due to inheritance of undesired traits genetically linked to the trait-of-interest. Pollen sequencing can provide a fast answer to the question how large a progeny should be to have a reasonable chance of removing the linkage drag, thanks to COs flanking the desired introgressed allele. Pollen of different hybrids can be evaluated as well, screening different parental combinations for CO frequencies neighboring the desired allele from the wild relative, and thereby their suitability for removing linkage drag. This makes the methodology described here not only useful to obtain scientific insight in recombination landscapes, CO hotspots and coldspots, but also of practical value in plant breeding.

5.4 Methods

5.4.1 Sequencing of pollen gametes

We produced F1 plants from crosses between *S. lycopersicum* cv. Heinz1706 and the following wild relatives: *S. pimpinellifolium* (CGN14498), *S. neorickii* (LA0735), *S. chmielewskii* (LA2663), *S.habrochaites* (LYC4), and *S. pennelli* (LA0716). The wild species served as the male parents. Mature pollen were collected from each hybrid and processed to isolate the high molecular weight DNA using the protocol in Fuentes *et al.* (2020). 10X Genomics libraries were constructed according to the

Chromium™ Genome v2 Protocol (CG00043) and then sequenced on an Illumina HiSeq 2500. Aside from the sequencing pool of pollen from these hybrids, we used the same protocols to sequence the inbreds of the parental tomato and the wild species.

5.4.2 Crossover detection

For detecting segregating markers in the hybrids, linked reads from the inbred wild parents were aligned against the *S. lycopersicum* cv. Heinz reference genome SL4.0 (Hosmani *et al.*, 2019) using *Longranger* (Marks *et al.*, 2019) and were subsequently processed using GATK HaplotypeCaller (Poplin *et al.*, 2017) with the recommended hard filtering to screen single nucleotide polymorphisms (SNPs). Heterozygous SNPs and other SNPs located in homopolymeric regions and regions prone to false positives due to inaccurate assembly or copy number variations, resulting in highly heterozygous alignments, were filtered out. Thereafter, for each hybrid, the linked reads from pollen gametes were aligned against SL4.0 using *Longranger* and were phased using the segregating markers as described in Fuentes *et al.* (2020). For each putative recombinant molecule, we applied filters on the resolution, spanning distance, block size, and the number of supporting reads, wild cards and markers per phased block. In the updated version of our pipeline (Fuentes *et al.*, 2020), filtering putative recombinant molecules with significant overlap with repeats and transposable elements was deprecated to enable analysis of correlation between COs and superfamilies of TEs. The number of overlapping crossover events between hybrids and their significance were determined using *bedtools* (Quinlan and Hall, 2010). To compare landscapes, Pearson's correlation matrix was computed on the CO count in 500-kb windows with a 50-kb step size.

5.4.3 Detection of coldspots

We counted the number of COs per hybrid in 10kb sliding windows and merged those windows with at least one CO and within 1kb distance of each other. The resulting set of genomic intervals are considered CO regions. Regions without COs spanning at least 1Mb are considered coldspots. To cluster coldspots from all hybrids, we first grouped those with at least 1 bp overlap. For each group, we built a graph with coldspots as nodes, connected by edges if they have a least 50% reciprocal overlap. Each graph was split into connected components (C) and then based on the genomic position, we computed the distance (p_k) between the leftmost and rightmost coldspot in each component. If p_k is at least 1.5 times the size of the smallest coldspot in C_k, the component was further regrouped by hierarchical clustering using a distance matrix d(i,j) = (f-2*length(i∩j))/(f-length(i∩j)), where i and j is the pair of coldspots in a component and f is the sum of their lengths. Complete linkage was used; the resulting dendrogram was cut at the height of 0.3. The resulting groups were used to define shared coldspots, which occur in at least two hybrids, and unique coldspots, *i.e.* those coldspots that

occur only in one hybrid. We also identified conserved coldspots or regions without CO in all five hybrids.

5.4.4 Detection and validation of SVs

Linked reads from inbreds of all the parental species were aligned to the reference genome and analyzed to detect SVs using *Longranger*. With the presence of heterozygous SVs in the parental genomes, it is possible that only the reference allele may have been inherited by the F1 plants. To determine for each hybrid whether the F1 plants inherited an SV allele causing heterozygosity between the homologous chromosomes during meiosis, we profiled SVs in the F1 pollen linked-reads. The pool of pollen included both recombinant and non-recombinant regions and represented alleles from both parental genomes of each F1 plant. Thereafter, SVs were reported if present in both the inbred and the corresponding pollen data, referred to as *parental SVs*. To further remove problematic regions, SVs between the Heinz reference and the Heinz inbred, which we refer to as *self SVs*, were detected. Lastly, we reported parental SVs, of the deletion (DEL) and inversion (INV) type, that do not overlap self SV. For SV validation, we compared the SL4.0 assembly against the existing assembly of *S. pennellii* (Bolger *et al.*, 2014a) using Syri (Goel *et al.*, 2019) and manually inspected randomly selected sets of DELs and INVs using Gepard (Krumsiek *et al.*, 2007).

5.4.5 Enrichment analysis

To determine the enrichment of COs in specific TE superfamilies, we generated 10,000 permutations of the CO data per hybrid using bedtools and computed the number of overlaps with transposable elements. We then compared the observed and the expected overlap with TE of these CO events. For detecting overrepresented motifs, we retrieved the genomic sequences spanning CO sites with a resolution above 0.002, including the 3-kb flanking regions, and analyzed these with the MEME suite (Bailey *et al.*, 2015) using default parameters. Furthermore, we generated a list of genes present in the CO and coldspot regions and subsequently ran Panther (Mi *et al.*, 2019) to identify enriched GO Terms. We also computed the number of resistance genes (Nieri *et al.*, 2017) and historical recombination hotspots (Fuentes *et al.*, 2022) in the CO coldspots.

5.4.6 Genotyping of population data

Using bwa mem (Li, 2013), we aligned a set of resequencing data for 357 accessions compiled in Fuentes *et al.* (2022) against the SL4.0 reference genome. SNPs were detected using GATK HaplotypeCaller and were further filtered using GATK joint-genotyping and hard filtering. We then selected biallelic SNPs with a minimum allele frequency of 0.05 and less than 10% missing data using bcftools (Danecek *et al.*, 2011) and imputed missing calls using Beagle v 5.1 (Browning *et al.*, 2018). To detect SVs, we ran Delly (Rausch *et al.*, 2012) for each accession and then we genotype SV sites across all accessions. We examined an inversion event

and filtered accessions with missing calls, retaining 292 accessions. With SNPs in this inversion region, we generated and visualized the neighbor-joining tree using Mega7 (Kumar *et al.*, 2016) and Figtree (http://tree.bio.ed.ac.uk/software/figtree/), respectively.

5.4.7 ACR detection

Tomato plants were grown and cultivated in a greenhouse with a photoperiod of 16 hours light and 8 hours dark, and a minimum temperature of 16°C. Only healthy four- to seven-week-old plants were used in all experiments. The youngest leaves (the most apical) were used to isolate somatic nuclei. Meiocytes were isolated from young flower buds containing anthers that were less than 2 mm in size. Microscopic analysis revealed that at this stage in anther development nearly all meiocytes are in prophase I.

For nuclei isolation, approximately 0.4 g of young tomato leaves, or anthers from 20 prophase I flower buds were collected and immediately chopped in 2mL pre-chilled lysis buffer (15mM Tris-HCl pH7.5, 20mM NaCl, 80mM KCl, 0.5mM spermine, 5mM 2-mercaptoethanol, 0.2% Triton X-100) until a homogenous suspension was obtained. The suspensions were filtered twice through Miracloth and subsequently loaded gently on the surface of 2mL dense sucrose buffer (20mM Tris-HCl pH 8.0, 2mM MgCl2, 2mM EDTA, 25mM 2-Mercaptoethanol, 1.7M sucrose, 0.2% Triton X-100) in a 15mL Falcon tube. The nuclei were centrifuged at 2200g at 4°C for 20 minutes and the pellets were resuspended in 500µL pre-chilled lysis buffer.

Nuclei were kept on ice during the entire sorting procedure. Nuclei were first stained with 4,6-Diamidino-2-phenylindole (DAPI) and examined for integrity and purity using a Zeiss Axioskop2 microscope. Once the integrity and purity of nuclei was confirmed, nuclei were sorted in a BD FACS Aria III sorter. A total of 50,000 nuclei were sorted based on their size, shape and the intensity of the DAPI signal, which indicates the ploidy levels of the nuclei. 2n nuclei were sorted from young leaf samples, while 4n nuclei, corresponding to meiocytes, were sorted from anther samples. After sorting, nuclei were once more checked for integrity and purity under a microscope. Nuclei were transferred from sorting tubes to LoBind Eppendorf tubes and centrifuged at 1000g at 4°C for 10 min and then washed with Tris-Mg Buffer (10mM Tris-HCl pH 8.0, 5mM MgCl2).

Tn5 integration was performed as previously published (Buenrostro *et al.*, 2015) on purified nuclei using the Nextera Illumina kit (Illumina, FC 121 1031) at 37 °C for 30 min. After tagmentation (insertion of the sequencing adapter into accessible chromatin), the tagged DNA was purified with a Qiagen MinElute PCR purification kit. To generate an ATAC-seq library for sequencing, tagged fragments were amplified by two successive rounds of PCR. In the first round of PCR, the fragments were amplified by only 3 PCR cycles using the NEBNext High-Fidelity 2xPCR Master Mix and the Custom Nextera PCR Primer 1 and barcoded sets of Primer 2. Subsequently, 2.5 µL of the PCR amplified DNA was subjected to

quantitative PCR to estimate the relative amount of successfully tagged DNA fragments and to determine the optimal number of amplification cycles for the second round of PCR. The latter was estimated by plotting fluorescence values against the number of cycles. The number of cycles required for the second PCR amplification equals the number of cycles that results in 25% of the maximum fluorescent intensity (Bajic *et al.*, 2018). ATAC-seq libraries generated were purified using AMPure XP beads (Beckman Coulter) and quantified using Qubit DNA high sensitivity assay in combination with Tapestation D1000 prior to sequencing.

Sequencing was carried out using an Illumina NextSeq 500. A snakemake analysis workflow (https://github.com/KoesGroup/Snakemake_ATAC_seq) was used for the analysis of the ATAC-seq dataset with the default parameters of the configuration files. Briefly, paired-end sequencing reads were trimmed to remove the Illumina adapter sequences using Trimmomatic 0.38 (Bolger *et al.*, 2014b). Only reads with a quality score (Phred) above 30 were kept and mapped to the SL4.0 version of the tomato genome, tomato chloroplast genome and tomato mitochondrial genome using Bowtie2 (Langmead and Salzberg, 2012). Only reads mapping to a unique position in the tomato genome were used for further analysis. Reads mapping to the tomato genome were then shifted to correspond to the real Tn5 binding location using the Deeptools alignmentSieve with the parameter " –ATACshift". ATAC peaks were called using the MACS2 algorithm (Zhang *et al.*, 2008; Feng *et al.*, 2012).

Reads mapping uniquely to the transposable element annotation were counted using bedtools. Read counts were normalized by the total number of reads in the library and then grouped by the transposable element classes. Heatmaps and clustering was performed using the pheatmap package 1.0.8 (CRAN).

5.5 Acknowledgements

This project was supported by the MEICOM Marie Sklodowska-Curie Innovative Training Network (ITN), H2020-MSCA-ITN-2017 Horizon 2020 Grant agreement number 765212 and by the Netherlands Top Consortium for Knowledge and Innovation (TKI project LWV19283). We also thank ENZA Zaden Research & Development B.V., Bejo Zaden B.V., Hortigenetics Research (S.E. Asia) Limited, Syngenta Seeds B.V., and KWS Saat SE & Co KGaA for support. The parental seed material was donated by the Centre for Genetic Resource, The Netherlands.

5.6 References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D.*, et al.* 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell,* 182**,** 145-161 e23.

Alseekh, S., Scossa, F., Wen, W., Luo, J., Yan, J., Beleggia, R., Klee, H. J., Huang, S., Papa, R. & Fernie, A. R. 2021. Domestication of Crop Metabolomes: Desired and Unintended Consequences. *Trends Plant Sci,* 26**,** 650-661.

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

Andolfo, G., D'Agostino, N., Frusciante, L. & Ercolano, M. R. 2021. The Tomato Interspecific NB-LRR Gene Arsenal and Its Impact on Breeding Strategies. *Genes (Basel),* 12.

Bailey-Serres, J., Parker, J. E., Ainsworth, E. A., Oldroyd, G. E. D. & Schroeder, J. I. 2019. Genetic strategies for improving crop yields. *Nature,* 575**,** 109-118.

Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. 2015. The MEME Suite. *Nucleic Acids Res,* 43**,** W39-49.

Bajic, M., Maher, K. A. & Deal, R. B. 2018. Identification of Open Chromatin Regions in Plant Genomes Using ATAC-Seq. *Methods Mol Biol,* 1675**,** 183-201.

Bolger, A., Scossa, F., Bolger, M. E., Lanz, C., Maumus, F., Tohge, T., Quesneville, H., Alseekh, S., Sorensen, I., Lichtenstein, G.*, et al.* 2014a. The genome of the stress-tolerant wild tomato species *Solanum pennellii*. *Nat Genet,* 46**,** 1034-8.

Bolger, A. M., Lohse, M. & Usadel, B. 2014b. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics,* 30**,** 2114-20.

Browning, B. L., Zhou, Y. & Browning, S. R. 2018. A One-Penny Imputed Genome from Next-Generation Reference Panels. *Am J Hum Genet,* 103**,** 338-348.

Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Current Protocols in Molecular Biology,* 109.

Casselman, A. L., Vrebalov, J., Conner, J. A., Singhal, A., Giovannoni, J., Nasrallah, M. E. & Nasrallah, J. B. 2000. Determining the physical limits of the *Brassica* S locus by recombinational analysis. *Plant Cell,* 12**,** 23-33.

Chitwood-Brown, J., Vallad, G. E., Lee, T. G. & Hutton, S. F. 2021. Characterization and elimination of linkage-drag associated with Fusarium wilt race 3 resistance genes. *Theor Appl Genet,* 134**,** 2129-2140.

Choi, K. & Henderson, I. R. 2015. Meiotic recombination hotspots - a comparative view. *Plant J,* 83**,** 52-61.

Choi, K., Reinhard, C., Serra, H., Ziolkowski, P. A., Underwood, C. J., Zhao, X., Hardcastle, T. J., Yelina, N. E., Griffin, C., Jackson, M.*, et al.* 2016. Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes. *PLoS Genet,* 12**,** e1006179.

Choi, K., Zhao, X., Tock, A. J., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J.*, et al.* 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in *Arabidopsis thaliana* transposons and gene regulatory regions. *Genome Research,* 28**,** 532-546.

Cong, B., Barrero, L. S. & Tanksley, S. D. 2008. Regulatory change in YABBY-like transcription factor led to evolution of extreme fruit size during tomato domestication. *Nat Genet,* 40**,** 800-4.

Crow, T., Ta, J., Nojoomi, S., Aguilar-Rangel, M. R., Torres Rodriguez, J. V., Gates, D., Rellan-Alvarez, R., Sawers, R. & Runcie, D. 2020. Gene regulatory effects of a large chromosomal inversion in highland maize. *PLoS Genet,* 16**,** e1009213.

Crown, K. N., Miller, D. E., Sekelsky, J. & Hawley, R. S. 2018. Local Inversion Heterozygosity Alters Recombination throughout the Genome. *Curr Biol,* 28**,** 2984-2990 e3.

Danecek, P., Auton, A., Abecasis, G., Albers, C. a., Banks, E., DePristo, M. a., Handsaker, R. E., Lunter, G., Marth, G. T., Sherry, S. T.*, et al.* 2011. The variant call format and VCFtools. *Bioinformatics (Oxford, England),* 27**,** 2156-8.

De Oliveira, A. S., Koolhaas, I., Boiteux, L. S., Caldararu, O. F., Petrescu, A. J., Oliveira Resende, R. & Kormelink, R. 2016. Cell death triggering and effector recognition by Sw-5 SD-CNL proteins from

resistant and susceptible tomato isolines to Tomato spotted wilt virus. *Mol Plant Pathol,* 17**,** 1442-1454.

Demirci, S., van Dijk, A. J., Sanchez Perez, G., Aflitos, S. A., de Ridder, D. & 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 Journal,* 89**,** 554-564.

Dominguez, M., Dugas, E., Benchouaia, M., Leduque, B., Jimenez-Gomez, J. M., Colot, V. & Quadrana, L. 2020. The impact of transposable elements on tomato diversity. *Nat Commun,* 11**,** 4058.

Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. 2012. Identifying ChIP-seq enrichment using MACS. *Nat Protoc,* 7**,** 1728-40.

Fransz, P., Linc, G., Lee, C. R., Aflitos, S. A., Lasky, J. R., Toomajian, C., Ali, H., Peters, J., van Dam, P., Ji, X.*, et al.* 2016. Molecular, genetic and evolutionary analysis of a paracentric inversion in *Arabidopsis thaliana*. *Plant J,* 88**,** 159-178.

Fuentes, R., Hesselink, T., Nieuwenhuis, R., Bakker, L., Schijlen, E., van Dooijeweert, W., Diaz Trivino, S., de Haan, J. R., Sanchez Perez, G., Zhang, X.*, et al.* 2020. Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. *Plant J,* 102**,** 480-492.

Fuentes, R. R., de Ridder, D., van Dijk, A. D. J. & Peters, S. A. 2022. Domestication Shapes Recombination Patterns in Tomato. *Mol Biol Evol,* 39.

Goel, M., Sun, H., Jiao, W. B. & Schneeberger, K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol,* 20**,** 277.

He, Y., Wang, M., Dukowic-Schulze, S., Zhou, A., Tiang, C. L., Shilo, S., Sidhu, G. K., Eichten, S., Bradbury, P., Springer, N. M.*, et al.* 2017. Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proc Natl Acad Sci U S A,* 114**,** 12231-12236.

Herget, T., Schell, J. & Schreier, P. H. 1990. Elicitor-specific induction of one member of the chitinase gene family in *Arachis hypogaea*. *Mol Gen Genet,* 224**,** 469-76.

Hosmani, P. S., Flores-Gonzalez, M., van de Geest, H., Maumus, F., Bakker, L. V., Schijlen, E., van Haarst, J., Cordewener, J., Sanchez-Perez, G., Peters, S.*, et al.* 2019. An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *bioRxiv*.

Huang, Z. & van der Knaap, E. 2011. Tomato fruit weight 11.3 maps close to fasciated on the bottom of chromosome 11. *Theor Appl Genet,* 123**,** 465-74.

Huang, Z., Van Houten, J., Gonzalez, G., Xiao, H. & van der Knaap, E. 2013. Genome-wide identification, phylogeny and expression analysis of SUN, OFP and YABBY gene family in tomato. *Mol Genet Genomics,* 288**,** 111-29.

Hulbert, S. H., Webb, C. A., Smith, S. M. & Sun, Q. 2001. Resistance gene complexes: evolution and utilization. *Annu Rev Phytopathol,* 39**,** 285-312.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I. & Paszkowski, J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature,* 472**,** 115-9.

Jiao, W. B. & Schneeberger, K. 2020. Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nat Commun,* 11**,** 989.

Kent, T. V., Uzunovic, J. & Wright, S. I. 2017. Coevolution between transposable elements and recombination. *Philos Trans R Soc Lond B Biol Sci,* 372.

Kirkpatrick, M. 2010. How and why chromosome inversions evolve. *PLoS Biol,* 8.

Kirkpatrick, M. & Barton, N. 2006. Chromosome inversions, local adaptation and speciation. *Genetics,* 173**,** 419-34.

Koenig, D., Jimenez-Gomez, J. M., Kimura, S., Fulop, D., Chitwood, D. H., Headland, L. R., Kumar, R., Covington, M. F., Devisetty, U. K., Tat, A. V.*, et al.* 2013. Comparative transcriptomics reveals patterns of selection in domesticated and wild tomato. *Proc Natl Acad Sci U S A,* 110**,** E2655-62.

Korunes, K. L. & Noor, M. A. F. 2019. Pervasive gene conversion in chromosomal inversion heterozygotes. *Mol Ecol,* 28**,** 1302-1315.

Krumsiek, J., Arnold, R. & Rattei, T. 2007. Gepard: a rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics,* 23**,** 1026-8.

Kuang, H., Padmanabhan, C., Li, F., Kamei, A., Bhaskar, P. B., Ouyang, S., Jiang, J., Buell, C. R. & 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.

Kumar, S., Stecher, G. & Tamura, K. 2016. MEGA7: Molecular Evolutionary Genetics Analysis Version 7.0 for Bigger Datasets. *Mol Biol Evol,* 33**,** 1870-4.

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

Lanfermeijer, F. C., Warmink, J. & Hille, J. 2005. The products of the broken Tm-2 and the durable Tm-2(2) resistance genes from tomato differ in four amino acids. *J Exp Bot,* 56**,** 2925-33.

Langmead, B. & Salzberg, S. L. 2012. Fast gapped-read alignment with Bowtie 2. *Nat Methods,* 9**,** 357-9.

Li, H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. 00**,** 3-3.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X.*, et al.* 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet,* 46**,** 1220-6.

Lowry, D. B. & Willis, J. H. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol,* 8.

Lye, Z. N. & Purugganan, M. D. 2019. Copy Number Variation in Domestication. *Trends Plant Sci,* 24**,** 352-365.

Marand, A. P., Jansky, S. H., Zhao, H., Leisner, C. P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A. J., Veilleux, R. E.*, 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. & Jiang, J. 2019. Historical Meiotic Crossover Hotspots Fueled Patterns of Evolutionary Divergence in Rice. *Plant Cell,* 31**,** 645-662.

Marks, P., Garcia, S., Barrio, A. M., Belhocine, K., Bernate, J., Bharadwaj, R., Bjornson, K., Catalanotti, C., Delaney, J., Fehr, A.*, et al.* 2019. Resolving the full spectrum of human genome variation using Linked-Reads. *Genome Res,* 29**,** 635-645.

Mercier, R., Mezard, C., Jenczewski, E., Macaisne, N. & Grelon, M. 2015. The molecular biology of meiosis in plants. *Annu Rev Plant Biol,* 66**,** 297-327.

Mi, H., Muruganujan, A., Ebert, D., Huang, X. & Thomas, P. D. 2019. PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools. *Nucleic Acids Res,* 47**,** D419-D426.

Mieulet, D., Aubert, G., Bres, C., Klein, A., Droc, G., Vieille, E., Rond-Coissieux, C., Sanchez, M., Dalmais, M., Mauxion, J. P.*, et al.* 2018. Unleashing meiotic crossovers in crops. *Nat Plants,* 4**,** 1010-1016.

Moyers, B. T., Morrell, P. L. & McKay, J. K. 2018. Genetic Costs of Domestication and Improvement. *J Hered,* 109**,** 103-116.

Moyle, L. C. 2008. Ecological and evolutionary genomics in the wild tomatoes (*Solanum sect. Lycopersicon*). *Evolution,* 62**,** 2995-3013.

Nieri, D., Di Donato, A. & Ercolano, M. R. 2017. Analysis of tomato meiotic recombination profile reveals preferential chromosome positions for NB-LRR genes. *Euphytica,* 213.

Ortiz-Barrientos, D., Engelstadter, J. & Rieseberg, L. H. 2016. Recombination Rate Evolution and the Origin of Species. *Trends Ecol Evol,* 31**,** 226-236.

Pan, Q., Deng, M., Yan, J. & Li, L. 2017. Complexity of genetic mechanisms conferring nonuniformity of recombination in maize. *Sci Rep,* 7**,** 1205.

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D.*, et al.* 2017. Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*.

Qiu, Z., Li, R., Zhang, S., Wang, K., Xu, M., Li, J., Du, Y., Yu, H. & Cui, X. 2016. Identification of Regulatory DNA Elements Using Genome-wide Mapping of DNase I Hypersensitive Sites during Tomato Fruit Development. *Mol Plant,* 9**,** 1168-1182.

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

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

Rodriguez, G. R., Munos, S., Anderson, C., Sim, S. C., Michel, A., Causse, M., Gardener, B. B., Francis, D. & van der Knaap, E. 2011. Distribution of SUN, OVATE, LC, and FAS in the tomato germplasm and the relationship to fruit shape diversity. *Plant Physiol,* 156**,** 275-85.

Rowan, B. A., Heavens, D., Feuerborn, T. R., Tock, A. J., Henderson, I. R. & Weigel, D. 2019. An Ultra High-Density *Arabidopsis thaliana* Crossover Map That Refines the Influences of Structural Variation and Epigenetic Features. *Genetics,* 213**,** 771-787.

Sauvage, C., Rau, A., Aichholz, C., Chadoeuf, J., Sarah, G., Ruiz, M., Santoni, S., Causse, M., David, J. & Glemin, S. 2017. Domestication rewired gene expression and nucleotide diversity patterns in tomato. *Plant J,* 91**,** 631-645.

Schouten, H. J., Tikunov, Y., Verkerke, W., Finkers, R., Bovy, A., Bai, Y. & Visser, R. G. F. 2019. Breeding Has Increased the Diversity of Cultivated Tomato in The Netherlands. *Front Plant Sci,* 10**,** 1606.

Schwander, T., Libbrecht, R. & Keller, L. 2014. Supergenes and complex phenotypes. *Curr Biol,* 24**,** R288-94.

Schwarzkopf, E. J., Motamayor, J. C. & Cornejo, O. E. 2020. Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. *BMC Genomics,* 21**,** 332.

Seah, S., Yaghoobi, J., Rossi, M., Gleason, C. A. & Williamson, V. M. 2004. The nematode-resistance gene, Mi-1, is associated with an inverted chromosomal segment in susceptible compared to resistant tomato. *Theor Appl Genet,* 108**,** 1635-42.

Seong, K., Seo, E., Witek, K., Li, M. & Staskawicz, B. 2020. Evolution of NLR resistance genes with noncanonical N-terminal domains in wild tomato species. *New Phytol,* 227**,** 1530-1543.

Serra, H., Lambing, C., Griffin, C. H., Topp, S. D., Nageswaran, D. C., Underwood, C. J., Ziolkowski, P. A., Seguela-Arnaud, M., Fernandes, J. B., Mercier, R.*, et al.* 2018. Massive crossover elevation via combination of HEI10 and recq4a recq4b during Arabidopsis meiosis. *Proc Natl Acad Sci U S A,* 115**,** 2437-2442.

Sharma, N., Sahu, P. P., Prasad, A., Muthamilarasan, M., Waseem, M., Khan, Y., Thakur, J. K., Chakraborty, S. & Prasad, M. 2021. The Sw5a gene confers resistance to ToLCNDV and triggers an HR response after direct AC4 effector recognition. *Proc Natl Acad Sci U S A,* 118.

Shen, C., Wang, N., Huang, C., Wang, M., Zhang, X. & Lin, Z. 2019. Population genomics reveals a fine-scale recombination landscape for genetic improvement of cotton. *Plant J,* 99**,** 494-505.

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

Shrestha, C. L., Oña, I., Muthukrishnan, S. & Mew, T. W. 2007. Chitinase levels in rice cultivars correlate with resistance to the sheath blight pathogen *Rhizoctonia solani*. *European Journal of Plant Pathology,* 120**,** 69-77.

Sim, S. C., Durstewitz, G., Plieske, J., Wieseke, R., Ganal, M. W., Van Deynze, A., Hamilton, J. P., Buell, C. R., Causse, M., Wijeratne, S.*, et al.* 2012. Development of a large SNP genotyping array and generation of high-density genetic maps in tomato. *PLoS One,* 7**,** e40563.

Sun, Y., Ji, K., Liang, B., Du, Y., Jiang, L., Wang, J., Kai, W., Zhang, Y., Zhai, X., Chen, P.*, et al.* 2017. Suppressing ABA uridine diphosphate glucosyltransferase (SlUGT75C1) alters fruit ripening and the stress response in tomato. *Plant J,* 91**,** 574-589.

Szinay, D., Wijnker, E., van den Berg, R., Visser, R. G. F., de Jong, H. & Bai, Y. 2012. Chromosome evolution in *Solanum* traced by cross-species BAC-FISH. *New Phytol,* 195**,** 688-698.

Termolino, P., Cremona, G., Consiglio, M. F. & Conicella, C. 2016. Insights into epigenetic landscape of recombination-free regions. *Chromosoma,* 125**,** 301-8.

Thieme, M., Lanciano, S., Balzergue, S., Daccord, N., Mirouze, M. & Bucher, E. 2017. Inhibition of RNA polymerase II allows controlled mobilisation of retrotransposons for plant breeding. *Genome Biol,* 18**,** 134.

Thompson, M. J. & Jiggins, C. D. 2014. Supergenes and their role in evolution. *Heredity (Edinb),* 113**,** 1-8.

Tieman, D., Zhu, G., Resende, M. F., Jr., Lin, T., Nguyen, C., Bies, D., Rambla, J. L., Beltran, K. S., Taylor, M., Zhang, B.*, et al.* 2017. A chemical genetic roadmap to improved tomato flavor. *Science,* 355**,** 391-394.

Tock, A. J. & Henderson, I. R. 2018. Hotspots for Initiation of Meiotic Recombination. *Front Genet,* 9**,** 521.

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

van der Knaap, E., Sanyal, A., Jackson, S. A. & Tanksley, S. D. 2004. High-resolution fine mapping and fluorescence in situ hybridization analysis of sun, a locus controlling tomato fruit shape, reveals a region of the tomato genome prone to DNA rearrangements. *Genetics,* 168**,** 2127-40.

van Rengs, W. M. J., Schmidt, M. H., Effgen, S., Le, D. B., Wang, Y., Zaidan, M., Huettel, B., Schouten, H. J., Usadel, B. & Underwood, C. J. 2022. A chromosome scale tomato genome built from complementary PacBio and Nanopore sequences alone reveals extensive linkage drag during breeding. *Plant J,* 110**,** 572-588.

Wang, X., Gao, L., Jiao, C., Stravoravdis, S., Hosmani, P. S., Saha, S., Zhang, J., Mainiero, S., Strickler, S. R., Catala, C.*, et al.* 2020. Genome of *Solanum pimpinellifolium* provides insights into structural variants during tomato breeding. *Nat Commun,* 11**,** 5817.

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

Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L.*, et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife,* 2**,** e01426.

Willing, E. M., Rawat, V., Mandakova, T., Maumus, F., James, G. V., Nordstrom, K. J., Becker, C., Warthmann, N., Chica, C., Szarzynska, B.*, et al.* 2015. Genome expansion of Arabis alpina linked with retrotransposition and reduced symmetric DNA methylation. *Nat Plants,* 1**,** 14023.

Wolters, A. M., Caro, M., Dong, S., Finkers, R., Gao, J., Visser, R. G., Wang, X., Du, Y. & Bai, Y. 2015. Detection of an inversion in the Ty-2 region between *S. lycopersicum* and *S. habrochaites* by a combination of de novo genome assembly and BAC cloning. *Theor Appl Genet,* 128**,** 1987-97.

Xu, C., Liberatore, K. L., MacAlister, C. A., Huang, Z., Chu, Y. H., Jiang, K., Brooks, C., Ogawa-Ohnishi, M., Xiong, G., Pauly, M.*, et al.* 2015. A cascade of arabinosyltransferases controls shoot meristem size in tomato. *Nat Genet,* 47**,** 784-92.

Yang, X., Caro, M., Hutton, S. F., Scott, J. W., Guo, Y., Wang, X., Rashid, M. H., Szinay, D., de Jong, H., Visser, R. G.*, et al.* 2014. Fine mapping of the tomato yellow leaf curl virus resistance gene Ty-2 on chromosome 11 of tomato. *Mol Breed,* 34**,** 749-760.

Yelina, N., Diaz, P., Lambing, C. & Henderson, I. R. 2015. Epigenetic control of meiotic recombination in plants. *Sci China Life Sci,* 58**,** 223-31.

Zervudacki, J., Yu, A., Amesefe, D., Wang, J., Drouaud, J., Navarro, L. & Deleris, A. 2018. Transcriptional control and exploitation of an immune-responsive family of plant retrotransposons. *EMBO J,* 37.

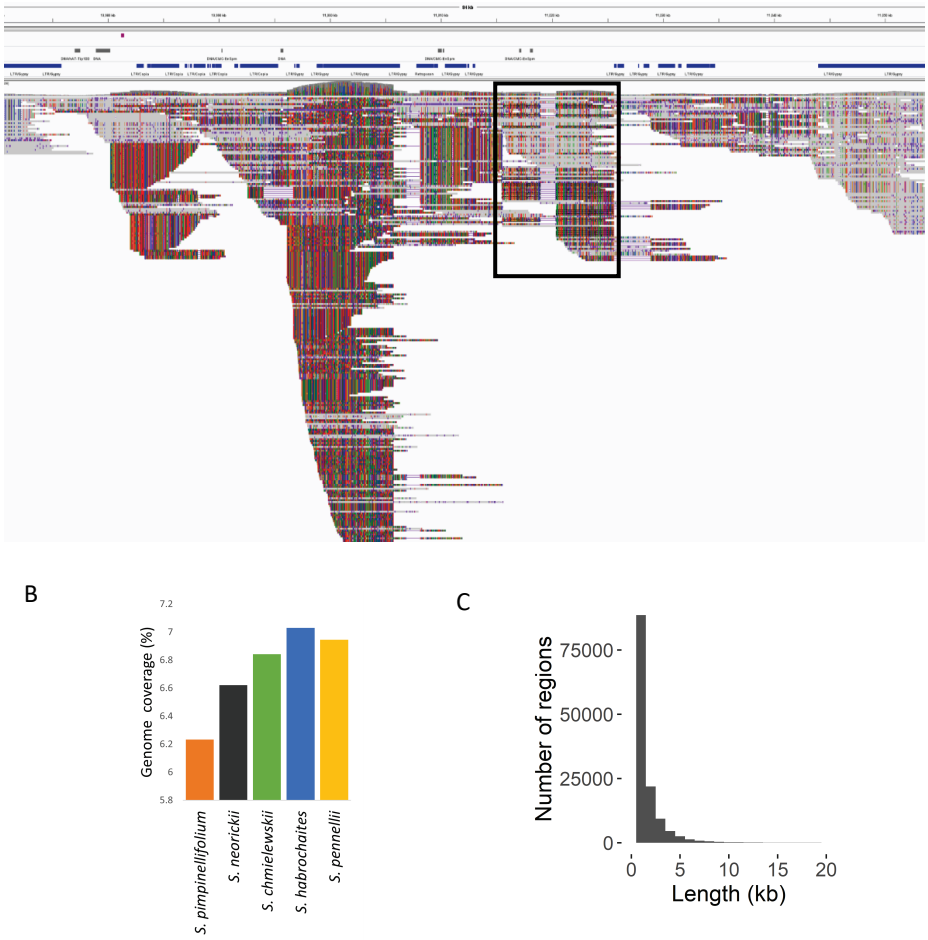
Zhang, Y., Liu, T., Meyer, C. A., Eeckhoute, J., Johnson, D. S., Bernstein, B. E., Nusbaum, C., Myers, R. M., Brown, M., Li, W.*, et al.* 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* 9**,** R137.

Zhu, G., Wang, S., Huang, Z., Zhang, S., Liao, Q., Zhang, C., Lin, T., Qin, M., Peng, M., Yang, C.*, et al.* 2018. Rewiring of the Fruit Metabolome in Tomato Breeding. *Cell,* 172**,** 249-261 e12.

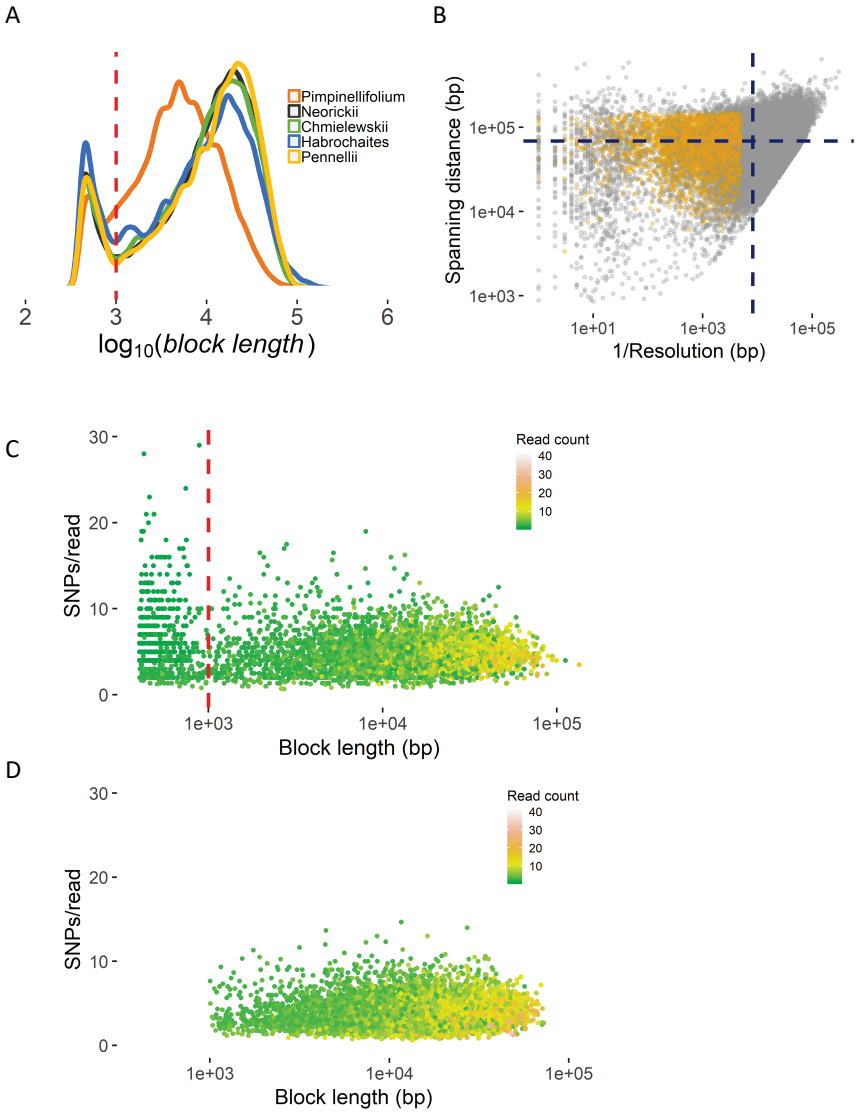
Zhu, Q., Maher, E. A., Masoud, S., Dixon, R. A. & Lamb, C. J. 1994. Enhanced Protection Against Fungal Attack by Constitutive Co–expression of Chitinase and Glucanase Genes in Transgenic Tobacco. *Nature Biotechnology,* 12**,** 807-812.

Ziolkowski, P. A., Underwood, C. J., Lambing, C., Martinez-Garcia, M., Lawrence, E. J., Ziolkowska, L., Griffin, C., Choi, K., Franklin, F. C., Martienssen, R. A.*, et al.* 2017. Natural variation and dosage of the HEI10 meiotic E3 ligase control Arabidopsis crossover recombination. *Genes Dev,* 31**,** 306-317.

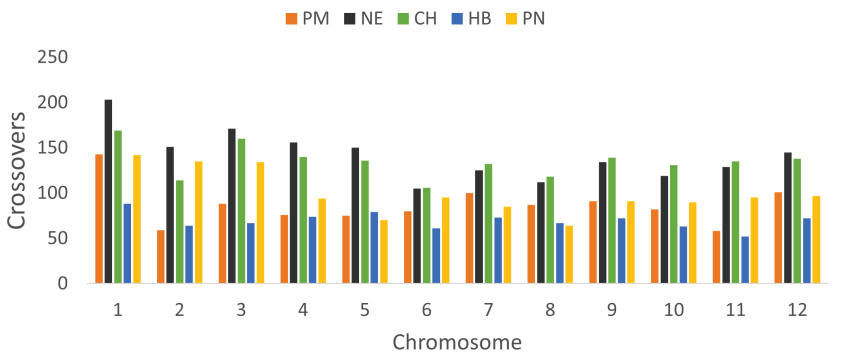
5.7 Supplementary figures



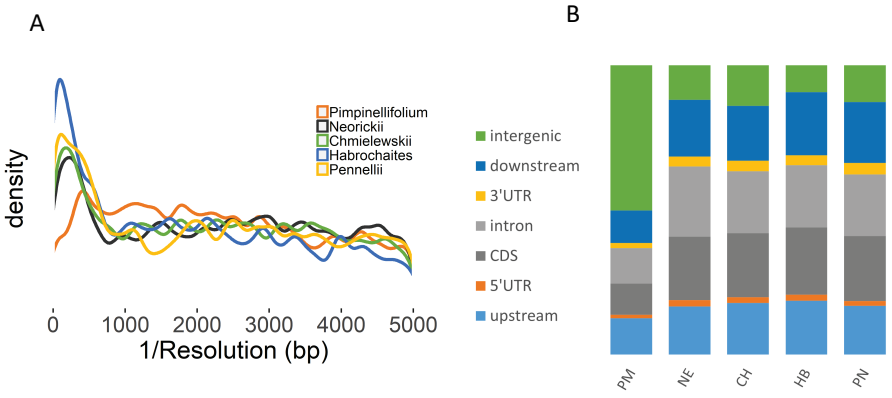
Supplementary Figure 5.1. False positive hotspots in pericentric heterochromatin. A) Regions with excessive levels of heterozygosity and read coverage causing false positive crossovers (black box). Possibly, these regions are collapsed genomic segments in the reference genome or part of a copy number variation. B) The coverage and C) length distribution of these regions.



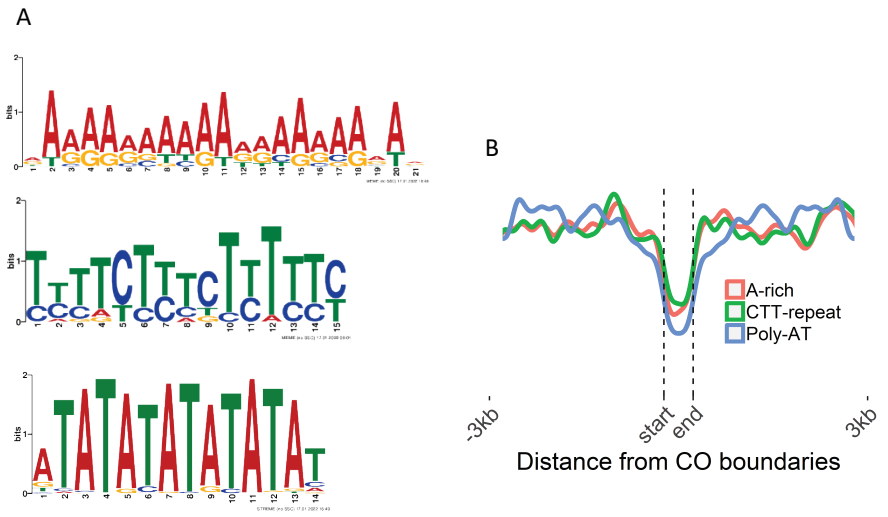
Supplementary Figure 5.2. Filtering criteria on recombination molecules. A) A lower limit on the size of a haplotype block (marked by a red vertical line) is set based on the distribution of all block sizes across all crosses. B) Distribution of resolution and spanning distance of each recombinant molecule. The yellow dots represent COs that passed the filtering. The blue lines mark the average spanning distance and resolution. C) Ratio of SNP and read count per haplotype block as a function of block length. Haplotype blocks with sizes below 1kb are supported by few reads with high SNP density which may result from mismapped reads, further supporting the cut-off for haplotype block size. D) Final set of COs that passed all filtering constraints.



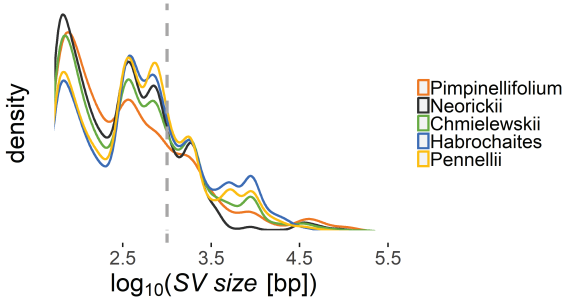
Supplementary Figure 5.3. Frequency of crossovers per chromosome.



Supplementary Figure 5.4. Crossover resolution and gene overlap. A) *S. pimpinellifolium* crossovers have lower resolution compared to the other groups, but between the resolution of 0.0002 to 0.001 (1kb to 5kb), the distributions are similar across the different crosses. B) Overlap of crossovers (resolution between 0.0002 to 0.001) with gene and intergenic regions.

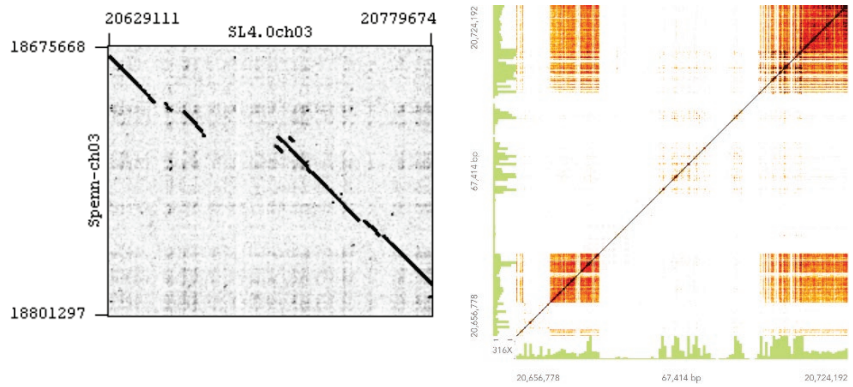


Supplementary Figure 5.5. Overrepresented motifs. A) Motifs found within and flanking CO sites. Of 1,267 COs with resolution of at least 0.002, only 8-28% have the motif within the CO sites while the rest contain multiple copies of the motifs in the flanking regions. B) Distribution of the motifs within and around high-resolution COs.

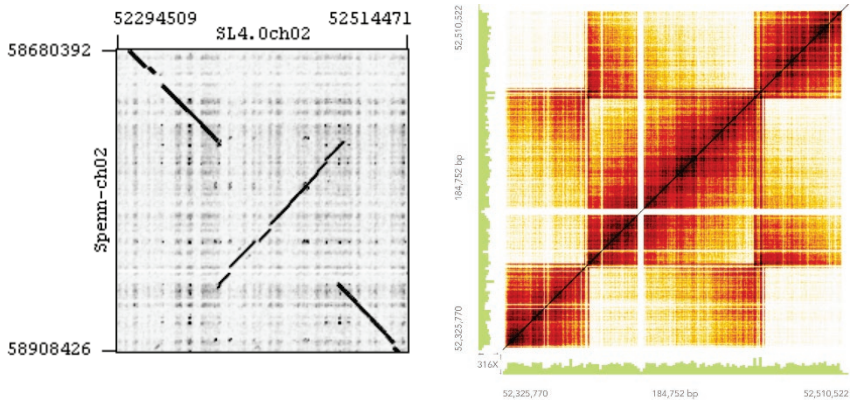


Supplementary Figure 5.6. Longer structural variants for more distant wild genomes. Distribution of SV sizes per population showing higher frequency of longer SVs for *S. habrochaites* and *S. pennellii*.

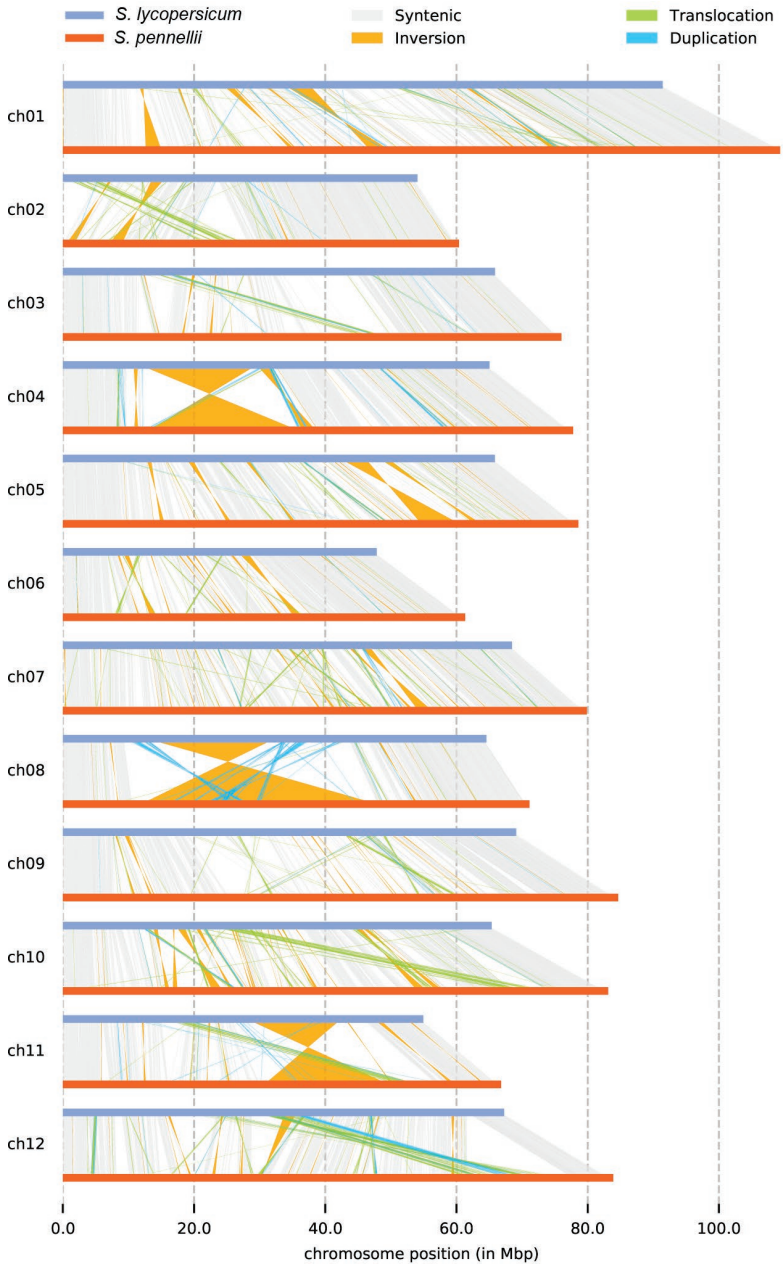
A



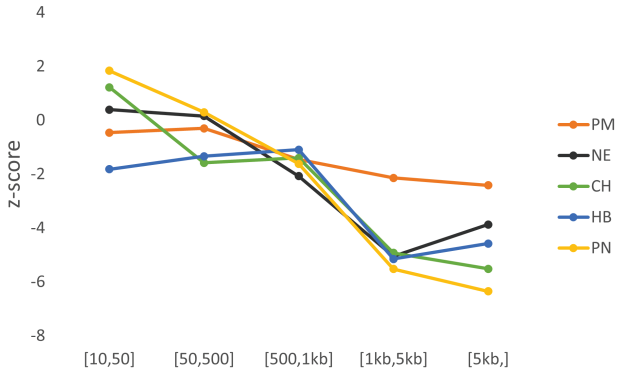
B



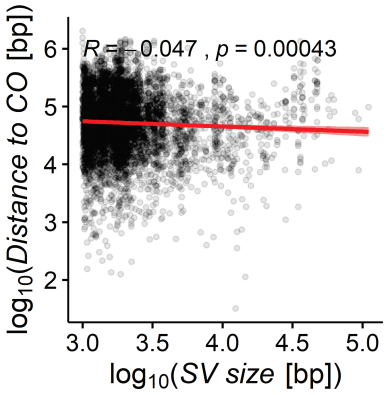
Supplementary Figure 5.7. Validation of structural variants. Examples of a (A) deletion and an (B) inversion that are validated by manual inspection. First, through dot plots between the assemblies of *S. lycopersicum c.v.* Heinz 1706 and *S. pennellii* genomes generated using *Gepard* (left). Second, through heatmap of overlapping barcodes between linked reads (10X Genomics) in the *S. pennellii* parental genome generated using *Loupe Browser*. The patterns in the top right and bottom right figures characterize a deletion and an inversion, respectively.



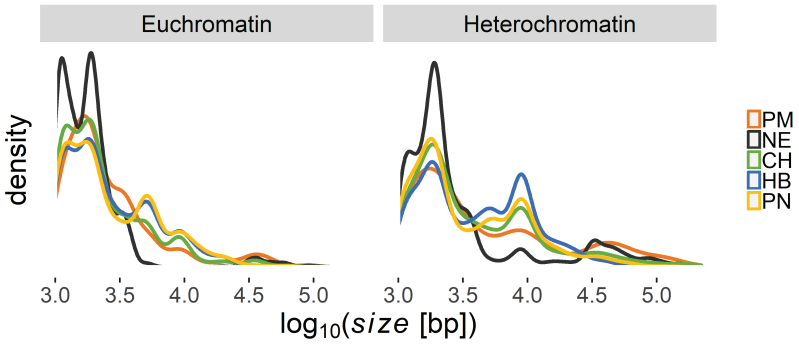
Supplementary Figure 5.8. Parental genome alignment. Alignment between the assemblies of *S. lycopersicum* and *S. pennellii* showing syntenic regions and rearrangements.



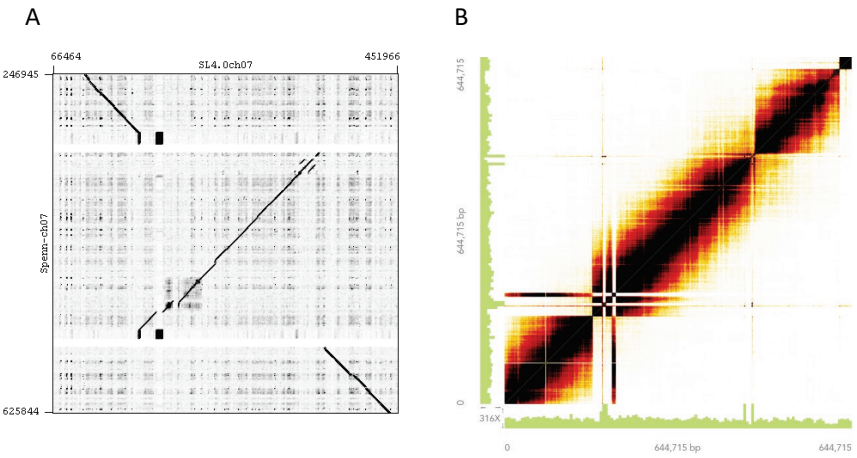
Supplementary Figure 5.9. Suppression of COs relative to the size of SVs. To determine the relation between SV size and CO suppression, SVs are first binned according to size. Afterwards, per bin, the overlap of SVs and COs in the observed data was compared against the overlap in 10,000 permutation sets. Only bin [1kb,5kb] and [5kb,] have significantly fewer COs in SVs than expected by chance for all populations.



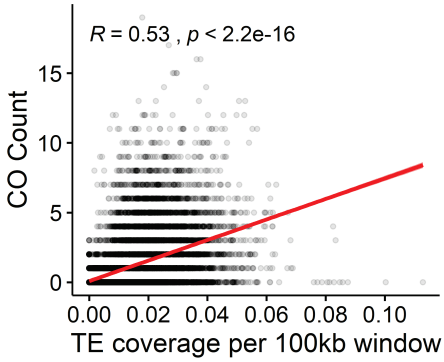
Supplementary Figure 5.10. Distance of SVs to COs by size. There is no association between SV size and the distance of the nearest CO.



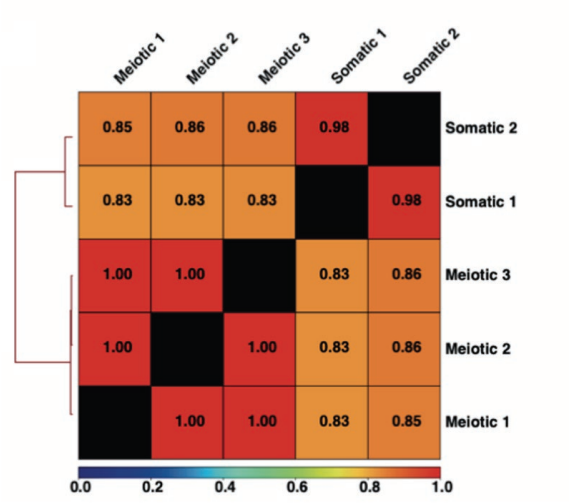
Supplementary Figure 5.11. Sizes of SVs. Pericentric heterochromatin contains longer SVs in most wild genomes.



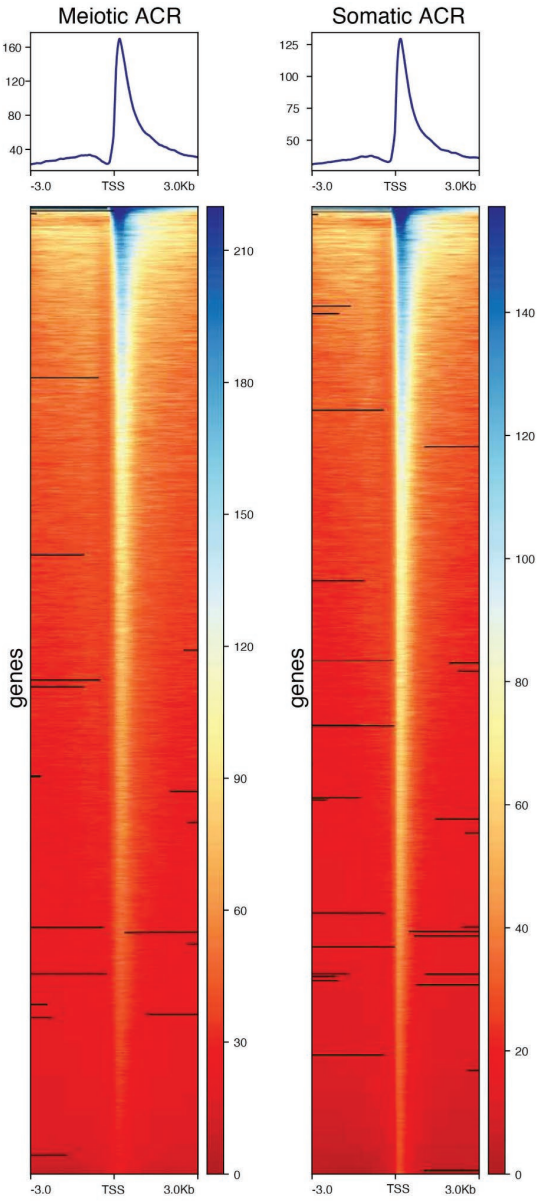
Supplementary Figure 5.12. Inversion in chromosome 7 short arm. A) A distal inversion between the short arm of *S. lycopersicum c.v.* Heinz 1706 and the *S.pennellii* assembly was visualized using a dot plot. B) Heatmap of overlapping barcodes between linked reads (10X Genomics) in the inversion region of *S. pennellii*. The figure was generated using *Loupe Browser*.



Supplementary Figure 5.13. TE and CO correlation. Spearman's rank correlation of crossover count and DNA transposons (*Stowaway* and *Tip100*) coverage in a sliding genome window. Each dot indicates a window.

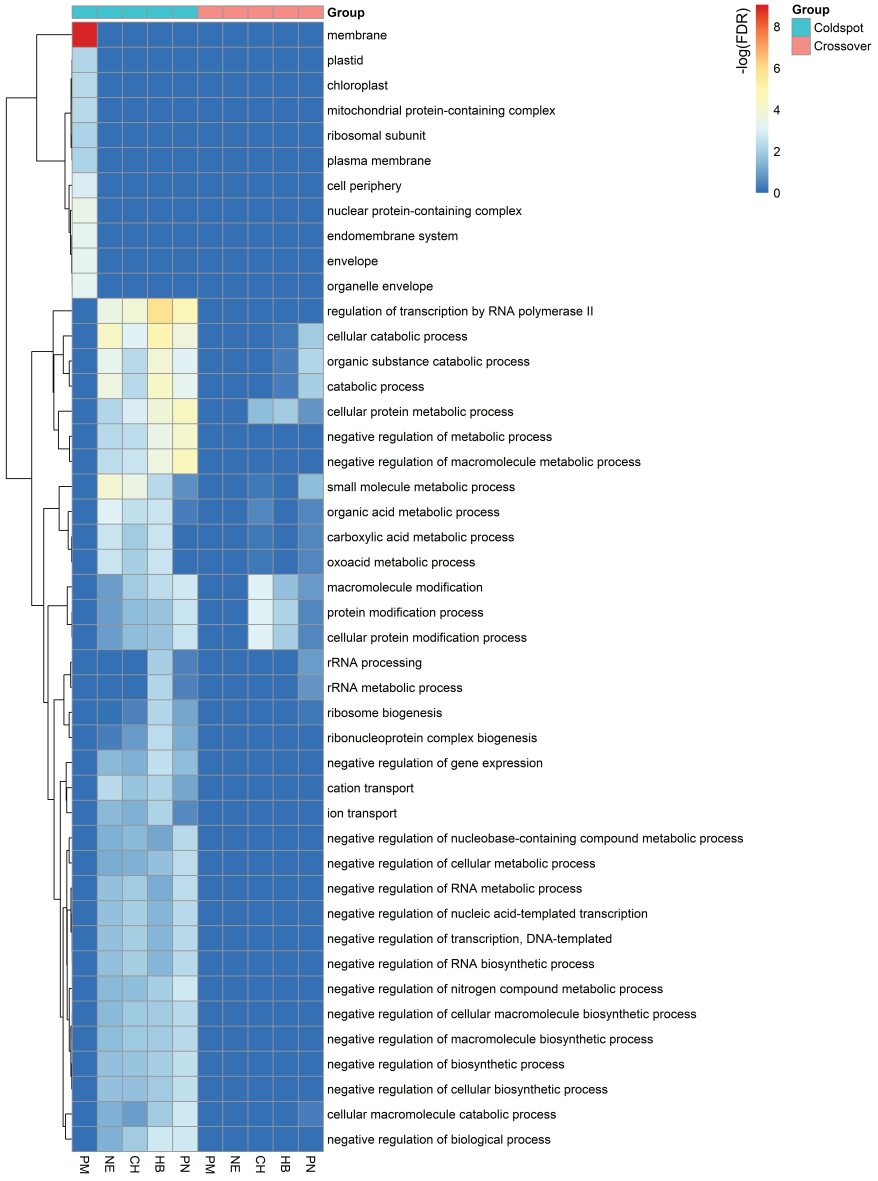


Supplementary Figure 5.14. Pearson correlation of ACRs. Comparison of read distribution over the genome between tissues and between biological replicates.

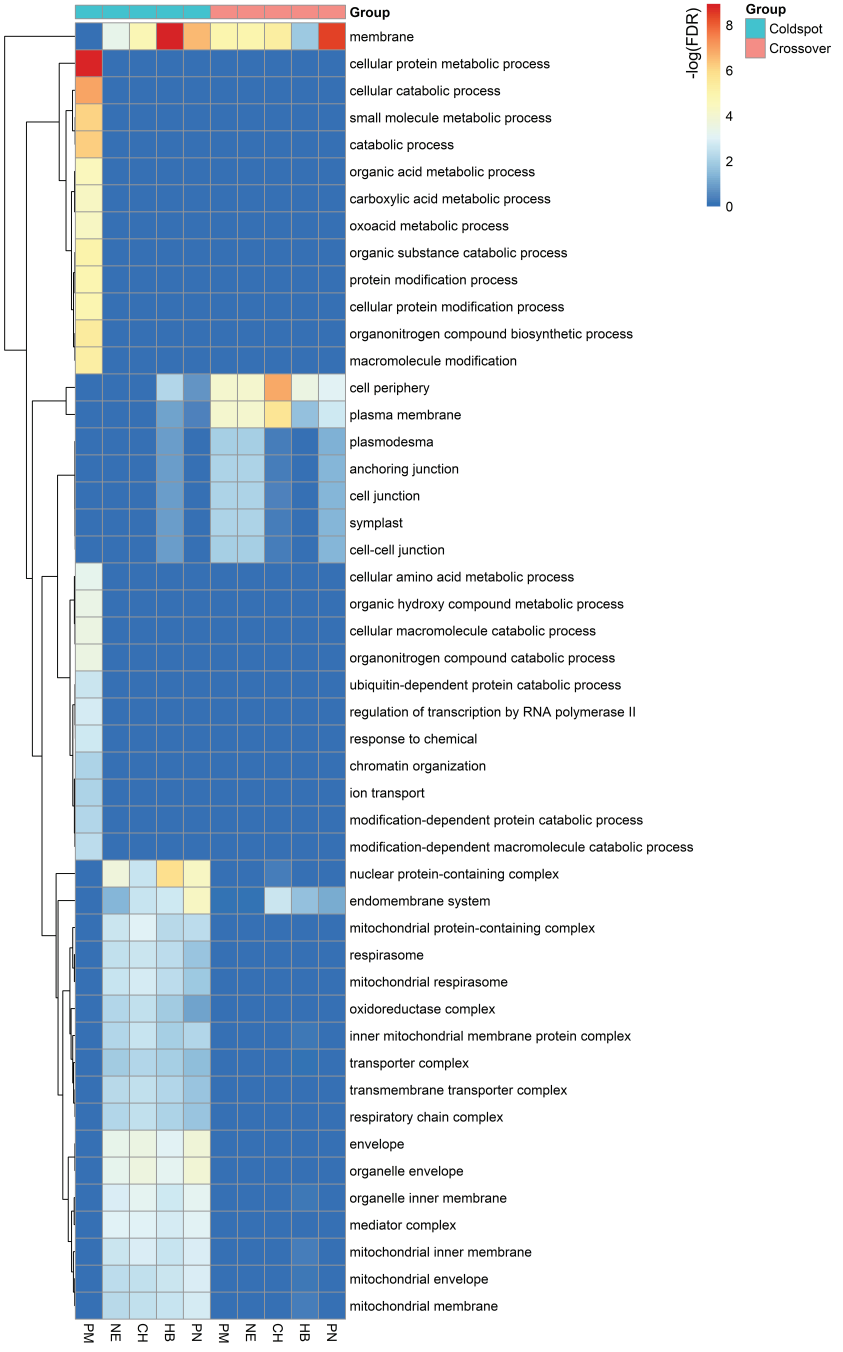


Supplementary Figure 5.15. ATAC-seq peaks at transcription start sites (TSS). Read density across ATAC-seq peaks at TSS and their 3kb flanking regions. Each row of the heatmap represents one gene. Coverage is normalized in RPKM.

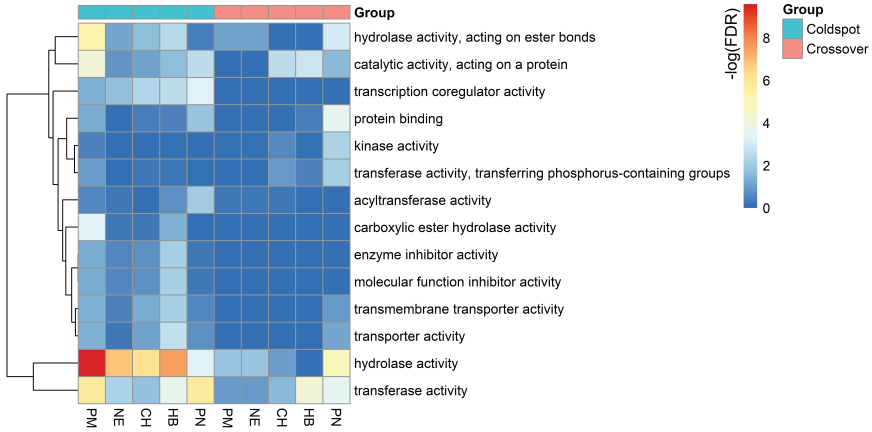
A


Group
membrane
plastid
chloroplast
mitochondrial protein-containing complex
ribosomal subunit
plasma membrane
cell periphery
nuclear protein-containing complex
endomembrane system
envelope
organelle envelope
regulation of transcription by RNA polymerase II
cellular catabolic process
organic substance catabolic process
catabolic process
cellular protein metabolic process
negative regulation of metabolic process
negative regulation of macromolecule metabolic process
small molecule metabolic process
organic acid metabolic process
carboxylic acid metabolic process
oxoacid metabolic process
macromolecule modification
protein modification process
cellular protein modification process
rRNA processing
rRNA metabolic process
ribosome biogenesis
ribonucleoprotein complex biogenesis
negative regulation of gene expression
cation transport
ion transport
negative regulation of nucleobase-containing compound metabolic process
negative regulation of cellular metabolic process
negative regulation of RNA metabolic process
negative regulation of nucleic acid-templated transcription
negative regulation of transcription, DNA-templated
negative regulation of RNA biosynthetic process
negative regulation of nitrogen compound metabolic process
negative regulation of cellular macromolecule biosynthetic process
negative regulation of macromolecule biosynthetic process
negative regulation of biosynthetic process
negative regulation of cellular biosynthetic process
cellular macromolecule catabolic process
negative regulation of biological process
PM
NE
CH
HB
PN
PM
NE
CH
HB
PN
-log(FDR)
8
6
4
2
0
Group
Coldspot
Crossover

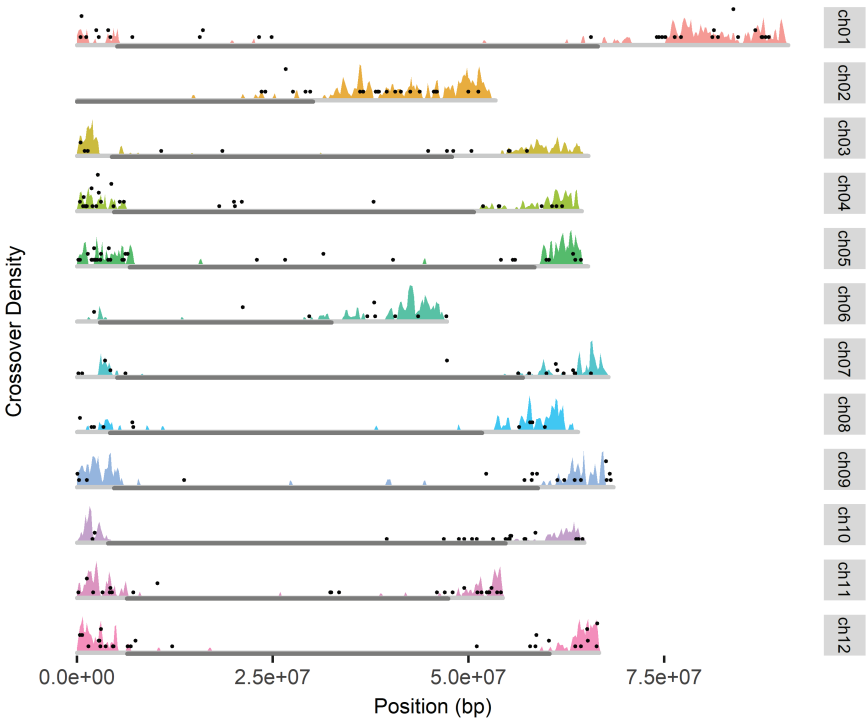
B


Group
membrane
cellular protein metabolic process
cellular catabolic process
small molecule metabolic process
catabolic process
organic acid metabolic process
carboxylic acid metabolic process
oxoacid metabolic process
organic substance catabolic process
protein modification process
cellular protein modification process
organonitrogen compound biosynthetic process
macromolecule modification
cell periphery
plasma membrane
plasmodesma
anchoring junction
cell junction
symplast
cell-cell junction
cellular amino acid metabolic process
organic hydroxy compound metabolic process
cellular macromolecule catabolic process
organonitrogen compound catabolic process
ubiquitin-dependent protein catabolic process
regulation of transcription by RNA polymerase II
response to chemical
chromatin organization
ion transport
modification-dependent protein catabolic process
modification-dependent macromolecule catabolic process
nuclear protein-containing complex
endomembrane system
mitochondrial protein-containing complex
respirasome
mitochondrial respirasome
oxidoreductase complex
inner mitochondrial membrane protein complex
transporter complex
transmembrane transporter complex
respiratory chain complex
envelope
organelle envelope
organelle inner membrane
mediator complex
mitochondrial inner membrane
mitochondrial envelope
mitochondrial membrane
PM
NE
CH
HB
PN
PM
NE
CH
HB
PN
-log(FDR)
8
6
4
2
0
Group
Coldspot
Crossover

C



Supplementary Figure 5.16. Functional enrichment in coldspot and CO regions. The minimum fold enrichment is 1.5. We separately reported the overrepresented terms by category: biological process (A), cellular location (B) and molecular function (C).



Supplementary Figure 5.17. Resistance genes across the tomato genome. The black dots representing the frequency of R genes is plotted with the recombination landscape of the *S. lycopersicum* x *S. pennellii* hybrid.

Chapter 6

Discussion

Meiotic recombination is integral to most sexually reproducing organisms, and is a major provider of genetic diversity. Hence, the study of recombination has been an important subject not only to understand evolution but also to aid breeders in developing better crops. There are, however, difficulties in profiling recombination and producing high-resolution crossover (CO) data. In this thesis, I developed new methods to detect recombination and studied COs with high resolution in newly generated tomato crosses as well as historical recombination in the context of tomato domestication. Below, I will discuss potential improvements and remaining challenges in the detection and study of meiotic recombination. I will also discuss future directions to further understand recombination and expound on their relevance and application to breeding.

6.1 Improvements and challenges

6.1.1 Crossover detection

Understanding recombination events and their distribution in the genome is crucial to improving and accelerating breeding, but the mere detection of COs has been labor intensive and costly, limiting it to major crops and in small numbers of crosses. In Chapter 2, I presented a method that relies on the sequencing of pooled pollen from an F1 hybrid to detect crossovers. It greatly reduces the cost and labor compared to the existing method, which relies on the generation and sequencing of many offspring plants. The significantly higher CO resolution compared to older methods enables us to perform analyses that link CO with genome features like genes, transposable elements, sequence motifs and structural variations (SVs). Our method also allows the comparison of local differences in recombination patterns between different crosses (Chapter 5), enabling the study of recombination barriers in interspecific hybrids. Similar methods of phasing linked reads to detect COs were developed and tested on other species like Arabidopsis, mouse and stickleback fish (Sun *et al.*, 2019; Dreau *et al.*, 2019). This demonstrated that CO profiling by sequencing a pool of gametes may be applied to a wide range of species, further enriching the understanding of recombination through cost-effective construction of recombination maps for non-model organisms.

The method I developed to detect COs relies on the availability of markers between the parental genomes. Marker density affects the resolution of CO, resulting in larger CO regions when the parental genomes are genetically closer to each other. This has been demonstrated in the comparison of resolution between the different hybrid crosses. The higher genetic divergence between *S. lycopersicum* and *S. pennellii* provided more markers than the cross between *S. lycopersicum* and its wild progenitor *S. pimpinellifolium*. The dependence on segregating markers makes this method less sensitive in detecting CO in crosses of genetically similar individuals. Currently, the method only uses homozygous SNPs in the parental genome as markers; including heterozygous SNPs and indels may help to improve sensitivity and resolution. The use of heterozygous SNPs requires accurate

assignment of parental origin if both parental genomes share a common allele. Genome assemblies must be haplotype-resolved to allow phasing of the heterozygous SNPs per parental genome and the parental assignment of haplotype instead of individual SNPs. Furthermore, it is still challenging to accurately detect indels, especially in low complexity regions. Resolving the detection issues of smaller mutations may increase the number of informative markers for phasing reads. Aside from the number of markers, high-quality SNP detection is also crucial to avoid false positive and negative COs. The quality of the reference genome influences SNP detection, especially in highly repetitive regions. Regions of collapsed copy number variation in the reference tend to produce false SNPs, causing problems in CO detection by introducing misleading signals of haplotype shifts. These are currently addressed by masking regions with excessively high heterozygosity and coverage. Since the linked read technology I used relies on short Illumina reads, it is still prone to misalignment which may lead to false SNP detection in the pollen data. In Chapter 3, the use of high-fidelity long reads allowed better alignment of reads and an overall higher SNP density.

One important aspect of our method is the use of pollen gamete pools to detect COs. It enables detection of many COs from the recombinant gametes of an F1 plant. From a single F1 plant, thousands of pollen may be sequenced as a pool, identifying thousands of recombination events. The method relies on linked-read sequencing to construct landscapes showing regions of high and low recombination. However, reads cannot be assigned to a gamete source, making it impossible to genotype the complete genome of every pollen and account for multiple COs occuring in the same gamete. This means the method does not allow construction of a genetic map, and cannot quantify changes in recombination frequency per gamete which could occur e.g. in meiotic gene mutants. To address this issue, barcodes could be added to uniquely identify the pollen source and thereby enable the grouping of DNA molecules from the same pollen. In turn, the sequence data can be used to find COs in every chromosome of each gamete. Alternatively, single-cell whole genome sequencing (10X sc-CNV) has been used in CO detection and genetic map construction (Campoy *et al.*, 2020), but is no longer commercially available. Single-cell RNA-seq or single-cell ATAC-seq may be used, but focusing on genes or accessible regions means that genome coverage is less than that of the sc-CNV method, reducing the sensitivity and the resolution with which COs can be detected. Given the lower resolution, both sc-RNA-seq and ATAC-seq may serve if only the frequency of CO occurrence and not the specific location of the COs in the genome is important.

The genome sequencing technology used is a key factor determining the accuracy of CO detection. Ideally, reads are accurate and long, to span many markers and help resolve detection issues in regions with low SNP density. Evaluating multiple sequencing technologies may thus help optimize CO prediction. In this study, I tested both linked reads and long read technologies, which have different cost and error rates. HiFi reads span longer physical regions than short

reads and more markers than linked reads, but also contain numerous indel errors in homopolymeric regions, causing issues in CO and gene conversion (GC) detection. Linked read molecules are longer than HiFi reads, but their physical coverage is limited by the sparse short reads. Recently, improvements have been proposed to long read sequencing (e.g PacBio Revio, Oxford Nanopore Technologies (ONT) Duplex) or alternative linked read sequencing methods (e.g TellSeq, stLFR) which may aid in better detection of both COs and GCs. Both the new PacBio and ONT upgrades promise better basecalling accuracy. Our method needs only a few adjustments to process input data from other sequencing platforms and the algorithm can easily be fine-tuned to the nature of the sequence data. When CO localization is less important and cost is an issue, it is better to explore TellSeq and stLFR. A recently developed method called "haplotagging" offers simple, rapid linked-read sequencing that allows whole-genome haplotyping of hundreds of individuals for significantly lower cost (Meier *et al.*, 2021). This method may be used to detect recombinant haplotypes, just like with 10X linked reads. On the other hand, HiFi can be used to identify GC when errors in homopolymeric regions can be resolved well, such as through the use of DeepConsensus (Baid *et al.*, 2023). Similarly, the new ONT Duplex sequencing approach may not only allow GC detection but should increase sensitivity as well, given the longer read length.

While read length and quality are important issues, one of the main challenges in meiosis research is the ability to profile recombination landscapes of many crosses in a cost-effective way. Although our method has reduced the costs of producing a recombination map, it is still not as cheap as we would like it to be. Generating recombination landscapes for every pair of species in the tomato clade can help unravel the recombination bottlenecks and incompatibilities, but it requires significant investments given the current cost of our method. Thus, the selection or development of cheaper sequencing solutions should be pursued.

As indicated in Chapter 2, our method relies on the germination of pollen to isolate the DNA, which has implications that may need to be addressed depending on the objective of the study. Compared to sequencing of offspring plants, sequencing of germinated pollen can account for COs detected from unviable gametes or for improper resolution of double-strand breaks (DSBs) such as unequal COs, which may have repercussions on plant growth and survival. However, it still cannot account for CO events in gametes with germination issues. Modifying the isolation protocol, such as using sonication or beads for pollen exine rupture, can circumvent this limitation. Differences in the exine hardness and pollen size among plant species can lead to inefficient DNA isolation. An important consideration for future studies is the evaluation and calibration of the isolation protocol or the lysis of pollen grain to free template DNA (Swenson and Gemeinholzer, 2021; Moore *et al.*, 2022).

It has been previously reported that there are differences in the frequency and localization of male and female COs (Kianian *et al.*, 2018; Fernandes *et al.*, 2018; Giraut *et al.*, 2011). Both Arabidopsis and maize show dramatic differences in

distribution of male and female COs along the chromosomes, which is vital in understanding the patterning of COs in the genome and role of male and female meiosis in gene evolution. In tomato, the frequency of female chiasma is higher than male, which may be due to the longer duration of female meiotic prophase (Havekes *et al.*, 1997). By comparing female COs to genomic features, we can check whether certain features influence the location of COs, possibly differentiating them from male COs. Since our method relies on pollen sequencing, we can only detect male COs. Future efforts should include the isolation and sequencing of gametes from ovules. Isolation of the high-molecular-weight DNA from ovules may be more complicated than from pollen, due to the need to dissect the flowers and separate other organs enveloping the ovary. In many pollinating crops, there are more pollen grains than ovules (Cruden, 2000; Gillet and Gregorius, 2020), which influences both the amount of DNA material that can be collected and sequenced without amplification and the number of female COs that can be detected from an F1 plant.

6.1.2 Gene conversions

Noncrossover events (NCOs) such as gene conversions (GCs) are also products of DSB resolution generating genetic diversity. Some DSBs are resolved using the homologous chromosome as template, producing GCs whenever polymorphisms are present, or COs in case the chromosomes have reciprocal exchanges of segments. Alternatively, the breaks may be resolved using the sister chromatid or the identical replicated copy of the chromosome, which does not generate GCs. A study in Arabidopsis shows that GC frequency is influenced by environmental cues such as elevated temperature (Sun *et al.*, 2012), suggesting that they can be modulated to increase genetic diversity or generate new alleles of genes. Finding GCs is thus important to better understand the factors that determine the fate of a DSB, either to resolve into a CO or a NCO, and the role of GCs in the evolution of a species (Chen *et al.*, 2007; Duret and Galtier, 2009; Cossu *et al.*, 2017; Harpak *et al.*, 2017). Yet, in contrast to COs, few studies focus on the detection and characterization of NCOs. A NCO can be detected only when the repair strand containing a polymorphism results in a gene conversion (GC). Due to the small genomic tracts, in most cases only one polymorphism is involved, and hence they are difficult to discover. Therefore, only a few GCs have been reported thus far. For example, a study on Arabidopsis meiotic tetrads detected far fewer GCs than COs (Wijnker *et al.*, 2013). In Chapter 2 and 3, few GCs were identified using our pipeline as well, but we expect there to be more than we can currently detect from the read data. It was reported that many GCs occur in repeats or duplicated regions (Xu *et al.*, 2008; Trombetta *et al.*, 2016; Cossu *et al.*, 2017; Harpak *et al.*, 2017), so detection may be hampered by misalignment of the short Illumina reads or by false variant calls. The emergence of long read technologies with higher base call accuracy may enhance GC detection, particularly in repeat regions. Future efforts must be exerted to improve the detection of GCs, which can advance the understanding of GC formation and functions in the genome.

6.1.3 Structural variants and transposable elements

Based on the results presented in Chapter 4 and 5, we found a significant role of structural variants (SVs) and transposable elements (TEs) in shaping the patterns of recombination, both during the evolution of tomato and in experimental hybrid crosses. This further emphasizes the importance of characterizing the structure of the genome and TE occurrence profiles. However, accurate detection of SVs is still a challenge, even in human genome studies for which there are less issues with inaccurate or incomplete assemblies compared to plants. Our SV profile of parental genomes in Chapter 5 is limited to deletions and inversions and does not account for large SVs and other SV types across the genome. The complexity and length of some SVs make it hard or even impossible to identify them by some sequencing technologies and algorithms. Moreover, the conventional mapping of reads to a single reference to detect SVs introduces reference bias and suffers from lower sensitivity in detecting more complex rearrangements. Newer approaches to detect SVs rely on the comparison of genome assemblies, but are limited to just a few species or individuals. The lowering cost and increasing base call accuracy of long read sequencing may soon permit assembly of more individuals, particularly breeding lines. In tomato, 100 accessions were recently sequenced using ONT, which allowed the discovery of 238,000 SVs (Alonge *et al.*, 2020) and assembly of 14 accessions. It is however limited to detecting SVs with lengths up to a few hundred kilobases, due to the inability of reads to span very large SVs and the limitations of the detection software on reference-aligned ONT reads. This means missing larger events that may have wider impact on recombination landscapes; such events might be particularly relevant to breeding if they contain multiple genes. Finding ways to improve SV detection, for example by combining technologies like long reads, optical mapping and higher-throughput chromosome conformation capture (Hi-C) (van Belzen *et al.*, 2021; Schopflin *et al.*, 2022; Uppuluri *et al.*, 2022) and more advanced algorithms (Zook *et al.*, 2020; Cleal and Baird, 2022; Lin *et al.*, 2022; Schikora-Tamarit and Gabaldon, 2022), may enable a comprehensive comparison of SV and CO regions. These technologies can complement each other to overcome their individual limitations in detecting larger SVs and specific SV types.

For finding SVs in heterozygous genomes or highly outcrossing wild relatives, it is suitable to generate haplotype-resolved assemblies and then perform pair-wise comparison of assemblies. Both the haplotype and zygosity information from the assemblies can indicate if an SV will result in CO suppression or not. A heterozygous SV in one individual may not always cause CO suppression when crossed with another individual, as this depends on the allele that is paired with the homologous chromosome. A complete catalog of SVs in the parental genomes would show cross-specific CO suppression and probably reveal syntenic regions that can facilitate recombination in finer detail. One possible follow up project to this thesis is the assembly of the genomes of all wild tomato relatives used in Chapter 5. This will not only support comparison of SVs and COs, but can even reveal if recombination in

one region can result in the loss and gain of a genomic segment in the resulting gametes. This gain and loss may affect duplicated and translocated regions and may be linked with reproductive incompatibility or reduced viability in hybrid crosses.

Similar to SVs, finding the presence and absence of TEs is challenging using short read-based sequencing technologies such as the 10X Genomics linked reads, because of misalignment problems. Furthermore, finding TE insertions requires targeted or localized assembly of the insertion sequence, which is problematic with short reads. Comparing the TE profiles of compatible species or individuals may reveal how TE insertions or deletions can contribute to divergence of the recombination landscape. To test the influence of specific TEs on promoting DSBs, two *Ty* element copies were deleted in *Saccharomyces cerevisiae* genome, which consequently reduced DSB levels in nearby regions (Sasaki *et al.*, 2013). Deleting or inserting a CO-associated TE in a CO hotspot or coldspot region may thus reveal its direct role in shaping the recombination landscape. CO hotspots overlapping a TE insertion in a gene promoter are particularly interesting to investigate, because of the possible impact on both CO incidence and gene activity. As we have reported in Chapter 5, regions of higher density of *Gypsy* and *Copia* elements tend to have low levels of recombination compared to random parts of the genome. Removing these retrotransposon clusters may modify the recombination patterns in nearby regions.

Most analyses in this study dealt with genetic information, but as mentioned above epigenetic features have been previously reported to associate strongly with recombination events. Nucleosome-free gene promoter and terminator regions, and repetitive elements have elevated COs compared to nucleosome-dense regions (Choi *et al.*, 2013; Choi *et al.*, 2018; Wijnker *et al.*, 2013). Furthermore, remodeling of epigenetic marks such as through *met1* mutants for loss of CG DNA methylation (Choi *et al.*, 2018) and *cmt3* mutants for loss of non-CG DNA methylation have led to heterochromatin decondensation (Underwood *et al.*, 2018), but only the latter altered the distribution of COs. Historical recombination hotspots overlap nucleosomes that have been found located in DNA hypomethylated gene promoters containing the histone variant H2A.Z (Choi *et al.*, 2013). On the other hand, compact chromatin conformation, such as in regions enriched with histone H3K9me, suppresses recombination (Termolino *et al.*, 2016). Apart from the local chromatin structure, it was found that 3D genome architecture, specifically the A and B compartments, which correspond to active and inactive chromosome regions, respectively, correlates with meiotic recombination (Golicz *et al.*, 2020; Jin *et al.*, 2021; Patel *et al.*, 2019). In mice, topologically associated domains (TADs), an organizational feature of interphase chromosomes, are lost in meiotic prophase while retaining A/B compartments (Patel *et al.*, 2019). Moreover, gene-dense A compartments are preferential sites for DSBs and COs. This implies that defining the boundaries of these compartments can help differentiate CO-rich and suppressed regions. It is worthwhile exploring the impact of the epigenetic configuration on the landscape of recombination, particularly for crosses between structurally divergent parental genomes.

We discussed in Chapter 5 that distinct TE superfamilies are associated with COs and that their chromatin accessibility or inaccessibility in meiocytes may determine whether they promote or inhibit recombination. Although it was reported that in maize *Mu* transposon insertions and meiotic recombination co-colocalize in open chromatin regions (Liu *et al.*, 2009), it is not clear if *Mu* insertions may alter the chromatin configuration. A recent study in mice revealed that specific TE subfamilies, particularly of LINEs, promoted strain-specific changes in chromatin accessibility (Ferraj *et al.*, 2023). Methylation in classes of TEs is regulated by small RNA (sRNA)-directed DNA methylation (RdDM) and can be disrupted by the loss of 23- to 24-nt mobile sRNAs (Lewsey *et al.*, 2016). Parent *et al.* (2021) found that small RNA (sRNA)-dependent differentially methylated regions (DMRs) are enriched in DNA transposons and sRNA-independent DMRs are enriched in retrotransposons (*Copia* and *Gypsy* elements). Another study reported meiocyte-specific sRNA in A. thaliana to be overrepresented in Helitron elements, but underrepresented in retrotransposons (Huang *et al.*, 2019). A mutation in maize *mop1*, which is a component of the RdDM pathway, resulted in redistribution of recombination towards the distal chromosomal regions (Zhao *et al.*, 2021). Apparently, the change in epigenetic profile of TEs can alter the recombination landscape, thus it is valid to speculate that new TE insertions or burst TE activity may have an impact on CO distribution. Exploring the epigenetic modification of TEs may thus provide understanding and opportunity to alter CO patterns, especially in plant species with high TE content.

6.1.4 Population-scale recombination rates

In Chapter 4, I collected resequencing data of tomato and wild relatives from previous studies and used it to estimate population-scaled recombination rates and find historical hotspots. One of the important findings presented is the effect of domestication on recombination patterns in tomato, particularly the loss of many recombination hotspots in agronomically important genomic regions. This demonstrates the use of readily available resequencing data of natural populations to study the evolution of recombination patterns in tomato, complementing the insights derived from experimental hybrid data without additional sequencing cost. Future work should focus on differentiating patterns of historical recombination in populations of domesticated and more distant wild relatives (e.g. *S. pennelli* accessions), which may help explain incompatibilities when crossing accessions coming from these populations. The abundance of resequencing data available in public databases makes it possible to extend this analysis of historical recombination to other species and to compare the diversity of recombination patterns between populations (i.e. spatially separated populations). This will reveal possible similarities or differences between species of TE superfamilies associating with COs (Marand *et al.*, 2017; Schwarzkopf *et al.*, 2020) and expose modifiers of CO hotspot positioning such as temperature, precipitation, isothermality and solar radiation (Dreissig *et al.*, 2019). Knowing the mechanism(s) underlying gain or loss

of hotspots between wild and domesticated species can help further understand how to modulate CO frequency and positioning, which may aid in accelerating breeding. From a broader perspective, comparing historical recombination between populations may enable us to examine if the changing climate has been shaping the landscape of recombination, particularly the finer changes which may be linked with particular genes. Such insights subsequently may enable us to better understand possible threats to the survival and productivity of crops.

Analyzing the natural population of wild and domesticated tomato uncovered regions of suppressed recombination across different accessions. Compared to CO data derived from a single hybrid cross, historical recombination allows to define common hotspots and coldspots based on multiple accessions. Coldspots may differ between hybrids (Chapter 5), but population-conserved coldspots may help better characterize genomic features suppressing recombination, such as in the pericentric heterochromatin. For example, in training machine learning models to predict COs, data from hybrid crosses are insufficient to properly define the negative or non-recombining regions. One essential finding in this thesis that can improve machine learning models is the coldspot information from the natural populations, which define CO suppression common among different accessions. Training the model using information from a single cross may only account for the recombination profile specific to a cross; combining the historical recombination data with COs detected in hybrids may allow for more accurate prediction in different crosses or even species.

6.2 Relevance to breeding and future directions

The study of recombination can aid breeding by exposing regions that allow or inhibit recombination. With the recombination landscape, either from hybrids or natural populations, it is possible to identify clusters of linked genes that recombination cannot break up, such as those in the pericentromere. Initiating recombination in these regions may create genetic diversity that can confer novel and advantageous phenotypes. For example, both the introgression of *Mi-1* and *Tm-2²* from *Solanum peruvianum* are associated with CO suppression due to the presence of inversions (Kaloshian *et al.*, 1998; Lanfermeijer *et al.*, 2005). Although they confer disease resistance in the susceptible cultivar, they introduce structural heterozygosity in large genomic regions, inhibiting recombination between hundreds of genes even after repeated backcrossing (Lin *et al.*, 2014). Using high-resolution recombination data, we can predict whether an introgression is accompanied by an unfavorable allele, i.e. linkage drag. Conventionally, breeders generate a large recombinant population and screen it to find their desired progenies; the presence of linkage drag may make it impossible to find these. This problem is particularly relevant for long-generation crops. To safeguard breeders from issues related to recombination barriers and linkage drag, we propose to annotate genomic rearrangements and TEs in breeding lines and use these to check the region of interest before attempting crossing. Investment in generating

recombination data, such as using our method, to analyze CO patterning should be considered to become a part of breeding programs. It is not feasible yet to do it for every cross due to the prohibitive cost, but the development of methods like ours may eventually render it affordable enough to routinely profile COs for each breeding scheme.

Recently, a super pan-genome of multiple species representing the major clades of both wild and cultivated tomato was released (Li *et al.*, 2023). This new resource can help us further improve our comparison of COs and SVs in Chapter 5, revealing in finer detail regions prone to recombine and regions devoid of COs due to structural heterozygosity. It can provide new insights on how TE expansion and contraction contribute to large-scale CO suppression, especially in pericentric heterochromatin. More importantly, these assemblies may reveal breeding bottlenecks by enabling the recognition of unwanted linkage and the underlying features causing the absence of recombination. Linkage drag became an even more interesting subject given the report that across these tomato genomes, 46% of gene families are dispensable or accession specific (Li *et al.*, 2023), some of which are located in SVs and possibly linked with alleles selected against during domestication. Furthermore, precise delineation of non-syntenic and repetitive regions in an assembly may help refine the boundaries of CO coldspots, which currently are not available with base-pair precision due to the use of a sliding-window to measure CO frequencies along the chromosome.

Understanding how recombination plays a role in the adaptation and survival of species in response to stresses allows breeders to harness it for crop improvement. We can expose some plant species to stresses and observe genomic regions prone to increased recombination as a result. For example, resistance genes undergo rapid diversification through elevated recombination to create new haplotypes, contributing to the generation of heritable variation (Hulbert *et al.*, 2001; Zhong and Priest, 2010). It was previously reported that certain TE families activate when the plant is under stress (Negi *et al.*, 2016; Roquis *et al.*, 2021), such as *Tnt1* in tobacco (Beguiristain *et al.*, 2001), *mPing* in rice (Naito *et al.*, 2009) and *ONSEN* in Arabidopsis (Cavrak *et al.*, 2014). In maize, at least four TE families have been transcriptionally activated in response to abiotic stress, contributing to the up-regulation of nearby genes (Makarevitch *et al.*, 2015). Heat stress can attenuate epigenetic regulation in TEs, decondensing heterochromatin and losing nucleosomes (Pecinka *et al.*, 2010), and can increase crossover frequency (Si *et al.*, 2015). It would be interesting to know if sudden activation of inactive TEs, especially under stresses like temperature, can modify the recombination locations and frequencies in a individual or its progenies (Ito *et al.*, 2011; Zervudacki *et al.*, 2018), substantiating the active role of TEs in the evolution of recombination landscape. This may not only expose the role of TEs in plant response to changing environments but can help us model the impact of global warming on important crops, particularly on their gametes.

Downstream analyses of our CO data uncovered features that are strongly associated with CO incidence or inhibition such as sequence motifs, distances from genes or promoters, TEs, SVs, etc. These features can be used to develop a machine learning (ML) model to predict CO regions in a specific cross and possibly in different species (Demirci *et al.*, 2018). A single model for multiple species that can accurately determine if a cross may have recombination issues will be a valuable tool to guide and accelerate breeding, informing breeders about potential incompatibility issues even before making the cross. The application of ML has been increasingly popular in breeding and has been producing valuable solutions to biological problems (van Dijk *et al.*, 2021). Extending the application of ML to the study of meiotic recombination may help answer relevant problems such as predicting open chromatin regions or three-dimensional genome architecture using genome sequences. It was previously demonstrated that epigenomic features are also highly predictive of recombination sites (Lian *et al.*, 2022; Peñuela *et al.*, 2022). There is, however, a lack of epigenomic data for large populations as compared to genomic data. Nonetheless, there are prediction models for chromatin state and genome architecture that rely on genome sequences (Wrightsman *et al.*, 2022; Zhou, 2022). Both open chromatin and genome structure are associated with CO distribution in the genome, thus the ability to predict them from existing sequence data may help predicting the landscape of recombination.

In Chapter 5 we studied the barriers to recombination by comparing multiple hybrids. To address recombination suppression in a specific region, we propose the use of putative bridge accessions by presenting a large inversion as an example. Resequencing data of wild and domesticated tomato was mined to identify accessions that do not cause a barrier in a genomic region. Future efforts should scale the investigation of potential barrier SVs to the whole genome and provide a computational solution that lets breeders find putative bridge accessions for any region of interest. It would be even better to routinely check if any introgression or breeding process will undesirably break adaptive linkages or introduce tight linkages, surveying important agronomic and resistance genes. This computational solution can even be extended, to recommend a breeding scheme that facilitates successive introgressions with less or no undesirable linkages and with a set of desirable alleles. The prospect is to design a crop with a mosaic of alleles from a series of guided introgressions, based on the profile of recombining and non-recombining regions. To provide the genetic diversity for mining, more wild accessions should be sequenced. This process of guided introgression would facilitate efficient and precise ways of designing crops.

6.3 References

Alonge, M., Wang, X., Benoit, M., Soyk, S., Pereira, L., Zhang, L., Suresh, H., Ramakrishnan, S., Maumus, F., Ciren, D.*, et al.* 2020. Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell,* 182**,** 145-161 e23.

Baid, G., Cook, D. E., Shafin, K., Yun, T., Llinares-Lopez, F., Berthet, Q., Belyaeva, A., Topfer, A., Wenger, A. M., Rowell, W. J.*, et al.* 2023. DeepConsensus improves the accuracy of sequences with a gap-aware sequence transformer. *Nat Biotechnol,* 41**,** 232-238.

Beguiristain, T., Grandbastien, M. A., Puigdomenech, P. & Casacuberta, J. M. 2001. Three Tnt1 subfamilies show different stress-associated patterns of expression in tobacco. Consequences for retrotransposon control and evolution in plants. *Plant Physiol,* 127**,** 212-21.

Campoy, J. A., Sun, H., Goel, M., Jiao, W. B., Folz-Donahue, K., Wang, N., Rubio, M., Liu, C., Kukat, C., Ruiz, D.*, et al.* 2020. Gamete binning: chromosome-level and haplotype-resolved genome assembly enabled by high-throughput single-cell sequencing of gamete genomes. *Genome Biol,* 21**,** 306.

Cavrak, V. V., Lettner, N., Jamge, S., Kosarewicz, A., Bayer, L. M. & Mittelsten Scheid, O. 2014. How a retrotransposon exploits the plant's heat stress response for its activation. *PLoS Genet,* 10**,** e1004115.

Chen, J. M., Cooper, D. N., Chuzhanova, N., Ferec, C. & Patrinos, G. P. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet,* 8**,** 762-75.

Choi, K., Zhao, X., Kelly, K. A., Venn, O., Higgins, J. D., Yelina, N. E., Hardcastle, T. J., Ziolkowski, P. A., Copenhaver, G. P., Franklin, F. C.*, 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., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J.*, et al.* 2018. Nucleosomes and DNA methylation shape meiotic DSB frequency in *Arabidopsis thaliana* transposons and gene regulatory regions. *Genome Research,* 28**,** 532-546.

Cleal, K. & Baird, D. M. 2022. Dysgu: efficient structural variant calling using short or long reads. *Nucleic Acids Res,* 50**,** e53.

Cossu, R. M., Casola, C., Giacomello, S., Vidalis, A., Scofield, D. G. & Zuccolo, A. 2017. LTR Retrotransposons Show Low Levels of Unequal Recombination and High Rates of Intraelement Gene Conversion in Large Plant Genomes. *Genome Biol Evol,* 9**,** 3449-3462.

Cruden, R. W. 2000. Pollen grains: Why so many? *Plant Systematics and Evolution,* 222**,** 143-165.

Demirci, S., Peters, S. A., de Ridder, D. & van Dijk, A. D. J. 2018. DNA sequence and shape are predictive for meiotic crossovers throughout the plant kingdom. *Plant J*.

Dreau, A., Venu, V., Avdievich, E., Gaspar, L. & Jones, F. C. 2019. Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat Commun,* 10**,** 4309.

Dreissig, S., Mascher, M. & Heckmann, S. 2019. Variation in Recombination Rate Is Shaped by Domestication and Environmental Conditions in Barley. *Mol Biol Evol,* 36**,** 2029-2039.

Duret, L. & Galtier, N. 2009. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu Rev Genomics Hum Genet,* 10**,** 285-311.

Fernandes, J. B., Seguela-Arnaud, M., Larcheveque, C., Lloyd, A. H. & Mercier, R. 2018. Unleashing meiotic crossovers in hybrid plants. *Proc Natl Acad Sci U S A,* 115**,** 2431-2436.

Ferraj, A., Audano, P. A., Balachandran, P., Czechanski, A., Flores, J. I., Radecki, A. A., Mosur, V., Gordon, D. S., Walawalkar, I. A., Eichler, E. E.*, et al.* 2023. Resolution of structural variation in diverse mouse genomes reveals chromatin remodeling due to transposable elements. *Cell Genomics*.

Gillet, E. M. & Gregorius, H. R. 2020. Effects of reproductive resource allocation and pollen density on fertilization success in plants. *BMC Ecol,* 20**,** 26.

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O. C. & Mezard, C. 2011. Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along chromosomes. *PLoS Genet,* 7**,** e1002354.

Golicz, A. A., Bhalla, P. L., Edwards, D. & Singh, M. B. 2020. Rice 3D chromatin structure correlates with sequence variation and meiotic recombination rate. *Communications Biology,* 3.

Harpak, A., Lan, X., Gao, Z. & Pritchard, J. K. 2017. Frequent nonallelic gene conversion on the human lineage and its effect on the divergence of gene duplicates. *Proc Natl Acad Sci U S A,* 114**,** 12779-12784.

Havekes, F. W., Jong, J. H. & Heyting, C. 1997. Comparative analysis of female and male meiosis in three meiotic mutants of tomato. *Genome,* 40**,** 879-86.

Huang, J., Wang, C., Wang, H., Lu, P., Zheng, B., Ma, H., Copenhaver, G. P. & Wang, Y. 2019. Meiocyte-Specific and AtSPO11-1-Dependent Small RNAs and Their Association with Meiotic Gene Expression and Recombination. *Plant Cell,* 31**,** 444-464.

Hulbert, S. H., Webb, C. A., Smith, S. M. & Sun, Q. 2001. Resistance gene complexes: evolution and utilization. *Annu Rev Phytopathol,* 39**,** 285-312.

Ito, H., Gaubert, H., Bucher, E., Mirouze, M., Vaillant, I. & Paszkowski, J. 2011. An siRNA pathway prevents transgenerational retrotransposition in plants subjected to stress. *Nature,* 472**,** 115-9.

Jin, X., Fudenberg, G. & Pollard, K. S. 2021. Genome-wide variability in recombination activity is associated with meiotic chromatin organization. *Genome Res,* 31**,** 1561-1572.

Kaloshian, I., Yaghoobi, J., Liharska, T., Hontelez, J., Hanson, D., Hogan, P., Jesse, T., Wijbrandi, J., Simons, G., Vos, P.*, et al.* 1998. Genetic and physical localization of the root-knot nematode resistance locus mi in tomato. *Mol Gen Genet,* 257**,** 376-85.

Kianian, P. M. A., Wang, M., Simons, K., Ghavami, F., He, Y., Dukowic-Schulze, S., Sundararajan, A., Sun, Q., Pillardy, J., Mudge, J.*, et al.* 2018. High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nat Commun,* 9**,** 2370.

Lanfermeijer, F. C., Warmink, J. & Hille, J. 2005. The products of the broken Tm-2 and the durable Tm-2(2) resistance genes from tomato differ in four amino acids. *J Exp Bot,* 56**,** 2925-33.

Lewsey, M. G., Hardcastle, T. J., Melnyk, C. W., Molnar, A., Valli, A., Urich, M. A., Nery, J. R., Baulcombe, D. C. & Ecker, J. R. 2016. Mobile small RNAs regulate genome-wide DNA methylation. *Proc Natl Acad Sci U S A,* 113**,** E801-10.

Li, N., He, Q., Wang, J., Wang, B., Zhao, J., Huang, S., Yang, T., Tang, Y., Yang, S., Aisimutuola, P.*, et al.* 2023. Super-pangenome analyses highlight genomic diversity and structural variation across wild and cultivated tomato species. *Nat Genet*.

Lian, Q., Solier, V., Walkemeier, B., Durand, S., Huettel, B., Schneeberger, K. & Mercier, R. 2022. The megabase-scale crossover landscape is largely independent of sequence divergence. *Nat Commun,* 13**,** 3828.

Lin, J., Wang, S., Audano, P. A., Meng, D., Flores, J. I., Kosters, W., Yang, X., Jia, P., Marschall, T., Beck, C. R.*, et al.* 2022. SVision: a deep learning approach to resolve complex structural variants. *Nat Methods,* 19**,** 1230-1233.

Lin, T., Zhu, G., Zhang, J., Xu, X., Yu, Q., Zheng, Z., Zhang, Z., Lun, Y., Li, S., Wang, X.*, et al.* 2014. Genomic analyses provide insights into the history of tomato breeding. *Nat Genet,* 46**,** 1220-6.

Liu, S., Yeh, C. T., Ji, T., Ying, K., Wu, H., Tang, H. M., Fu, Y., Nettleton, D. & 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.

Makarevitch, I., Waters, A. J., West, P. T., Stitzer, M., Hirsch, C. N., Ross-Ibarra, J. & Springer, N. M. 2015. Transposable Elements Contribute to Activation of Maize Genes in Response to Abiotic Stress. *PLoS Genetics,* 11.

Marand, A. P., Jansky, S. H., Zhao, H., Leisner, C. P., Zhu, X., Zeng, Z., Crisovan, E., Newton, L., Hamernik, A. J., Veilleux, R. E.*, et al.* 2017. Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biol,* 18**,** 203.

Meier, J. I., Salazar, P. A., Kucka, M., Davies, R. W., Dreau, A., Aldas, I., Box Power, O., Nadeau, N. J., Bridle, J. R., Rolian, C.*, et al.* 2021. Haplotype tagging reveals parallel formation of hybrid races in two butterfly species. *Proc Natl Acad Sci U S A,* 118.

Moore, M. A., Scheible, M. K. R., Robertson, J. B. & Meiklejohn, K. A. 2022. Assessing the lysis of diverse pollen from bulk environmental samples for DNA metabarcoding. *Metabarcoding and Metagenomics,* 6.

Naito, K., Zhang, F., Tsukiyama, T., Saito, H., Hancock, C. N., Richardson, A. O., Okumoto, Y., Tanisaka, T. & Wessler, S. R. 2009. Unexpected consequences of a sudden and massive transposon amplification on rice gene expression. *Nature,* 461**,** 1130-1134.

Negi, P., Rai, A. N. & Suprasanna, P. 2016. Moving through the Stressed Genome: Emerging Regulatory Roles for Transposons in Plant Stress Response. *Front Plant Sci,* 7**,** 1448.
Parent, J. S., Cahn, J., Herridge, R. P., Grimanelli, D. & Martienssen, R. A. 2021. Small RNAs guide histone methylation in Arabidopsis embryos. *Genes Dev,* 35**,** 841-846.
Patel, L., Kang, R., Rosenberg, S. C., Qiu, Y., Raviram, R., Chee, S., Hu, R., Ren, B., Cole, F. & Corbett, K. D. 2019. Dynamic reorganization of the genome shapes the recombination landscape in meiotic prophase. *Nat Struct Mol Biol,* 26**,** 164-174.
Pecinka, A., Dinh, H. Q., Baubec, T., Rosa, M., Lettner, N. & Mittelsten Scheid, O. 2010. Epigenetic regulation of repetitive elements is attenuated by prolonged heat stress in Arabidopsis. *Plant Cell,* 22**,** 3118-29.
Peñuela, M., Gallo-Franco, J. J., Finke, J., Rocha, C., Gkanogiannis, A., Ghneim-Herrera, T. & Lorieux, M. 2022. Methylation in the CHH Context Allows to Predict Recombination in Rice. *International Journal of Molecular Sciences,* 23.
Roquis, D., Robertson, M., Yu, L., Thieme, M., Julkowska, M. & Bucher, E. 2021. Genomic impact of stress-induced transposable element mobility in Arabidopsis. *Nucleic Acids Res,* 49**,** 10431-10447.
Sasaki, M., Tischfield, S. E., van Overbeek, M. & Keeney, S. 2013. Meiotic recombination initiation in and around retrotransposable elements in *Saccharomyces cerevisiae*. *PLoS Genet,* 9**,** e1003732.
Schikora-Tamarit, M. A. & Gabaldon, T. 2022. PerSVade: personalized structural variant detection in any species of interest. *Genome Biol,* 23**,** 175.
Schopflin, R., Melo, U. S., Moeinzadeh, H., Heller, D., Laupert, V., Hertzberg, J., Holtgrewe, M., Alavi, N., Klever, M. K., Jungnitsch, J.*, et al.* 2022. Integration of Hi-C with short and long-read genome sequencing reveals the structure of germline rearranged genomes. *Nat Commun,* 13**,** 6470.
Schwarzkopf, E. J., Motamayor, J. C. & Cornejo, O. E. 2020. Genetic differentiation and intrinsic genomic features explain variation in recombination hotspots among cocoa tree populations. *BMC Genomics,* 21**,** 332.
Si, W., Yuan, Y., Huang, J., Zhang, X., Zhang, Y., Zhang, Y., Tian, D., Wang, C., Yang, Y. & Yang, S. 2015. Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants. *New Phytol,* 206**,** 1491-502.
Sun, H., Rowan, B. A., Flood, P. J., Brandt, R., Fuss, J., Hancock, A. M., Michelmore, R. W., Huettel, B. & Schneeberger, K. 2019. Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination. *Nat Commun,* 10**,** 4310.
Sun, Y., Ambrose, J. H., Haughey, B. S., Webster, T. D., Pierrie, S. N., Munoz, D. F., Wellman, E. C., Cherian, S., Lewis, S. M., Berchowitz, L. E.*, et al.* 2012. Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in *Arabidopsis thaliana*. *PLoS Genet,* 8**,** e1002968.
Swenson, S. J. & Gemeinholzer, B. 2021. Testing the effect of pollen exine rupture on metabarcoding with Illumina sequencing. *PLoS One,* 16**,** e0245611.
Termolino, P., Cremona, G., Consiglio, M. F. & Conicella, C. 2016. Insights into epigenetic landscape of recombination-free regions. *Chromosoma,* 125**,** 301-8.
Trombetta, B., Fantini, G., D'Atanasio, E., Sellitto, D. & Cruciani, F. 2016. Evidence of extensive non-allelic gene conversion among LTR elements in the human genome. *Sci Rep,* 6**,** 28710.
Underwood, C. J., Choi, K., Lambing, C., Zhao, X., Serra, H., Borges, F., Simorowski, J., Ernst, E., Jacob, Y., Henderson, I. R.*, 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.
Uppuluri, L., Wang, Y., Young, E., Wong, J. S., Abid, H. Z. & Xiao, M. 2022. Multiplex structural variant detection by whole-genome mapping and nanopore sequencing. *Sci Rep,* 12**,** 6512.
van Belzen, I., Schonhuth, A., Kemmeren, P. & Hehir-Kwa, J. Y. 2021. Structural variant detection in cancer genomes: computational challenges and perspectives for precision oncology. *NPJ Precis Oncol,* 5**,** 15.
van Dijk, A. D. J., Kootstra, G., Kruijer, W. & de Ridder, D. 2021. Machine learning in plant science and plant breeding. *iScience,* 24**,** 101890.
Wijnker, E., Velikkakam James, G., Ding, J., Becker, F., Klasen, J. R., Rawat, V., Rowan, B. A., de Jong, D. F., de Snoo, C. B., Zapata, L.*, et al.* 2013. The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*. *Elife,* 2**,** e01426.

Wrightsman, T., Marand, A. P., Crisp, P. A., Springer, N. M. & Buckler, E. S. 2022. Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks. *Plant Genome,* 15**,** e20249.

Xu, S., Clark, T., Zheng, H., Vang, S., Li, R., Wong, G. K., Wang, J. & Zheng, X. 2008. Gene conversion in the rice genome. *BMC Genomics,* 9**,** 93.

Zervudacki, J., Yu, A., Amesefe, D., Wang, J., Drouaud, J., Navarro, L. & Deleris, A. 2018. Transcriptional control and exploitation of an immune-responsive family of plant retrotransposons. *EMBO J,* 37.

Zhao, M., Ku, J. C., Liu, B., Yang, D., Yin, L., Ferrell, T. J., Stoll, C. E., Guo, W., Zhang, X., Wang, D.*, et al.* 2021. The mop1 mutation affects the recombination landscape in maize. *Proc Natl Acad Sci U S A,* 118.

Zhong, W. & Priest, N. K. 2010. Stress-induced recombination and the mechanism of evolvability. *Behavioral Ecology and Sociobiology,* 65**,** 493-502.

Zhou, J. 2022. Sequence-based modeling of three-dimensional genome architecture from kilobase to chromosome scale. *Nat Genet,* 54**,** 725-734.

Zook, J. M., Hansen, N. F., Olson, N. D., Chapman, L., Mullikin, J. C., Xiao, C., Sherry, S., Koren, S., Phillippy, A. M., Boutros, P. C.*, et al.* 2020. A robust benchmark for detection of germline large deletions and insertions. *Nat Biotechnol,* 38**,** 1347-1355.

Summary

Meiotic recombination is a biological process essential for the shuffling of genetic materials and balanced segregation of chromosomes. Through meiotic recombination, breeders can create new allele combinations in crops conferring desirable traits, which is becoming urgent in the face of changing climate conditions, food production and food security. Determining the factors that influence the process of meiosis is therefore beneficial in predicting or even manipulating recombination events. Changes in the distribution of crossover events in the genome can provide insights into their role in species evolution and the bottlenecks in breeding. Classical methods like cytological techniques and offspring genotyping have been developed to profile and study recombination in different species, but the low number and coarse resolution of recombination events that can be studied limit subsequent analyses. More recent approaches rely on the generation and sequencing of offspring populations, which is quite laborious and expensive. In this thesis, I focus on developing and using computational methods to profile meiotic recombination in natural and experimental populations and perform downstream analyses to decipher genetic factors influencing recombination events.

In **Chapter 1**, I provide an overview of the recombination mechanism and its role in species evolution. The connection between recombination and plant breeding is also discussed, in particular the function of recombination in the development of crops and the efforts to manipulate it to accelerate breeding. I further explore different methods to detect recombination, from early cytology-based techniques to sequencing-based algorithms, and expound on their limitations. In **Chapter 2**, I present a method based on linked-read sequencing to detect recombination in a pool of pollen gametes from an F1 interspecific tomato hybrid plant, reducing the labor and cost of developing a recombination map compared to current approaches. This high-throughput method produces high-resolution crossovers and gene conversions, enabling comparison with gene features and sequence motifs. I extend this method to utilize long reads, specifically PacBio hi-fidelity (HiFi), to phase DNA molecules and detect crossovers (**Chapter 3**). The resulting resolution of crossovers is superior to when using linked reads.

Historical recombination in a natural population provides valuable information on species evolution. In **Chapter 4**, I generate the recombination landscape for populations of wild, early-domesticated and heirloom tomato. Comparing these populations, I find both conserved genome-wide and local changes in recombination rates. Loss and divergence in historical recombination hotspots during tomato domestication is associated with selective sweeps and with genetic changes and structural variants. Specific transposable element superfamilies are associated with elevated or reduced numbers of recombination hotspots, hinting at a role in shaping recombination patterns.

Constructing recombination maps for a single cross by screening an offspring population requires considerable time and resources, which makes the comparison of multiple maps between different crosses demanding. This prohibits exploring the

diversity of recombination patterns in different crosses and their contributions to breeding bottlenecks. In **Chapter 5**, I use the pollen-based method of chapter 2 to profile recombination in five interspecific hybrids and observe varying patterns of recombination in specific genomic regions. Structural variants and transposable elements are associated with divergent recombination patterns between hybrids, causing bottlenecks in breeding by crossover suppression. I propose to find putative bridge accessions that can circumvent the recombination bottleneck in a region of interest, aiding the development of better varieties by avoiding linkage between unwanted pairs of alleles due to genomic rearrangements.

Finally, I conclude this thesis in **Chapter 6** by describing the ongoing challenges in the detection of crossovers and gene conversions, and other associated-features like structural variants and transposable elements. I also discuss the relevance of recombination studies to breeding and set out some future directions in this research area.

List of publications

Fuentes, R.R., de Ridder, D., van Dijk, A.D.J. & Peters, S.A. Domestication Shapes Recombination Patterns in Tomato. Molecular Biology and Evolution 39, 2022.

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

Fuentes, R.R. *et al*. Meiotic recombination profiling of interspecific hybrid F1 tomato pollen by linked read sequencing. Plant J 102, 480-492, 2020.

Fuentes, R.R. *et al*. Structural variants in 3,000 rice genomes. Genome Res 29, 2019.

Wang, W.*, Mauleon, R.*, Hu, Z.*, Chebotarov, D.*, Tai, S.*, Wu, Z.*, Li, M.*, Zheng, T.*, **Fuentes, R.R.***, Zhang, F.*, Mansueto, L.*, Copetti, D.*, … Hamilton, R.S., Wing, R.A, Wei, C., Ruan, J., Zhang, G., Alexandrov, N., McNally, K.L., Li, Z., Leung, H. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature, 2018.

Mansueto, L., **Fuentes, R.R.**, Chebotarov, D. *et al.* SNP-Seek II: A resource for allele mining and analysis of big data in Oryza sativa. Current Plant Biology. 7-8:16-25, 2016.

Mansueto, L., **Fuentes, R.R.**, Borja, F.N. *et al*. Rice SNP-seek database update: new SNPs, indels, and queries. Nucleic Acids Res. 45, D1075–D1081, 2016.

Alexandrov, N., Tai, S., Wang, W., Mansueto, L., Palis, K., **Fuentes, R.R.**, Ulat, V.J., Chebotarov, D., Zhang, G., Li, Z., Mauleon, R., Hamilton, R.S., McNally, K.L. SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 43, D1023–D1027, 2014.

In preparation

Fuentes, R.R., Nieuwenhuis, R., Chouaref, J., Hesselink, T., van Dooijeweert, W., van den Broeck, H., Schijlen, E., Schouten, H.J., Bai, Y., Fransz, P., Stam, M., de Jong, H., Trivino, S.D., de Ridder, D., van Dijk, A.D.J., Peters, S.A. Aberrant patterns of recombination in interspecific tomato hybrids reveal breeding bottlenecks.

*shared first authorship

Acknowledgements

This whole endeavor of doing PhD has never crossed my mind, let alone study in a foreign land. But it became a reality through the many people who supported and guided me along the way. This thesis is dedicated to each of them.

First, I am thankful to my three supervisors, Dick, Sander and Aalt-Jan, for this 5-year journey. Despite my doubts on myself due to the lack of formal training in biology, they believed and guided me in my quest to answer interesting puzzles in meiosis. Sometimes, or probably many times, they needed to remind me not to deviate too far from my research objectives when I'm carried away by my curiosity to answer new questions. They taught me to be critical on my research and has inspired me to continue doing science in my future career. Three great brains in every meeting are both a treat and a challenge, but it made the whole process more fruitful and memorable.

I also extend my gratitude to the members of the Chair Group Bioinformatics and the Applied Bioinformatics who in many ways made this journey enjoyable. Collaborations and discussions with them led not only in solving some of my puzzles, but to broadening my knowledge of biology. The weekly insightful meetings with MEICOM collaborators are instrumental in deepening my understanding of the mechanisms of recombination and in fostering new research ideas. And I also thank the whole MEICOM community for providing me invaluable training and exposure to experts in the field.

As a foreign student, it was difficult at first to adjust in a new environment. The WUR Filipino community became my new home in the Netherlands. They accompany me in lunches, biking, parties, and tours. I thank all the Tita's of Wageningen for all the parties and feeding program for hungry Filipino students. And of course, I won't forget all the karaoke sessions over the years, practicing our favorite Filipino hobby responsibly.

While finishing my thesis, I joined Charlie and his lab to continue my quest. They made my relocation to a new country easy and has supported me in finishing my thesis. They have been teaching me to do some cool work in the lab. Joining the institute allowed me to meet a community of scientists focused on a similar research topic, lavishly feeding my curiosity. The coming years will be full of exciting scientific discoveries together.

The support of my friends back home and the KUMU squad not only encouraged me to continue my studies but helped me survived during the pandemic. It felt like we are just meters and not thousand of miles apart. The bonding we shared online lightened the load of doing a thesis while confined in a room for months. It is truly a unique experience. Salamat!

Lastly, I thank my parents and siblings for being my inspiration in pursuing this. I hope I made you all proud. Salamat sa moral support nyo para matapos ko 'tong pangarap na 'to. May maisasabit na uli sa wall. Para sa inyo 'to.

The research described in this thesis was financially supported by MEICOM Marie Sklodowska-Curie Innovative Training Network (ITN), H2020-MSCA-ITN-2017 Horizon 2020 Grant agreement number 765212. Part of the research was financed by the Netherlands Top Consortium for Knowledge and Innovation (TKI project LWV19283).

Cover design by Roven Rommel Fuentes
Cover images by Yazhong Wang