Mining genetic diversity for tomorrow's agriculture

Yu-Ming Hsu

▶ **To cite this version:**

Yu-Ming Hsu. Mining genetic diversity for tomorrow's agriculture. Plants genetics. Université Paris-Saclay, 2022. English. NNT : 2022UPASL048 . tel-03813496

HAL Id: tel-03813496
https://theses.hal.science/tel-03813496

Submitted on 13 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

université PARIS-SACLAY

Mining genetic diversity for tomorrow's agriculture

Valoriser la diversité génétique pour l'agriculture de demain

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 577 : Structure et dynamique des systèmes vivants (SDSV)
Spécialité de doctorat : Sciences de la vie et de la santé
Graduate School : Life Science and Health
Référent : Faculté des sciences d'Orsay

Thèse préparée dans les unités de recherche
GQE - Le Moulon (Université Paris-Saclay, INRAE, CNRS, AgroParisTech)
et **Institut des Sciences des Plantes de Paris Saclay (IPS2)**
(Université Paris-Saclay, Université d'Évry, Université Paris Cité, CNRS, INRAE),
sous la direction de **Matthieu FALQUE**, Ingénieur de recherche (HDR),
et la co-direction d'**Olivier C. MARTIN**, Directeur de recherche

Thèse soutenue à Paris-Saclay, le 7 septembre 2022, par

Yu-Ming HSU

Composition du Jury

Cecile FAIRHEAD Professeure, Université Paris-Saclay	Présidente
Martin HOWARD Professeur, John Innes Centre	Rapporteur & Examinateur
Mathias LORIEUX Directeur de recherche, Institut de Recherche pour le Développement	Rapporteur & Examinateur
Beth ROWAN Project Scientist, University of California, Davis	Examinatrice
Piotr ZIOLKOWSKI Professeur associé, Adam Mickiewicz University	Examinateur
Matthieu FALQUE Ingénieur de recherche, Université Paris-Saclay	Directeur de thèse

THESE DE DOCTORAT

NNT : 2022UPASL048

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Titre : Valoriser la diversité génétique pour l'agriculture de demain

Mots clés : biostatistique, informatique, inférence statistique, sélection végétale

Résumé : La diversité génétique est la clé de la sélection végétale. Ainsi, la compréhension des facteurs contribuant à la diversité génétique et cette diversité elle-même ouvrirait la voie à l'amélioration des cultures. Au cours de ma thèse, j'ai utilisé la modélisation quantitative et des approches bioinformatiques pour étudier à la fois la recombinaison méiotique, un facteur à l'origine du remaniement du génome, et la diversité génétique de deux cultures importantes, la tomate et l'arachide. Pour la recombinaison méiotique, individuellement, chacune des caractéristiques génomiques/épigénomiques explique mal les paysages croisés chez Arabidopsis thaliana. Au lieu de cela, un état épigénétique résumé, se référant à 10 états de chromatine, est capable de révéler la tendance de la distribution des crossovers. De plus, j'ai découvert qu'un niveau intermédiaire polymorphismes nucléotidiques simples (SNP) entre les homologues recrute plus de crossovers par rapport aux séquences homologues identiques, et les régions intergéniques d'une taille inférieure à 1,5 kb suppriment les crossovers. Pris ensemble, ces effets ont été intégrés dans un modèle quantitatif qui peut prédire le paysage de recombinaison reproduisant une grande partie des variations dans les données de crossing-over expérimentales.

Dans le cadre de deux autres projets liés aux cultures, j'ai évalué la diversité génétique des arachides cultivées à Taiwan par l'approche RAD (restriction site associated DNA) en utilisant 31 accessions. Mes résultats indiquent que les accessions mondiales ont une plus grande diversité génétique que les accessions locales, ce qui suggère que de nouvelles ressources génétiques devraient être introduites dans les programmes de sélection actuels pour améliorer la diversité génétique. Enfin, j'ai travaillé sur l'identification de la résistance au flétrissement bactérien (BW) chez la tomate cultivée, l'une des maladies les plus destructrices de cette culture. J'ai utilisé les données de la séquence du génome entier de six lignées de tomates résistantes et de neuf lignées sensibles au flétrissement bactérien pour identifier les polymorphismes spécifiques aux lignées résistantes. Parmi les polymorphismes spécifiques à la résistance affectant 385 gènes, le marqueur Bwr3.2dCAPS situé dans l'*Asc* (Solyc03g114600.4.1) s'est avéré être associé de manière significative à la résistance à la BW, mais néanmoins il n'explique pas entièrement le phénotype de résistance. Enfin, ces recherches successives, motivées respectivement par la biologie fondamentale et par la science appliquée de la sélection, fournissent de nouvelles perspectives qui peuvent aider les stratégies futures d'amélioration des cultures.

université PARIS-SACLAY

ÉCOLE DOCTORALE
Structure et dynamique des systèmes vivants (SDSV)

Title : Mining genetic diversity for tomorrow's agriculture

Keywords : biostatistics, computation, statistical inference, plant breeding

Abstract : Genetic diversity is the key ingredient fueling gains during plant breeding programs. Thus, understanding the structure of a germplasm's genetic diversity as well as the factors shaping it pave the way for crop improvement. During my thesis, I utilized quantitative modeling and bioinformatic approaches to study both meiotic recombination, a factor driving genome reshuffling, and the genetic diversity of two important crops, tomato and peanut. For meiotic recombination, individual genomic/epigenomic features have weak predictive power regarding the distribution of crossovers in Arabidopsis thaliana. Instead, a summarized epigenetic status, referring to 10 chromatin states, is able to reveal the associated landscape rather well. Furthermore, I found that intermediate levels of single nucleotide polymorphisms (SNPs) between homologs leads to more crossovers compared to the case of near identical sequences, and that intergenic regions of size less than 1.5 kb tend to suppress crossovers. Taken together, I integrated these effects into a quantitative model that can predict recombination landscapes and that reproduces much of the variation seen in the experimental crossover data.

Moving on to two other projects related to crops, I assessed the genetic diversity of cultivated peanuts in Taiwan by the restriction site associated DNA (RAD) approach using 31 accessions. My results indicate that worldwide accessions have greater genetic diversity than local accessions, suggesting that novel genetic resources should be introduced into the present breeding programs for enhancing the genetic diversity. Lastly, I worked on the identification of resistance against Bacterial wilt (BW) in cultivated tomato, one of the most destructive diseases in this crop. I used the whole genome sequence data of six BW resistant and nine BW susceptible tomato lines to identify polymorphisms specific to resistant lines. Among resistant-specific polymorphisms affecting 385 genes, the marker Bwr3.2dCAPS located in the *Asc* (Solyc03g114600.4.1) was shown to be significantly associated with the BW resistance but nevertheless it does not fully explain the resistance phenotype. Lastly, These successive investigations, motivated respectively by fundamental biology and by applied breeding science, provide new insights that can help future strategies for crop improvement.

Acknowledgement

Time flies. It has been about four years since I pursued my PhD study. When I got to the end of this journey, it reminded me of a sentence that I learned while in junior high school. According to Dr. Zhifan CHEN, “I always receive too much help from others compared to my help to others, because there are too many people for which I need to be grateful, I had better thank the god.” Even with this feeling, I still want to thank a lot of people that helped me during this period.

First, I would like to express my gratitude to Martin HOWARD and Mathias LORIEUX for being “rapporteurs” to read and evaluate my thesis manuscript, and I also want to thank Beth ROWAN, Cecile FAIRHEAD, and Piotr ZIOLKOWSKI for being jury members of my defense to evaluate my thesis work.

Second, I would like to express my gratitude to my thesis committee members, Mathilde GRELON, Pierre SOURDILLE, and Mathieu ROUSSEAU-GUEUTIN, for giving me many valuable opinions to improve one of three projects in my thesis work, modeling of meiotic recombination. And I also appreciate having the chance to collaborate with Derek, Hung-Yu, and Shen-Shan for the other two projects. This collaboration experience allows me to learn more and make my thesis more fruitful.

I would like to thank my colleagues in the RAMDAM team at GQE-Le Moulon and the REGARN team at IPS2: Xavier, Elise, Adrien, Aurelie, Jeremie, Christine, Céline C, Céline S, Caroline, Thomas B, Andana, Richard, Gautier, Thomas R, Michel H, Michel P, Francisco, Soledad, Jerôme, Lea-Lou, Ivana, Marianne. I appreciate very much that all of you treat me well and bring joy, energy, and an active attitude into the daily environment so that I enjoy the

time with you.

Normally, it would be lucky for PhD students to meet a great supervisor during their studies. I am luckier than them because I have two, Olivier MARTIN and Matthieu FALQUE, with both knowledge in science and wisdom in dealing with my difficulties in my study. Both of you provided me with a comfortable environment for discussing science and other issues. Whenever I have any questions or annoying problems, I know you will be supporting me. It is impossible to express my gratitude based on these few sentences, but your support in the past few years will be an important part of my life.

I would like to thank Martin CRESPI. Your passion for science and life impresses me. I will never forget how you master BBQ. Plus, your suggestions and support help me a lot to improve my thesis work.

I want to express my great gratitude for the financial support from the University of Paris Saclay and the Ministry of Education in Taiwan. Besides, I also want to thank Sandrine at Doctoral School ED577, your assistance in many complicated administrative processes is helpful and warm during this journey.

Finally, I want to thank my family and friends. Even though this journey is tough, I know I will never be alone because of your support. Last but not the least, Ching-Hui, my beloved, thanks for all the support and company these years. Without you, I couldn’t finish my thesis and become a better person.

Table of contents

Abbreviations

Ac	*Activator*
AMS	accelerator mass spectrometry
BLAP18	BLM-TOPIIIα-RMI2
BLAP75	BLM-TOPIIIα-RMI1
BW	bacterial wilt
CCA1	*CIRCADIAN CLOCK ASSOCIATED 1*
CMT2	Chromomethylase2
CMT3	Chromomethylase3
CO	crossover
DDM1	Decreased DNA Methylation1
dHJ	double-Holliday junction
DRM2	Domains Rearranged Methylase2
Ds	*Dissociation*
DSB	double strand break
dw3	*dwarf3*
egl	endoglucanase
EM	electron microscopy
EMBRAPA	Empresa Brasileira de Pesquisa Agropecuária
FDR	first division restitution
FIGL1	FIDGETIN-Like-1
FLC	*FLOWERING LOCUS C*
Fst	genetic differentiation
IBPGR	International Board for Plant Genetic Resources
ICRISAT	International Centre for Research in the Semi-Arid Tropics
IH	inter-homologous
Indel	Insertion/deletion
IPGRI	International Plant Genetic Resources Institute
ITS	internal transcribed spacer
K_S	the number of synonymous substitutions per synonymous site
LHY	*LATE ELONGATED HYPOCOTYL*
LINE	long interspersed element
LTR	long terminal repeat

MAGIC	Multiparent Advanced Generation Inter-Cross
Met1	Methyltransferase1
MLSA	multilocus sequence analysis
MMR	mismatch repair
MRN	Mre11, Rad50, and Nbs1
MRX	Mre11, Rad50, and Xrs2
NBPGR	National Bureau of Plant Genetic Resources
Nc	census population size
NCO	noncrossover
ncRNA	non-coding RNA
NDR	nucleosome-depleted region
NLD	*NOT LIKE DAD*
OCRI	Oilseed Crops Research Institute
PIC	polymorphism information content
Pol III	polymerase III
RAC1	RESISTANCE TO ALBUGO CANDIDA1
RFLP	restriction fragment length polymorphism
RSSC	*R*. *solanacearum* species complex
SC	synaptonemal complex
SDR	second division restitution
SDSA	synthesis dependent strand annealing
SINE	short interspersed element
SNP	Single Nucleotide Polymorphism
SSR	simple sequence repeat
STR	Sgs1-Top3-Rmi1
SV	structural variation
TAD	topologically associated domain
TAMU	Texas AgriLife Research Center
TE	transposable element
TIR	terminal inverted repeat
TOPOII	Topoisomerase II
TopoVIA	type II topoisomerase
ts/tv	transition/transversion
TSD	target size duplication
USDA-ARS	United States Department of Agriculture
WHD	winged-helix domain
Wva700	West Virginia 700

List of Figures

0. Genetic diversity and thesis objectives

0.1 Genetic diversity: the key ingredient fueling plant breeding

By definition, genetic diversity is the extent of distinct DNA sequences between individuals (chromosomes) of a species (population). From an evolutionary perspective, genetic diversity is shaped by spontaneous mutations, genetic drift and selection that can make populations or species better adapted to changes in their environments. In agriculture, particularly in plant breeding, exploiting genetic diversity is crucial for creating novel varieties with larger yield and improved traits. As defined by Poehlman and Sleper (1995), plant breeding is "the art and science of improving the heredity of plants for the benefit of humankind", which is a process to exploit genetic variation for selecting better individuals.

Plant domestication, an ancestral mode of plant breeding, can be traced to about 10,000 years ago, for instance in the Fertile Crescent where wild emmer wheat (*Triticum dicoccoides*) was one of the first cereals domesticated. In the cereal domestication process, undesirable traits like seed shattering and small grain sizes were removed by the artificial selection (Purgganan and Fuller, 2009). To date, more than 1,000 plant species have been domesticated, and around 200 agronomic and horticultural crops are consumed by people in daily life (Xu, 2010).

In 1865, Gregor Mendel established “Laws of Inheritance" after carrying out hybridization experiments using garden peas. Briefly, he made crosses using different pure lines and observed the ratio of phenotypes obtained within F_1 and F_2 offspring during successive seasons. From these experiments, Mendel drew conclusions that are now referred to as his laws: 1) some alleles mask others (i.e., are dominant); 2) each gene contributes one allele to the gamete; and 3) alleles controlling different traits segregate independently to form different gametes (this law is known not to be true if genes controlling different traits are linked). These findings significantly motivated breeders to utilize

different hybridization strategies to combine various genetic resources to improve crops, which advanced the development of plant breeding.

Shull (1908) proposed the concept of “heterosis”, also named as hybrid vigor later, by making the first single-cross hybrid corn using selected inbred lines; this made a huge contribution to plant breeding methods for open-pollinated species. On the other hand, various plant breeding methods improving self-pollinated plants were developed using hybridization as well. By virtue of these methods, Borlaug and Chang (Khush, 2001), respectively, developed semi-dwarf and high-yield wheat and rice varieties based on incorporating different genetic resources, leading to the so-called “Green Revolution”.
In plant breeding, the exploitation of genetic diversity is not always restricted to the same species or the current gene pools. Stadler (1928) discovered that exposing seeds to radiation increased the mutation rate in barley, and Blakeslee and Avery (1937) proved that chromosome doubling and polyploidy can be induced by colchicine which facilitates the production of new crops by interspecies hybridization.

From the 1980s to now, many molecular tools have been developed to significantly accelerate the breeding process and improve its efficiency, including different types of molecular marker developments and the complete genome sequencing of crops. In addition to the systematic improvements of breeding methods, the plant-breeding community realized that enlarging the germplasm is also crucial for long-term crop improvement. Thus, the International Board for Plant Genetic Resources (IBPGR) (renamed presently as the International Plant Genetic Resources Institute, IPGRI) was established to gather, evaluate and maintain plant genetic resources.

All in all, genetic diversity is the “fuel” of plant breeding. In the first chapter, different sources that shape genetic diversity will be reviewed. For the second chapter, meiotic recombination, one of the important mechanisms driving

genetic improvement in breeding programs, will be described. For the third chapter, I will cover uses of genetic diversity in plant breeding and will present two cases I worked on, namely peanut and tomato.

0.2 Thesis objectives: three projects to investigate genetic diversity questions

As mentioned in the previous paragraphs, plant breeding is a process that improves crops, meeting the needs of humankind. In this process, plant breeders are able to exploit their toolbox to create, evaluate, and manipulate genetic diversity for better designing and developing new varieties.

First, *evaluating* genetic diversity is one of the important processes for breeders to understand their breeding materials after setting the breeding goal, because breeders should have comprehensive ideas about the materials in the germplasm collection before starting breeding programs. The evaluation of genetic diversity can rely on morphological, biochemical, cytological and molecular markers. Thanks to advances in molecular tools, the molecular marker has become a common and efficient tool for evaluating genetic diversity. In addition, identification of candidate genes for target phenotypes of desirable traits can facilitate breeding programs. Breeders often use hybridization to accumulate target phenotypes, and select progenies with desirable traits. For example, when breeders carry out resistance breeding programs, they often use gene pyramiding to combine multiple resistance genes into one genotype for the following selection. Thus, the candidate genes identified by different approaches can indeed help breeders to better design breeding programs. Finally, in any breeding program, breeders select progenies depending on genome reshuffling contributed by meiotic recombination. This specific phenomenon occurring in meiosis generates different combinations of alleles which can be used for the selection. If one can elucidate the underlying mechanism controlling the number and distribution of crossovers, breeders should benefit from this knowledge to better manipulate the genetic variation of their breeding materials, and then select ideotypes with more efficiency.

In my thesis work, I started from modeling meiotic recombination in *Arabidopsis thaliana* to establish a quantitative model that can predict fine-scale structuring of recombination landscapes. Then, I was involved in a project for evaluating peanut genetic diversity in Taiwan. Based on RAD-seq data of 31 accessions, I not only made the diversity analysis but transformed SNPs into a set of KASP markers which can be used for further breeding usage. Finally, I used whole-genome sequencing data of resistant and susceptible tomato lines to identify a candidate resistance gene against bacterial wilting disease, and this result can be also applied in the future breeding work. These three separate projects begin with different aspects but share the goal of developing tools for better evaluating and manipulating genetic diversity for future plant breeding.

1. The forces shaping genetic diversity

1.1 Mutations

A mutation corresponds to having a change in the nucleotide sequence of the DNA defining an organism's genome. Mutations can arise at small-scales: point mutations refers to cases with just one nucleotide being changed, while changes at a small number of nucleotides generally correspond to small insertions or deletions, called indels. Mutations can also arise at larger scales for instance via segmental duplications or they can even induce changes at the chromosomal scale, sometimes leading to changes in the karyotype. Here, I will introduce different types of mutations from small-scale ones to large-scale ones, and then I will explicitly cover the kinds of changes induced by transposable elements (TEs).

1.1.1 Point mutations and small insertions/deletions

If a single nucleotide is altered, inserted or deleted in a DNA sequence, one calls it a point mutation. Point mutations can arise spontaneously, associated or not with DNA replication, and they can also be induced by chemical mutagens or irradiation. For the four common DNA bases, thymine and cytosine are pyrimidines, having a one-ring structure, while adenine and guanine are purines, having a two-ring structure. Point mutations are categorized as "transitions" and "transversions". When a purine is replaced by another purine or a pyrimidine is replaced by another pyrimidine, the mutation is a "transition". On the other hand, if a purine is replaced by a pyrimidine or vice versa, the mutation is a "transversion". Since transition (respectively transversion) is the change of one base to another within the same (respectively different) chemical category, for each base there is only one possible transition while there are two possible transversions. However these mutations are not all equiprobable. For instance, there is a bias in favor of transitions, so the transition/transversion (ts/tv) ratio ranges from 1.02 to 1.68

in plant species (Batley et al., 2003; Kujur et al., 2015; Li et al., 2015; Kanfany et al., 2020).

Mutations corresponding to point substitutions have different impacts on biological function if they arise in coding vs non-coding regions. Nevertheless, mutations located in introns can alter gene expression and thus influence downstream pathways. For example, in *Arabidopsis*, *FLOWERING LOCUS C* (*FLC*) and its antisense transcript *COOLAIR*, a non-coding RNA, coordinate together for the FLC expression that therefore regulates the flowering time. SNP259, located in the intron of *COOLAIR*, is just next to the acceptor splice site, and a T nucleotide in SNP259 of one *COOLAIR* haplotype resulted in the alternative splicing of *COOLAIR* that further increases the FLC expression which is associated with late flowering time (Li et al., 2015). When point mutations occur in exons of genes, they can be either synonymous or nonsynonymous. Synonymous mutations alter the codon but do not affect the associated amino acid, a feature of the genetic code that has degeneracies. On the other hand, nonsynonymous mutations modify codons and lead to different amino acids. Within nonsynonymous mutations, the ones that result in nonfunctional proteins are referred to as missense mutations, while a mutation that leads to a premature stop codon is referred to as a nonsense mutation. An associated example is the *GS3* locus in rice identified by Fan et al. (2006). The gene therein controls rice grain length and encodes a putative transmembrane protein with 232 amino acids. A C-to-A nonsense mutation in the second exon of this gene causes a 178-aa truncation. That mutation is shared by all large-grain varieties, indicating that this mutation has been important in rice domestication.

If a DNA sequence has the addition or removal of one or more nucleotide base pairs, one has an insertion or deletion, respectively. Insertion/deletions (Indels) can result from DNA polymerase slippage during the replication in the presence of repetitive DNA sequences, but they also occur via activity of

transposable elements or via errors arising during meiotic recombination. Because an amino acid is defined by three nucleotides, the reading frame is often shifted by Indels if they occur in a protein-coding gene. That type of mutation is called a frameshift mutation, it will generally lead to a nonfunctional protein. For instance, in maize, Gilles et al. (2017) identified a frameshift mutation caused by a 4-bp insertion in the gene *NOT LIKE DAD* (*NLD*) which leads to a truncated protein. This mutation is used in inducer lines as it is responsible for triggering gynogenesis, a form of asexual reproduction, that is useful for plant breeders to fix allelic combinations. Another case involving a much larger Indel was identified in sorghum *dwarf3* (*dw3*) mutants (Multani et al., 2003). In that study, the authors discovered that there is a 882-bp duplication in the fifth exon among sorghum plants showing the dwarfing phenotype, duplication caused by unequal crossovers. The deletion of this duplication reverted the plants to the tall phenotype.

It is generally assumed that mutations occur randomly and independently of the DNA context, and before other factors like selection or genetic drift that can change the final frequency of genetic variants. In 1952, Joshua Lederberg used the replica plating technique to demonstrate that the streptomycin-resistant mutations in a bacteria population arose before exposing bacteria to the antibiotic. Recently, Monroe et al. (2022) challenged this idea by analyzing *Arabidopsis* mutation-accumulation lines of Arabidopsis produced by single-seed descent for 24 generations without using natural populations that could be confounded by other factors. They found that mutations including Single Nucleotide Polymorphism (SNP) and indels occur half as frequently in gene bodies than in regions outside genes. Furthermore, according to that study, genes with more conserved functions, defined as essential genes, have about one-third reduction of mutation rate compared to non-essential genes.

1.1.2 Large-scale mutations: genetic variation of chromosomal structure and numbers

In nature, there are several types of chromosomal abnormalities associated with chromosomal rearrangements involving large scale deletions, duplications, inversions or translocations (Figure 1). When a chromosome becomes broken at two places, the repair mechanism can lead to an “inversion”. Inversions are sometimes produced by the activity of transposable elements but they can also be induced artificially via irradiation by X-rays or Gamma rays. In addition, inversions can be categorized into two types depending on whether the inverted segment contains or not the centromere. Pericentric inversions contain the centromere, whereas paracentric inversions lack the centromere. In *Arabidopsis*, Zapata et al. (2016) identified 47 large scale inversions when comparing the L*er* and Col-0 assembled genome. Among these variants, the largest one is a 1.2 Mb inversion located on chromosome 4. Meiotic recombination is suppressed in this region, thereby preventing genetic exchanges there between chromosomes with and without the inversion. Those authors thus classified 409 worldwide accessions into two groups with 383 accessions having the L*er*-allele and 26 accessions having the Col-0 allele (Col-0 is thus an outlier when considering these worldwide accessions). Interestingly, when this 1.2-Mb inversion in Col-0 was reverted by the CRISPR/Cas9 system, the local recombination within this region was restored (Schmidt et al., 2020).

A chromosomal translocation occurs if a segment of one chromosome is exchanged with a segment of another (non-homologous) chromosome. If the exchange between two non-homologous chromosomes doesn’t lose any genetic material, this is considered as the reciprocal translocation. Otherwise one says that the translocation is nonreciprocal. Another type of translocation, named Robertsonian translocation, occurs when two acrocentric

chromosomes undergo a reciprocal exchange that leads to one metacentric chromosome and one small chromosome. This genetic phenomenon was used in wheat breeding programs to produce wheat-barley translocation lines with improved traits (Türkösi et al., 2018).

In addition to variation in chromosomal structure, there are cases where the number of chromosomes gets changed. Then, whether or not genetic material is added or lost, one refers to this situation as “aneuploidy”. Aneuploidy mainly results from improper chromosome segregation during meiosis. The normal progression in meiosis I has homologous chromosomes pair, synapse, recombine and then separate for the first division. In meiosis II it is the sister chromosomes that separate for the second division. If any of these steps fail, one can encounter aneuploidies. For instance not all homologs will synapse (asynapsis), or homologous chromosomes may separate prematurely (desynapsis). Both asynapsis and desynapsis can result in univalents that are usually observed in metaphase I, leading to gametes with unbalanced chromosomes that finally result in the creation of aneuploids (Cai & Xu, 2007; Ross et al., 1997). A classic study of aneuploidy was performed using Jimson weed (*Datura stramonium*). Blakeslee (1922) found that *Datura*, a diploid species with 12 pairs of chromosomes, exhibited changed phenotypes when there was an additional copy of any of the chromosomes (these plants were thus trisomic). There are cases of plant breeding exploiting such aneuploidies, in particular for barley and wheat (Türkösi et al., 2016; 2018).

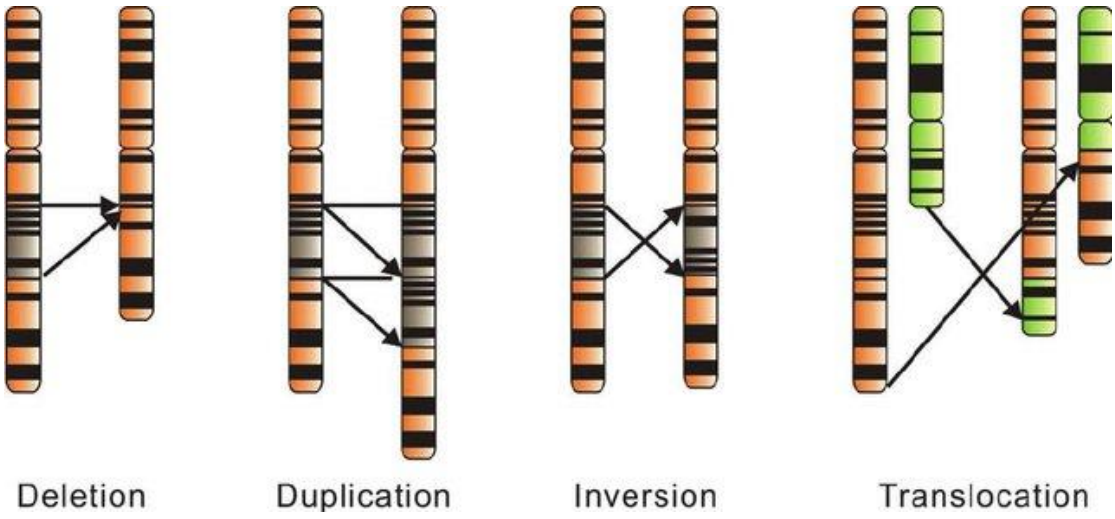


Figure 1. (Adapted from Cai & Xu, 2007) The types of chromosomal abnormalities

1.1.3 Mutations induced by transposable elements

Mutations can be induced by a DNA sequence that moves from one place to another, and even creating different copies through this process; this type of mobile DNA sequence is called a “transposable element” or “TE”. Transposable elements were first discovered by Barbara McClintock (1948) in maize (*Zea mays*). She identified the *Ac*/*Ds* system by studying the relationship between the breakage of chromosome 9 and the changed color of maize kernels. In this system, the *Ac* (*Activator*) factor has to be present with the *Ds* (*Dissociation*) factor in the genome to stimulate the chromosome breakage caused by the *Ds* element. Based on current knowledge of the mechanisms of transposition, transposable elements can be categorized into class I (retrotransposons) operating via a “copy-and-paste” procedure and class II (DNA transposons) operating via a “cut-and-paste” procedure (Wicker et al., 2007). A transposable element that can move by itself (without relying on genes of other TEs) is autonomous, otherwise it is non-autonomous (Wicker et al., 2007).

Class I TEs perform the transposition through an RNA intermediate produced by transcription of one of the TE's copies in the genome. The RNA intermediate is then reverse transcribed and the produced DNA is inserted into the genome, thereby creating a new copy. Retrotransposons can be classified into LTR retrotransposons and non-LTR retrotransposons depending on whether they have long terminal repeat (LTR) elements. LTR retrotransposons, sometimes also called retrovirus-like elements, are evolutionarily closely related to retroviruses. These LTRs are located at the two extremities of the TE. LTR retrotransposons usually have *gag* and *pol* genes that encode a structural protein of the virus capsid and a reverse transcriptase/integrase, respectively. For the non-LTR retrotransposons, there are two main classes: long interspersed elements (LINEs) and short interspersed elements (SINEs). Lacking LTRs, LINEs are autonomous retrotransposons that produce the reverse transcriptase and nuclease for their transposition, and such TEs can reach several kilobases in length. On the contrary, SINEs, ranging from 80 to 500 bp, are non-autonomous, relying on the gene products of LINEs for their transposition. SINEs possess the polymerase III (Pol lll) promoters in their head sequences. Class II TEs, based on "cut and paste", can be divided into two subclasses based on the number of cleaved DNA strands arising during transposition. TIR TEs are distinguished by various lengths of terminal inverted repeats (TIRs) and different sizes of target size duplication (TSD). Interestingly, they can also increase their number by transposition during chromosome replication from one replicated region to another unreplicated one (Greenblatt and Brink, 1962). Different from TIR TEs with double-strand DNA cleavage, Helitron TEs transpose by cutting only one strand through a rolling-circle system without creating TSDs (Wicker et al., 2007).

Transposable elements occupy quite variable proportions of the genomes of plant species. The genome of *A. thaliana*, rice and maize contain about 10%, 20% and 85% of TEs, respectively. Historically, TEs were called "junk DNA"

since they seem not to have obvious functions like protein-coding genes. However, more recent evidence indicates that TEs can often play a role in the regulation of genes and strongly influence chromatin status. The regulation of TEs largely entails their silencing since mobility and even more proliferation of TEs is potentially dangerous for genome integrity. TEs located near or within genes tend to lead to suppression of the expression of those genes (Dubin et al., 2018). Clearly, it is not surprising that a TE inserted within an exon of a gene can produce a loss of function, but even when inserted within introns TEs can disturb the gene function by the altered methylation patterns or by leading to alternative splicings (Ong-Abdullah et al., 2015). For instance, in oil palm, a hypomethylated LINE retrotransposon that resides in the intron of the homeotic gene *DEFICIENS* modifies the splicing, causing abnormal fruits with reduced yield. When a TE is inserted close to the promoter of a gene, it can either disrupt or enhance the transcription rate, affecting more generally the regulation of that gene. In rice, Naito et al. (2009) identified that the DNA transposon *mPing* is generally not inserted within exons, unlike *Tos17* that tends to insert into exons and thereby disrupting those genes (Miyao et al., 2003). In that study, the authors found that the *mPing* insertion leads to the upregulation of 111 out of 710 studied genes with these TE insertions at 1 to 5 kb upstream from the corresponding transcription start site. Furthermore, the *mPing* insertions contribute to the stress inducibility by cold and salt. A more recent study in *Arabidopsis* showed that the insertion of *ONSEN*, a LTR-copia retrotransposon, serves as a promoter and enhancer that specifically activates two adjacent genes under heat stress (Roquis et al. 2021). Moreover, the novel *ONSEN* insertions lead to transcriptional changes including activated/deactivated gene expression, alternative splicing, creation of non-coding RNAs (ncRNAs), antisense transcription and the fused transcript with TEs and genes, suggesting that such novel TE insertions sometimes provide individuals with more complex regulation mechanisms that can be of use for increased resilience to environmental changes.

1.2 Polyploidy

Polyploidy is another type of genetic variation involving a change of chromosome numbers but in contrast to aneuploidy the number change is the same for all chromosomes. Based on the source, polyploids created within one species are referred to as autopolyploids, while polyploids arising from different species are referred to as allopolyploids. Polyploidy occurs more often in plants than animals, and it can naturally result from skipping a cell division step in meiosis or mitosis. It can be artificially induced by chemicals. Consider for instance the case of cell division in meiosis. Normally, meiosis I followed by meiosis II produces four gametes with a haploid set of chromosomes since there is just one round of DNA replication and there are two successive cell divisions. If there is a failure in cell divisions, the process can lead to the production of gametes with unreduced chromosomes, a phenomenon called “meiotic restitution”. Meiotic restitution can be of two types: first division restitution (FDR) or second division restitution (SDR) based on the division that fails. Both SDR and FDR produce two gametes with unreduced chromosomal sets (Figure 2), and are considered as a major source of polyploid production (Ramanna & Jacobsen, 2003; Cai & Xu, 2007).

Polyploidization is of importance for plant evolution, domestication and breeding. Many crops are polyploids, including wheat, potatoes, bananas, cotton and peanuts. Cultivated wheat (2n = 42, AABBDD), *Triticum aestivum*, is a classical example of an allopolyploid that arose without any artificial induction. Initially, the *Triticum urartu* (2n = 14, AA) was pollinated by *Aegilops speltoides* (2n = 14, BB), and underwent natural chromosome doubling to create *Triticum turgidum* (2n = 28, AABB), a progenitor of durum wheat. Then, this AABB allotetraploid was hybridized with wild goat grass (2n = 14, DD), *Aegilops tauschii*, and went through another chromosome doubling to create the present bread wheat (Rosyara et al., 2019).

A breakthrough in producing synthetic polyploids occured when Blakesll and Avery (1937) discovered the potential of colchicine for inducing polyploidy. A landmark synthetic allopolyploid crop is triticale (x *Triticosecale* Wittmack): it is a very successful man-made crop, obtained from the cross between wheat and rye followed by the induction of chromosome doubling by colchicine. There are hexaploid and octoploid triticales which were synthesized by hybridizing hexaploid wheat (*T. aestivum*; 6x = 42) or the tetraploid durum wheat (*T. turgidum*; 4x = 28), respectively, with the cultivated diploid rye (*Secale cereale* L.; 2x = 14). In addition, polyploids can not only be direct targets for creating genetic variation, but can also form a bridge for transferring genetic material between two species, a process called bridge crossing (Dewey, 1980).

From the perspective of evolution, polyploidization often leads to transgressive phenotypes and vigor superior to that of their diploid progenitors (Van de Peer et al., 2009). In general, these extra chromosome sets in polyploids lead to increased cell size and thus larger organs, an advantage that is selected for in plant breeding (Alix et al., 2017). The multiplication of chromosomes also produces “genome redundancy” that will buffer against deleterious alleles (Soltis and Soltis, 2000). Interestingly, the transcriptomic levels in polyploid species don’t follow the ploidy change (Song et al., 2020). Presently, even though it is not fully understood how progenitor genomes precisely shape the molecular mechanisms of polyploid individuals, the genomic, transcriptomic and epigenomic changes brought *via* polyploidization provide potential heterosis, e.g. for yield or for stress resistance (Sattler et al., 2016; Van de Peer et al., 2021). For example, the allotetraploid obtained from the cross between *A. thaliana* and *A. arenosa* epigenetically represses *CIRCADIAN CLOCK ASSOCIATED 1* (*CCA1*) and *LATE ELONGATED HYPOCOTYL* (*LHY*), producing more chlorophyll and starch than its parents (Ni et al., 2009). Wang et al. (2018) studied the 3D genome architectures of diploid and tetraploid cotton, and found that allopolyploidization affected the

switching of A/B compartments (A and B compartments refers to open and closed chromatin, respectively) and led to the reorganization of topologically associated domains (TADs), with corresponding greater complexity of transcriptional regulation.

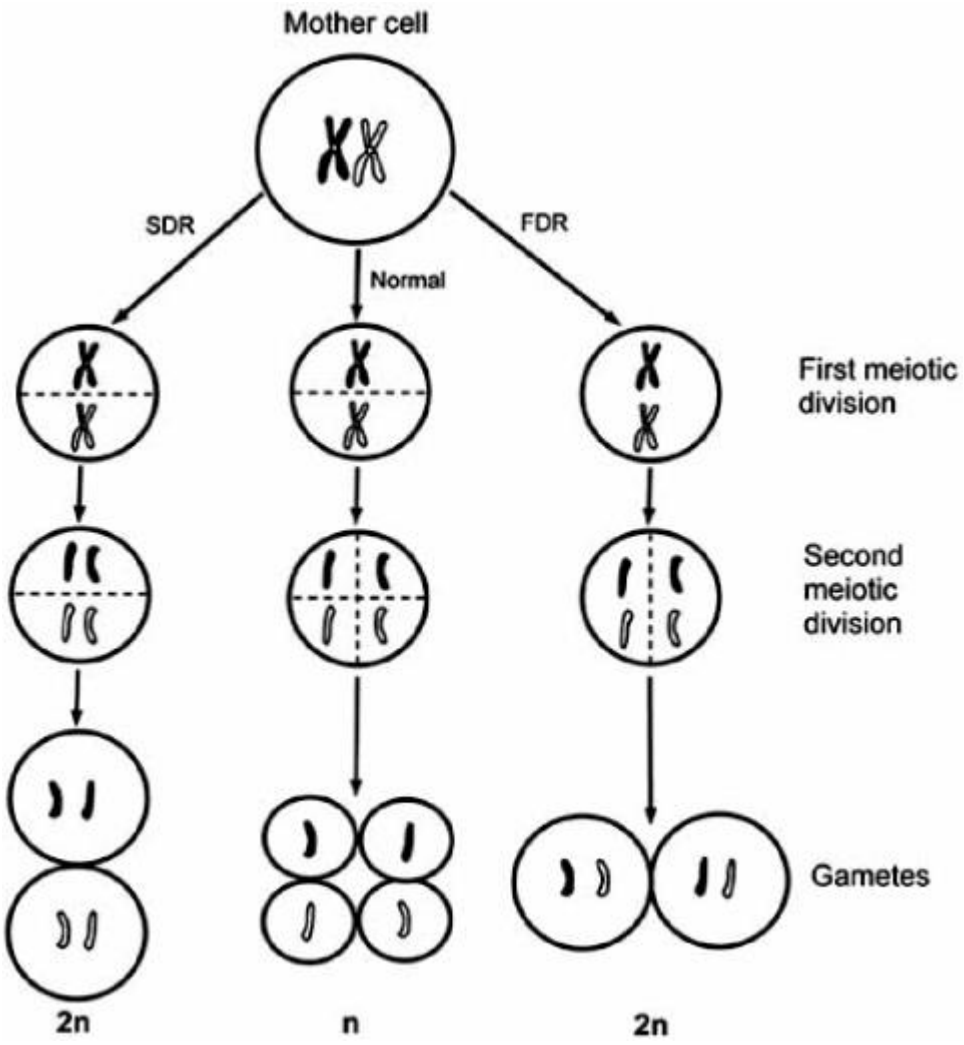


Figure 2. (Adapted from Cai & Xu, 2007) The comparison between normal division and abnormal division at the first or second division during meiosis. In this diagram, the middle shows the normal successive divisions that lead to 4 haploid gametes. Both first and second division restitution (FDR/SDR) give rise to unreduced gametes.

1.3 Genetic variation produced by meiotic recombination

Meiosis is a specialized process in gametogenesis that produces one round of DNA replication and two successive rounds of cell division, which means that meiosis allows organisms to produce gametes with half the number of chromosomes compared to their somatic cells. This process plays a key role in the maintenance of chromosome numbers among species since the fertilisation phase of sexual reproduction combines two gametes and thus doubles the number of chromosomes. During meiosis, there is a phenomenon referred to as meiotic recombination, which can produce “crossing overs”, which allow genetic reciprocal exchanges between homologous chromosomes. The detailed mechanisms operating during meiosis and driving meiotic recombination will be described in chapter 2, this section will mainly introduce the genetic variation brought about by meiotic recombination.

Meiotic recombination has a profound effect on genetic variation. It occurs at the prophase I of meiosis, and allows the genetic exchange of alleles between homologous chromosomes. Because this process drives genome reshuffling and produces new combinations of different alleles, it is recognized as the heart of crop selection programs (Wijnker & de Jong, 2008). Many plant breeding selection methods rely on such random recombination to reshuffle alleles and produce variation to be selected for. For example, mass selection, a plant breeding method for open pollinated plants, is based on random matings between adult individuals in the field. Breeding can also be better controlled by selecting elite individuals for the next round of random mating until the goal of the breeding program is reached. The spirit of breeding is to use the genome reshuffling provided by meiotic recombination to combine beneficial alleles together. If we take backcross selection as another example, breeders use it to introduce one or two genes of interest from a donor parent into a recurrent parent, leading after sufficient number of generations to a line with most of the genetic background being from the recurrent parent but also

including regions containing the desired genetic material from the donor parent. In the process of backcross selection, the recurrent parent is used at each generation and crossed with the latest progeny, the goal being to reduce as much as possible the contribution of the donor except for the region of interest to be introgressed. Nowadays, although breeders are able to use molecular markers to precisely pinpoint the genetic content of progenies, one still has to rely on the random recombination events produced by meiosis to purge most of the donor genome.

In addition to being used in breeding programs, meiotic recombination allows geneticists to produce different ressources key for their genetic studies. Positional cloning is a fundamental approach in forward genetics. This method starts from a phenotype of interest and aims to identify specific regions or even genes responsible for that phenotype, exploiting different kinds of mapping populations. In terms of the number of parents used, mapping populations can be classified into biparental and multiparental mapping populations. For biparental mapping populations in plants, one can for instance start from the cross between one parent with the phenotype of interest (eg. the one with the resistance) and another without that phenotype (eg. the one with susceptibility), and make a F_2 population from selfing F_1 individuals. Within a F_2 population, each individual is made up of different combinations and fractions of the biparental genetic backgrounds because of meiotic recombination. Then, this F_2 population can be used for conducting QTL mapping for instance, to delimit intervals where allelic variation is associated with the phenotypic variation. Although QTL mapping is a powerful tool to dissect genetic variation, there are limitations to the mapping, resolution being generally limited by the number of recombinations or by the level of genetic diversity in the region of interest, especially in biparental populations. The first limitation can be overcome by having larger populations. Note that it is possible to include more rounds of recombination (going beyond F_2 individuals) but if this is done only by selfing the gain is modest because

homozygosity sets in rapidly. For the second limitation, geneticists have come to utilize natural populations rather than controlled crosses so as to have a broader phenotypic diversity. In such a context one can perform genome-wide association studies to identify intervals linked to the target phenotype (Korte & Farlow, 2013) without having to generate any crosses. In natural populations, since they have gone through many rounds of meiotic recombination, past recombination events are dense and thus allow for much higher resolution than F_2 populations. Another type of population that exploits meiotic recombination is the Multiparent Advanced Generation Inter-Cross (MAGIC) one that aggregates the genetic variation from multiple parents and produces associated recombinant inbred lines. Depending on the number of chosen founder lines, several generations of intercrossing are conducted for combining the genetic background of founder lines together. Then, individuals corresponding to mixtures of the founder lines will be used to produce recombinant inbred lines by selfing, resulting in a MAGIC population. This artificial multiparental population has higher genetic diversity than biparental populations and it solves the problem of population structure arising in the natural population that tends to confound GWAS approaches (Scott et al., 2020). Taken together, all of the methods mentioned above rely on meiotic recombination to provide novel combinations of different alleles; these resources can be used to better understand the relationship between genes and phenotypes and can also be exploited for crop improvement.

During meiosis, homologous chromosomes may pair yet be misaligned. If such misaligned regions have high sequence identity, crossovers may form between them corresponding to “unequal” crossing over (Cai & Xu, 2007). Unequal crossing over generates segmental duplications and deletions and is considered as an important factor that influences genome evolution and variation. The requirement for high sequence identity may seem quite limiting but in practice transposable elements come in many copies and so generate such situations quite frequently. Unequal crossing over can take place in both

intragenic and intergenic regions. In plants, a number of studies have shown that disease resistance genes are often organized in clusters with many similar copies. That is exactly the expected signature of unequal crossing overs, favored by gene families. For resistance loci, the novel haplotypes or combinations of such genes helps to keep up with fast evolving pathogen populations. A good example of this occurs in maize: the *Rp1* region is a classic example of the result of unequal crossing overs. This complex locus is located in the distal end of the short arm of chromosome 10 and is responsible for acquisition of resistance to the fungus *Puccinia sorghi*. Even though *Rp1* is a dominant resistant locus, Bennetzen et al. (1988) surprisingly found the presence of susceptible progenies from the test cross between *Rp1* homozygous lines and a *rp1*/*rp1* line, probably because of the meiotic instability that led to the *Rp1* inactivation. By studying the abnormal exchange flanking markers of *Rp1* homologs, Sudupak et al. (1993) concluded that meiotic instability resulted from unequal crossing over. Furthermore, this phenomenon was also proved based on BACs from B73 for the sequence analysis of *Rp1* homologs that showed the chimeric structure within genes (Ramakrishna et al., 2002). In tomato (*Lycopersicon esculentum*), several *Cladosporium fulvum* resistance genes were identified, including *Cf-4* and *Cf-9* that were introgressed from *L. hirsutum* and *L. pimpinellifolium*, respectively, into cultivated tomato. These two genes are located in a 36-kb region, and sequence analysis suggests that *Cf-4* and *Cf-9* are probably derived from a common gene, in line with what is expected if unequal crossing over shaped that region. Interestingly, the unequal crossing over taking place in the associated intergenic intervals generated recombinants with different resistance specificities (Thomas et al., 1997).

Unequal crossing over also shapes repeated sequences in the genomes on different scales. TEs arise in many copies and are thus considered repeated sequences; they almost always occupy a large fraction of the genomes for many plants. Among TEs, retrotransposons contribute most to genome

expansion. However, unequal crossing over can act both ways, increasing or decreasing the number of copies. In rice, 11 families, 1,000 events in total, of LTR retrotransposons were investigated for their sequence structures. The result showed that more than 75% of elements are solo LTRs and fragmented, and that this was driven by unequal crossing over and illegitimate recombination (recombination arising from chromosomes not sharing homology). And these two variant forms of recombination accounted for the removal of more than 190 Mb of LTR-retrotransposon DNA over the past 8 million years (Ma et al., 2004). Note that simple sequence repeats (SSR), also named as microsatellites, are another type of repeated sequence. Different studies indicate that unequal crossing over could be one of the mechanisms that creates novel SSR loci by gaining or deleting repeats (Innan et al., 1997; Oliveira et al., 2006).

1.4 Genetic variation driven by changes in allele frequencies

In previous sections, I concentrated on different sources driving changes in the genome. In this section, I will focus on sources of change in allele frequencies that produce genetic diversity from a population genetics perspective.

Before introducing the forces that can change allele frequencies in populations, an important principle should be described beforehand, that is, the Hardy-Weinberg principle. In 1908, Hardy and Weinberg proposed that the frequencies of genotypes in a random-mating population can be predicted by the allele frequencies whenever mutation, random genetic drift, natural selection and migration can be ignored. Furthermore, the frequencies of both genotypes and alleles of a population will remain the same through generations if not disturbed by those other factors, so this principle is also referred to as the Hardy-Weinberg equilibrium. In a population satisfying Hardy-Weinberg equilibrium, suppose a gene has two alleles "*A"* and "*a"* of frequencies *p* and *q*, respectively. The frequencies of the (diploid) genotypes *AA*, *Aa*, *aa* are then predicted as p^2 (*p* x *p*), 2*pq* (2 x *p* x *q*) and q^2 (*q* x *q*). However, populations in natural environments exhibit rarely the Hardy-Weinberg equilibrium because of a number of factors that we now cover and that drive evolutionary changes and shape the genetic diversity among or between species.

1.4.1 Genetic drift

The production of gametes is associated with a *sampling* of alleles between homologous chromosomes. As a result, the allelic content of offspring is stochastic (one can say that every offspring is unique). This stochasticity in allelic frequencies extends to populations, an extreme case arising when the population is of size two and produces two progenies. Assume there is neither selection nor mutation and that at the considered locus both parents are heterozygous (genotypes Cc), so the allele frequencies for C and c at the parental generation are 0.5. However, following the Mendelian independent assortment, the probability of recovering those same allele frequencies in the two progeny is only 6/16, indicating that allele frequencies will change in 10 out 16 random realizations when going from one generation to the next. This is a general phenomenon independent of the initial frequencies. For instance if one assumes that the initial allele frequencies of C and c are 0.75 and 0.25, respectively, the probability of maintaining those frequencies in the two offspring is now 4/16. There is also a non-zero probability that one allele will be completely lost amongst the progenies (Figure 3A). Having a larger population will reduce the size of the fluctuations but will not remove them, so one concludes that allele frequencies typically change from one generation to the next due to random sampling. Over multiple generations, such fluctuations can lead to allele fixation, thus changing quite fundamentally the genetic makeup of the population.

The stochastic behavior of allelic frequencies is referred to as “genetic drift”. That “force” plays a more important role in populations of small-sizes: as the population size becomes large the relative size of the fluctuations decrease and the allele frequencies depart less and less from their mean. Wright (1931) studied genetic drift and showed that the frequency of heterozygotes (Cc in our previous example) denoted as “*H*” tends to decrease in a finite population. He quantified this effect and proposed the mathematical formula $H_{t+1} = H_t$ (1 -

$1/(2N_t)$) where H_t and N_t are the degree of heterozygosity and the number of individuals at generation t, respectively, and H_{t+1} is the *average* degree of heterozygosity at generation t+1. This equation thus describes the average effect of genetic drift, not the actual effect (which is stochastic). But clearly smaller populations are more sensitive to genetic drift. To illustrate this, assume that in the initial population the two alleles have equal frequencies. If the population size is 16, H will typically drop below 0.1 after 50 generations while if the population size is 1024, H is expected to remain above 0.4 for over 200 generations (Figure 3B).

From Wright's formula, we can say that the strength of genetic drift is inversely proportional to the population size. Note that the population here is an idealized and panmictic population. This type of population is also called a Wright-Fisher population: all of its individuals have equal probability to act as parents during reproduction. However, this idealized population is unrealistic in the real world since there are often factors that make populations violate the assumption of the idealized one, such as the occurrence of mutation, migration, and preferences in matings. Thus, a concept of "effective population size" (denoted as N_e) has been introduced whereby Wright's equation still describes the effect of drift in such modified populations if one replaces the actual population size by N_e. This suggestion has a long history. Indeed, in 1931, Ronald Fisher and Sewall Wright defined N_e as "the number of breeding individuals in an idealized population that would show the same amount of dispersion of allele frequencies under random genetic drift or the same amount of inbreeding as the population under consideration". Since the magnitude of genetic drift depends on N_e and population size affects the maintenance of genetic diversity, N_e can be also interpreted as the size of an idealized population with the same genetic diversity as the population of interest. There is an estimator, named as Watterson estimator, that predicts the genetic diversity of a random-mating population based on the combination of N_e, the mutation rate per site per generation (μ) and the scaling factor

depending on ploidy. While considering diploid organisms, this estimator equals $4N_e\mu$ (Ellegren & Galtier, 2016).

The concept of effective population size is often utilized to understand the history of crop domestication. The crop domestication is a process whereby wild ancestors undergo selection by humans, then become the current crops. Through this long process, descendents of wild ancestors acquire morphological modifications to fit various human requirements. For example, several phenotypes (traits) are considered as major domestication targets, including the size of grains (fruits), seed shattering and dormancy, and the plant architecture (of use for large-scale cultivation). Gaut et al. (2018) classified domestication into four stages. Stage 1 starts from the wild ancestors with substantial genetic diversity during which ancient people began to manage these ancestors that could somewhat influence the genetic diversity of wild ancestors. Stage 2 corresponds to initiating cultivation with purpose; ancestors of current domesticated crops often suffer from genetic bottlenecks that significantly modify the allele frequency of domestication genes and decrease the genetic diversity and effective population size. Stage 3 refers to the expansive domestication in more places, so this process can lead to various adaptations in different environments. Stage 4 corresponds to having organized breeding programs for these domesticated crops (Figure 4). Even though not all domesticated plants exhibit the same trend as shown in Figure 4, cereal crops frequently have more noticeable genetic bottlenecks (more drastically reduced population sizes) than perennial crops. In addition, even though stage 2 is considered as generating an abrupt reduction of N_e, some studies based on different demographic inference approaches showed that there are probably protracted N_e declinations during stage 1 due to stresses and human management (Gaut et al., 2018). In addition, the drop of N_e during stage 1 and 2, leading to an increased genetic drift, can eventually result in the accumulation of deleterious mutations at high frequencies, a phenomenon referred to as “the cost of domestication”. Note that in

populations of large N_e such mutations can be more easily removed by selection and so genetic drift doesn't act much in such situations. On the contrary, once N_e is small, the frequency of these deleterious mutations can rise substantially because of the larger magnitude of genetic drift and less effective selection pressures (Gaut et al., 2018; Moyer et al., 2018).

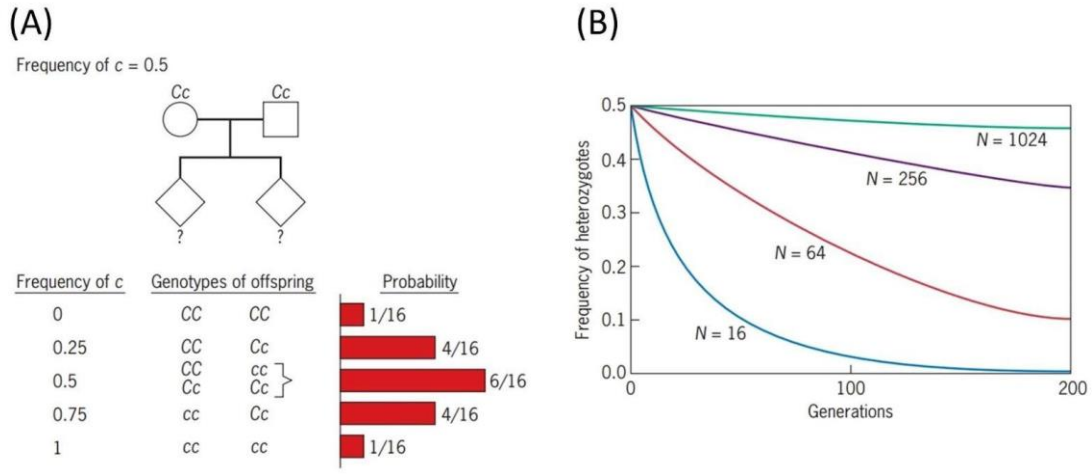


Figure 3. (Adapted from Principles of Genetics, 6th edition) Schematic illustrations for genetic drift. (A) The probability of the frequency of allele c from two offspring produced by two parents both having the Cc heterozygous genotype. (B) Assuming the same allele frequency of 0.5 for both alleles within the initial population, the plot shows the decrease in average H (the frequency of heterozygotes) due to genetic drift as predicted by Wright's formula for different population sizes.

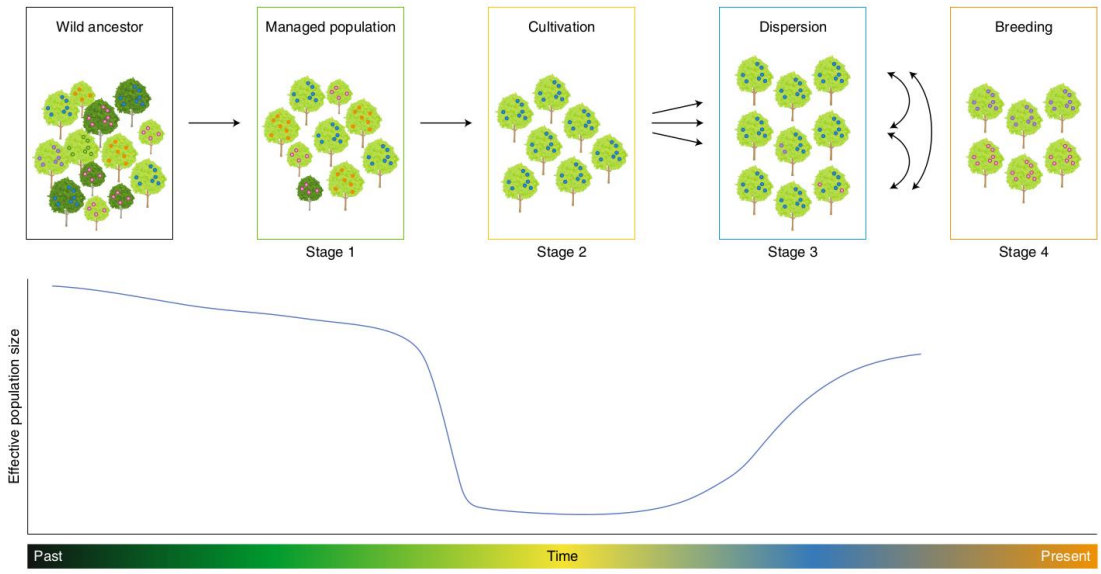


Figure 4. (Adapted from Gaut et al., 2018) The domestication processes decomposed into four successive stages. On the top, the figure shows the progression along four stages starting from the wild ancestors. On the bottom, the curve represents the effective population size as a function of time. In the beginning, the wild ancestor has a larger effective population size than the following stages, which means that it has considerable genetic diversity. During stage 1 (Human management), the genetic diversity may be somewhat modified due to the non random management of particular genotypes. The effective population size reduces significantly from stage 1 to stage 2 leading to genetic bottlenecks that increase the strength of genetic drift. In parallel, frequencies of interesting alleles increase. Stage 3 and 4 represent the diffusion of domestication and the systematic use of breeding programs. The first increases effective population size via the expansion of the regions of cultivation, allowing domesticated genotypes to shape their genetic makeup to be better adapted to different environments.

1.4.2 Natural selection

In 1859, Charles Darwin published his theory of evolution founded on “natural selection”. He pointed out that inevitable selection by the environment will act on the heritable variation within populations. Through this process, the variation best adapted to the environment will be transmitted to successive generations, while variations not fitting that environment will be removed. In other words, individuals with advantageous phenotypes will have larger survival and reproduction rates in the given environment, and they will be gradually prevalent in populations through many generations if those phenotypes are partly heritable. This process leads to the characterization and evolution of distinct populations among species. Furthermore, once populations within species accumulate enough variations, they may see drops in cross fertility rates and associated formation of new species.

Geneticists often refer to the ability to survive and reproduce as “fitness”. Let’s illustrate how fitness can influence allele frequencies at one locus. In our example we consider an insect species that mates at random and whose color is controlled by a locus with two alleles, “*A”* and “***a”***. The allele “*A”* is dominant and leads to dark grey individuals while allele “a” is recessive and leads to light gray individuals. In a forest-type environment, the trees provide *AA* or *Aa* insects better protection since the dark form, similar to the color of the trunk, camouflages them from predators, while *aa* insects are more easily seen and thus subject to predation. However, in an open-field environment the opposite situation arises, *AA*/*Aa* insects are more easily seen than *aa* insects. Thus, the fitness of each genotype depends on the environment. Let us define the relative fitness by comparing fitness to that of the advantageous genotype. The relative fitness of the advantageous genotype is then 1 and we denote by 1-s the relative fitness of the disadvantageous genotype. This parameter *s* is named as “selection coefficient”, referring to the intensity of natural selection for eliminating such genotypes in the population. When *s* is

large, it indicates that natural selection removes that genotype more strongly. Here is a table of fitnesses for the three genotypes of insects in their two habitats for our example:

Genotype	*AA*	*Aa*	*aa*
Phenotype	dark grey	dark grey	light grey
relative fitness in forest	1	1	$1 - s_1$
relative fitness in field	$1 - s_2$	$1 - s_2$	1

Even though the relative fitness doesn't give the absolute reproduction rate of the three genotypes, we can still know how natural selection influences the weaker phenotypes according to the value of *s*. If s_1 is 1, it means the *aa* genotype is completely lethal. If s_1 is 0.1, it means natural selection slightly reduces the frequency of the *aa* genotype, leading to extinction of that genotype through sufficiently many generations. Here is the table showing the genetic makeup and the contribution to the next generation of the three genotypes living with a forest (the initial allele frequencies of *A* and *a* are taken as 0.5):

Genotype	*AA*	*Aa*	*aa*
Phenotype	dark gray	dark gray	light gray
Relative fitness in forest	1	1	1 - 0.1 = 0.9
Frequency (before selection)	0.25	0.5	0.25
Relative contribution to the next generation	0.25	0.5	0.25 x 0.9 = 0.225

Proportional contribution to the next generation	0.256	0.513	0.231

Initially, *AA*, *Aa* and *aa* genotypes occupy 25%, 50% and 25% of the total insect population, respectively, in this forest. However, 10% of *aa*-genotype individuals are removed by natural selection (predators). Thus, the relative contributions of *AA*, *Aa* and *aa* genotypes to the next generation are 0.25, 0.5 and 0.225, respectively, summing up to 0.975. Then, the normalization of these relative contributions shows that the three genotypes now have frequencies 0.256 (*AA*), 0.513 (*Aa*) and 0.231 (*aa*) and so the allele frequencies of *A* and *a* become 0.513 and 0.487, respectively. Through many generations in such a framework, the frequency of *a* will decrease and eventually reach 0. In effect natural selection can drive the fixation of different alleles given various environments, thereby shaping allelic characteristics of populations belonging to a given species.

Our example corresponds to a case of negative or purifying selection that removes disadvantageous alleles from populations which can lead to the fixation of particular genotypes. On the other hand, natural selection can also act on the maintenance of genetic diversity if the heterozygous genotype has higher fitness than the homozygous ones. That situation is referred to as balancing selection. It arises in particular in the case of two alleles of the hemoglobin gene, denoted as HBB^S and HBB^A. People having the homozygous HBB^SHBB^S genotype suffer from sickle cell anemia due to the damaged form of red blood cells. However, people with the homozygous HBB^AHBB^A genotype have the normal red blood cells. Interestingly, it was discovered in West Africa that people having the heterozygous HBB^SHBB^A genotype are more resistant to the parasites leading to malaria than people with the homozygous HBB^AHBB^A genotype. Thus, even though the mutant HBB^S allele leads to a very serious (sometimes lethal) disease in the homozygous context, that allele is nevertheless maintained in the population

because of its beneficial effect in the heterozygote, a direct illustration of balancing selection.

So far we have only considered single locus situations because of the underlying simplicity. Generally when selection operates in the context of multiple loci the mathematics is far more complex. Nevertheless a case that is easily understood qualitatively is that of “genetic hitch-hiking”. This situation occurs when the selection acts on one locus and thereby drags along with it selection in the flanking regions, even if those regions themselves do not contribute to fitness. Over generations, the selection at that central locus shapes the haplotypes in its neighborhood (Ellegren & Galtier, 2016). For genetic hitch-hiking, there are two key factors influencing the genetic variation produced near the region under selection; the intensity of selection there and the local recombination rate. The higher the selection pressure, the less genetic variation will be maintained in this region. Across generations, the flanking regions will be shuffled by recombination, and the genetic variation at a given distance will be larger the more recombination events can arise there (Ellegren & Galtier, 2016). Corbett-Detig et al. (2015) utilized genomic resequencing data of 40 animal and plant species to define the relationship between natural selection and recombination rate that influences neutral genetic diversity (meaning no effect on fitness). Specifically, they used body sizes and species as proxies for census population size (Nc), and proved that natural selection eliminated more neutral genetic variation of species with larger Nc than the ones with small Nc. Finally, they concluded that neutral genetic diversity among species can be reduced by natural selection.

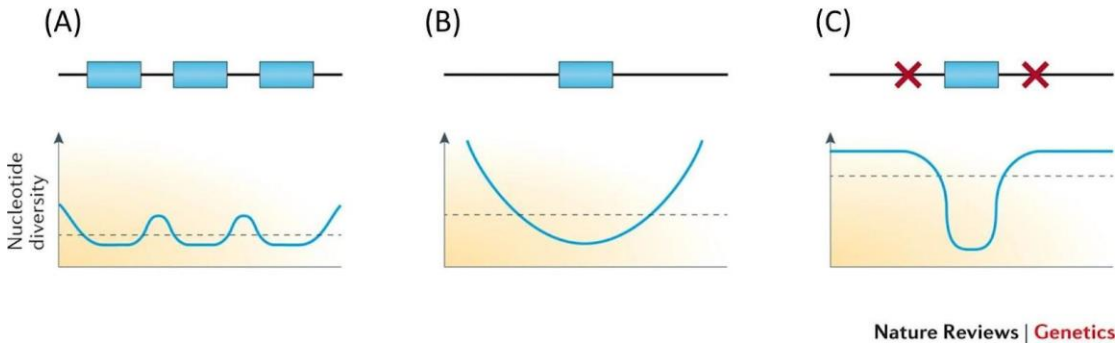


Figure 5. (Adapted from Ellegren & Galtier, 2016) The genomic distribution of a population's genetic diversity is shaped by the combination of selection and recombination. When a region containing one or multiple genes is under selection, the genetic diversity in this region behaves differently according to the selection pressure and the recombination rate in that region. (A) Case where the region has multiple genes under selection, so the linked selection acts heavily on the whole region, severely reducing the genetic diversity there. (B) Case where the region has only one locus under selection, so the genetic diversity in flanking regions increases as one moves away from the locus under selection. (C) Case where the recombination rate is high in the two flanking regions of the selection target. Then linked selection will be broken (there is no genetic hitch-hiking effect), and the genetic diversity of the flanking regions of the selection target can be maintained.

1.4.3 Gene flow

Another factor that alters the allele frequencies is gene flow, also named as gene migration. Gene flow is the transfer of genetic material from one population to another, typically arising because of migrating individuals. Since gene flow brings new genetic content to a population, it can reduce the genetic divergence between populations that might have been established through accumulation of mutations, selection or genetic drift. Gene flow highly depends on the mobility of organisms, so animals might be expected to have higher degree of gene flow than plants. However, plants are quite subject to gene flow because of the way pollination arises and also because seeds are subject to much dispersal (e.g. via transport by wind, water and animals). The extent of gene flow can be measured by genetic differentiation (Fst). If a collection of demes (sub-populations typically separated geographically) has no gene flow, Fst will be 1. On the other hand, when there is significant migration per generation among the demes, Fst will reduce significantly, indicating that there is a lot of shared genetic variation between the demes. Populations with limited gene flow will more easily diverge from other populations, sometimes leading to speciation. In plants, different forces influence gene flow. First, outcrossing plants lead to more gene flow than selfing plants, so outcrossing plants tend to have little population structure while selfing plants tend to fix different variants in separate populations (Wright et al., 2008). In addition, the architecture of plants and the way their seeds can disperse will also influence population differentiation. It was demonstrated that outcrossing trees pollinated by wind have lower Fst values than mixed-mating (the combination of selfing and outcrossing) and non-woody plant species pollinated by insects, indicating that outcrossing trees, generally having greater longevity and sizes than non-woody plant species, are more effective to connect within and between subpopulations by gene flow. (Gamba & Muchhala, 2020).

2. Meiosis and crossover formation

2.1 Phases and stages in meiosis

Meiosis begins in a mother cell after its S phase (DNA replication) and is divided into meiosis I and meiosis II, both leading to a cell division and thus to a reduction of ploidy. Both meiosis I and meiosis II involve four stages: prophase I/II, metaphase I/II, anaphase I/II and telophase I/II. Among these, prophase I is the most complex and time-consuming. According to a previous study (Bennett, 1971), the duration of prophase I for 12 different plant species occupies 50.0% to 89.6% of the total duration of meiosis.

Prophase I is further divided into five stages. The earliest one is called “leptotene”, referring to “leptonema” that is “thin threads” in Greek. Because of the replication that arose just prior to meiosis, each chromosome is attached to an identical sister and so these are referred to as “sister chromatids”. During leptotene these chromosomes begin to condense and form threads that can be seen under the microscope. Then, one enters the second stage of prophase I, the zygotene stage, referring to zygonema or “paired threads” in Greek. Indeed, during this stage, homologous chromosomes start to get close and to become paired and ultimately undergo “synapsis” where the elements of the pairs are no longer visible as separate entities in standard microscopy. Synapsis is generally facilitated by a structure called “synaptonemal complex” between two paired chromosomes. The synaptonemal complex (SC) consists of two lateral elements associated, one for each of chromosomes and one central element sandwiched by two lateral elements, and the SC is thought to not only mediate synapsis but also the crossover formation in eukaryotes (Carpenter, 1975). However, it was discovered that the SC is not a prerequisite for the formation of crossovers (Storlazzi et al., 1996). As the synapsis progresses, the paired homologous chromosomes become thicker, leading to the pachytene stage, pachynema corresponding to “thick threads" in Greek. In such paired homologs, since each side has two sister chromatids, one refers to this structure as a “bivalent”. Note that it contains 4 chromatids

(DNA molecules). At this stage, the homologous chromosomes are fully paired, and some of them start to make bridges that can lead to crossing over. The mechanism of crossing over, responsible for meiotic recombination, is a process that starts with the formation of double strand breaks (DSBs) in leptonema, and ends in repairing those DSBs either by crossovers or by non crossovers (typically gene conversions). By the end of the pachytene stage, crossovers are finalized. The molecular details of these different steps are quite complex, but if we consider only the end result we note that meiotic recombination is an important source of genetic variation since it facilitates the genome reshuffling by exchanging genetic materials of homologous chromosomes. After the pachytene stage, the next one is the diplotene stage, diplonema ("two threads" in Greek). In this stage, the SC complex starts to be pulled apart, and the homologous chromosomes separate, except at the places where crossover events have been produced. Each crossover will lead to what is called a “chiasma” (plural is “chiasmata”) where the two homologs remain in contact, locally forming an “X” (thus the name chiasma). Then, one enters the final stage of prophase I, diakinesis ("movement through" in Greek), during which the chromosomes condense still further and the chiasmata become particularly clear. In this stage, the chiasmata are the only attachment between homologues. Finally, when the nucleolus disappears, the nuclear membrane begins to disintegrate and the spindle apparatus forms, one has reached the end of diakinesis.

During metaphase I, the nuclear membrane completely disappears and spindle microtubules, oriented perpendicularly to the chromosomes, attach to the kinetochores on each side of the bivalents and then drive those bivalents to migrate to the spindle equator. In contrast to the situation arising in mitosis, where sister chromatids have oppositely oriented kinetochores and thus are attached by spindle microtubules to both poles, in meiosis I the sister chromatids of one homologous chromosome have their kinetochores pointing in the same direction and so attach via the spindle microtubules to just one

pole. The dynamics of the system is such that after trials and errors the two homologous chromosomes become attached via spindle microtubules to opposite poles. That property is key for separating the homologues to opposite poles during anaphase, otherwise aneuploidies will arise. Since these orientations are random, the separation of homologs will give at each pole a mixture of paternal and maternal chromosomes. The random orientation of bivalents is also the basis of the independent assortment of chromosomes. During anaphase I, the paired chromosomes separate from each other, migrating toward opposite spindle poles, mediated by the shortening of the microtubules that remain attached to the kinetochores. When the two sets of chromosomes arrive at their respective spindle poles, telophase I begins. During this stage, the spindle apparatus is taken apart, and the nuclear envelopes appear again around each set of chromosomes. Telophase I is followed by cytokinesis that produces two daughter cells containing chromosomes consisting of two chromatids that are no longer identical because crossovers have led to exchange of material between homologs.

Meiosis II is the second cell division within meiosis. It largely resembles mitotic division even though creating a significantly different result (haploid gametes). In prophase II, the nuclear envelope and nucleoli disappear, and chromosomes begin to condense again. In addition, centrosomes move to opposite poles, and the spindle apparatus is set up for the next stage. In metaphase II, chromosomes are aligned in the spindle equator, and the two kinetochores acquired by centromeres of each chromatid are attached by spindle microtubules from two opposite spindle poles, so this time the two kinetochores face different poles. Then, these chromatids are separated and migrate toward opposite poles during anaphase II. In the end, telophase II, similar to telophase I, leads to disassembling the spindle microtubules, nuclear envelope formation, and finally one obtains four haploid daughter cells, each with a complete set of chromosomes.

2.2 The steps leading to meiotic recombination

2.2.1 The formation of DNA double-strand breaks

As the initiation step of meiotic recombination, DNA double-strand breaks (DSBs) are induced during prophase I of meiosis. This evolutionarily conserved process involves multiple proteins (Bergerat et al., 1997; Keeney et al., 1997). The Spo11 protein, homologous to the A subunit of the type II topoisomerase (TopoVIA) from the archaeon *Sulfolobus shibatae*, catalyzes DSBs (Bergerat et al., 1997). It consists of two domains including a DNA-binding core having a winged-helix domain (WHD) and a TOPRIM domain found in various topoisomerases and primases. The Spo11 protein forms a transient covalent bond between itself and DNA via one of its tyrosines. This tyrosine is strongly conserved in Spo11 orthologs and among TopoVIA, across many different species (Bergerat et al., 1997; Cervantes et al., 2001; Hartung et al., 2007; Malik et al., 2007). The endonucleolytic cleavage catalytic activity leads to the resection of the DNA strand bound by Spo11 protein. In *S. cerevisiae*, these proteins are the MRX complex (Mre11, Rad50, and Xrs2) and Sae2. In other species, they are the MRN complex (Mre11, Rad50, and Nbs1) and CTIP. The two 5' strand ends are then resected by 5′ to 3′ exonucleases (de Massy, 2013) (Figure 6). In *S. cerevisiae*, Exo1 and Mre11 also have the 5′ to 3′ and 3′ to 5′ exonuclease activity, respectively, that can perform the strand resection from 5' or 3' end (Garcia et al., 2011). Multiple *Spo11* paralogs have been found within species in different cases. In mice, *Spo11α* and *Spo11β* are two major isoforms that probably have distinct functions, and the partially fertile phenotype acquired by the expression of only *Spo11β* suggests that *Spo11α* possibly regulates the formation of late-forming DSBs (Kauppi et al., 2011). In *A. thaliana*, two of three *Spo11* paralogs, *Spo11-1* and *Spo11-2*, are involved in forming DSBs (Grelon et al., 2001; Hartung et al., 2007; Stacey et al., 2006). In *O. sativa*, a number of Spo11 paralogs were identified, and *OsSpo11-1* and *OsSpo11-4* are

necessary for meiosis to progress to completion (An et al., 2011; Yu et al., 2010).

2.2.2 Homology search

After the resection by the exonuclease, the resected strands are further protected by the Rpa protein. Then, proteins of the RecA family replace the Rpa protein for forming nucleofilaments that can catalyze the search for a homologous sequence on another chromosome and produce a heteroduplex for repair and then exchange of DNA molecules (Figure 6). In *S. cerevisiae*, The RecA protein family includes Rad51 and Dmc1 recombinase, sharing 54% and 45% amino acid identity with humans (Masson & West, 2001). Furthermore, mutation analyses showed that the Rad51 recombination complex is independent of Dmc1 but that the Dmc1 recombination complex coexists with Rad51, suggesting that these two homologs probably have distinct roles even though a number of structural parameters of Rad51 and Dmc1 filaments are very similar (Bishop, 1994; Bishop et al., 1992; Sheridan et al., 2008). Unlike somatic recombination, meiotic recombination has a strong bias in the choice of template for DNA repair: there is a clear preference for the homologous chromosome over the sister. That bias can be justified a posteriori by the obligatory crossover rule: one needs to have inter-homologous (IH) templates to ensure proper chromosome segregation (de Massy et al., 2013; Mercier et al., 2015).

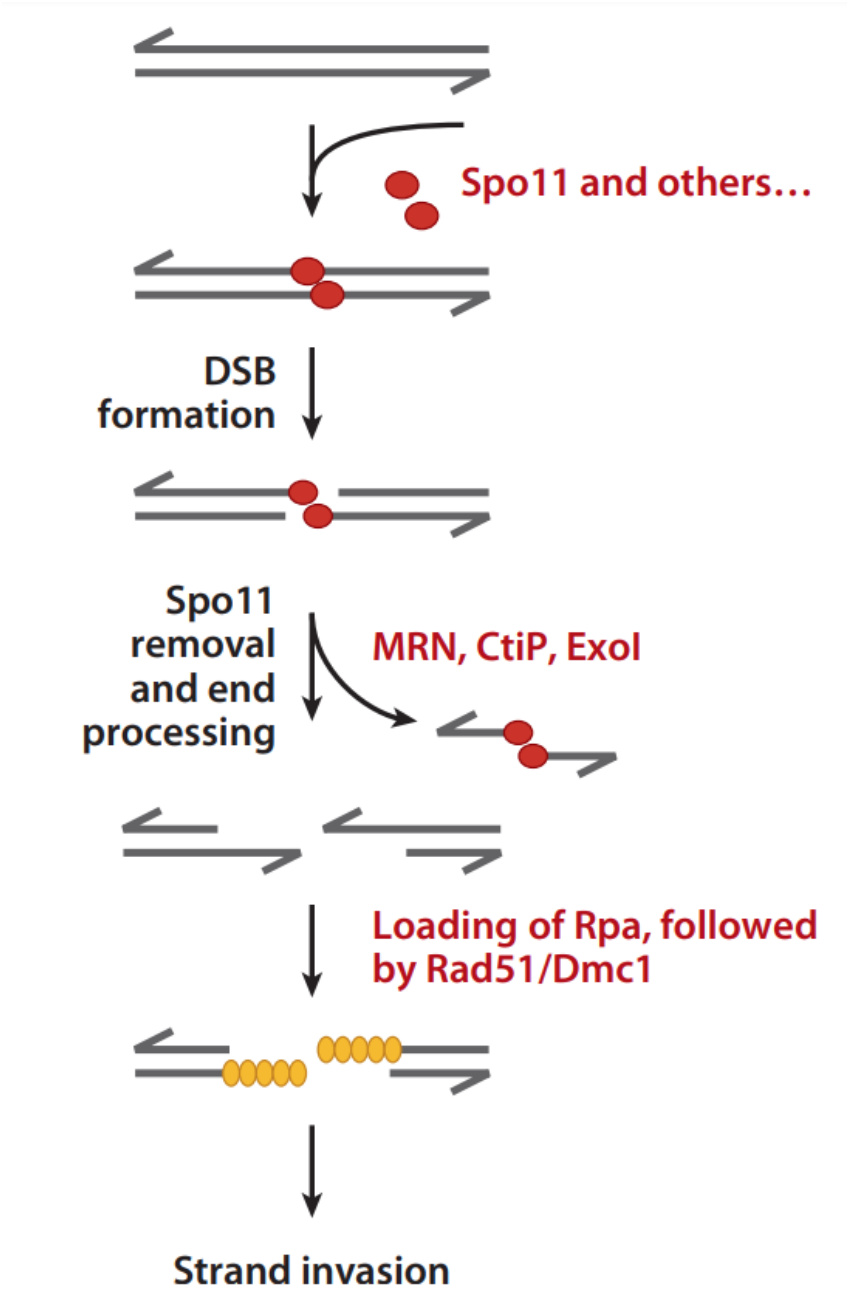


Figure 6. (adapted from de Massy, 2013) The formation of DNA double strand breaks (DSBs). The Spo11 proteins form dimers that link covalently to DNA with the assistance of other proteins to form a transient structure. Then, Spo11 proteins are removed e.g. by the MRX complex (Mre11, Rad50, and Xrs2 protein and Sae2 in yeast). Then, the maturation of the DSB depends on a 5' to 3' exonuclease that performs resection, and on the protection of the 3' strand by the Rpa complex. That is followed by the replacement of that complex by the strand exchange proteins Rad51/Dmc1.

2.2.3 The pairing of homologous chromosomes

In leptotene, after the RecA filament is formed, it must find a sequence homologous to the associated single strand of DNA, and that is associated with a search that involves DNA/chromatin/chromosome movements. In cytological studies, “bridge” structures can be observed associated with such interhomolog interactions (Albini & Jones, 1987). In such a bridge structure, a "leading" DSB end and the homologous regions from the homologous chromosome form a nascent D-loop structure, suggesting that this "leading" DSB acts as a "tentacle" to search for its partner in the chromosome. This homology-searching system occurs in a chromatin loop that at its base is tethered to an axis of proteins forming the axial element (Kim et al., 2010; Panizza et al., 2011; Storlazzi et al., 2010). In vitro experiments showed that the searching process can be finished rapidly (Yancey-Wrona & Camerini-Otero, 1995). Nevertheless, in vivo the search for homologous regions generally is a lengthy process, perhaps because there is so much to explore. If the DSB-mediated pairing process is carried out without any pre-disposition such as global pairing, coupling or clustering, it may lead to the chromosome entanglements (Zickler & Kleckner, 2015). In addition to the "tentacle" hypothesis, "stirring forces" also help homologous regions to find each other. Several contributing stirring forces are non-thermally driven motion (for instance relying on the cytoskeleton), chromatin remodeling, DNA/RNA metabolism, or prophase chromosome structure assembly. During leptotene, there is a noticeable feature related to the spatial organization of chromosomes called the "bouquet". This structure is characterized by telomeres that are attached to a localized area of the nuclear envelope, facilitating the pairing of homologs (Scherthan, 2001). However, this configuration may not systematically play a major role since the coalignment is finished before the bouquet formation in some species (Zickler, 2006).

2.2.4 The synaptonemal complex

Following coalignment, synapsis takes place in zygotene and a special structure, called the synaptonemal complex (SC), mediates synapsis that in effect zips the two homologous chromosomes to one another. In the canonical meiosis program, the complete SC formation and the dissociation of SC define the pachytene and diplotene stage, respectively (Zickler & Kleckner, 2015). The SC, a conserved tripartite protein structure, contains a central region between two axial elements that each anchors two sister chromatids. This central part consists of transverse filaments, including for instance the Zip1 protein having in yeast a coiled-coil domain, and it plays a crucial role in mediating the recombination complexes, allowing them to transit from the on-axis position (before and during coalignment) to a between-axis position within the SC's central region (Hunter, 2015; Zickler & Kleckner, 2015). The correlation between the SC and recombination complexes was first identified using electron microscopy (EM), showing that "recombination nodules" are located in the central regions of SC (Carpenter, 1975). This picture was further supported by immunolocalization of recombination proteins (Moens et al. 2002; Higgins et al. 2004; de Boer et al. 2006). In *Sordaria macrospora*, Espagne et al. (2011) identified Sme4, a component of the SC, that is required for relocalizing the recombination complexes including Rad51, Mer3, and Msh4 from the chromosome axes (lateral elements) to more central regions. It has been shown in many organisms that the SC is necessary for recombination complexes and more generally that the SC facilitates the maturation of crossovers from DSBs (Börner et al. 2004; Storlazzi et al. 2010; Qiao et al. 2012; Yokoo et al. 2012; Reynolds et al. 2013).

2.2.5 The mechanisms for repairing double-strand breaks

Crossover intermediates are repaired according to different mechanisms (Figure 7). After DSB formation and homology search by the filament, there can be strand invasion leading to a D-loop. Some of these intermediates will lead to polymerisation of the single strand using the homologous template. If the end gets ligated to the other filament one obtains a double-Holliday junction (dHJ). Then, depending on the way such dHJ are resolved (the cleavages can arise in topologically inequivalent ways), the DSBs will be repaired as crossovers (CO) or noncrossovers (NCO) (Figure 8). For CO formation, two pathways, ZMM-dependent and ZMM-independent, are separately responsible for class I and class II COs. ZMM-dependent (class I) COs are subject to CO interference, a phenomenon suppressing the CO occurring closeby, while the ZMM-independent COs (class II) seem to be noninterfering COs (Mercier et al., 2015; Pyatnitskaya et al., 2019).

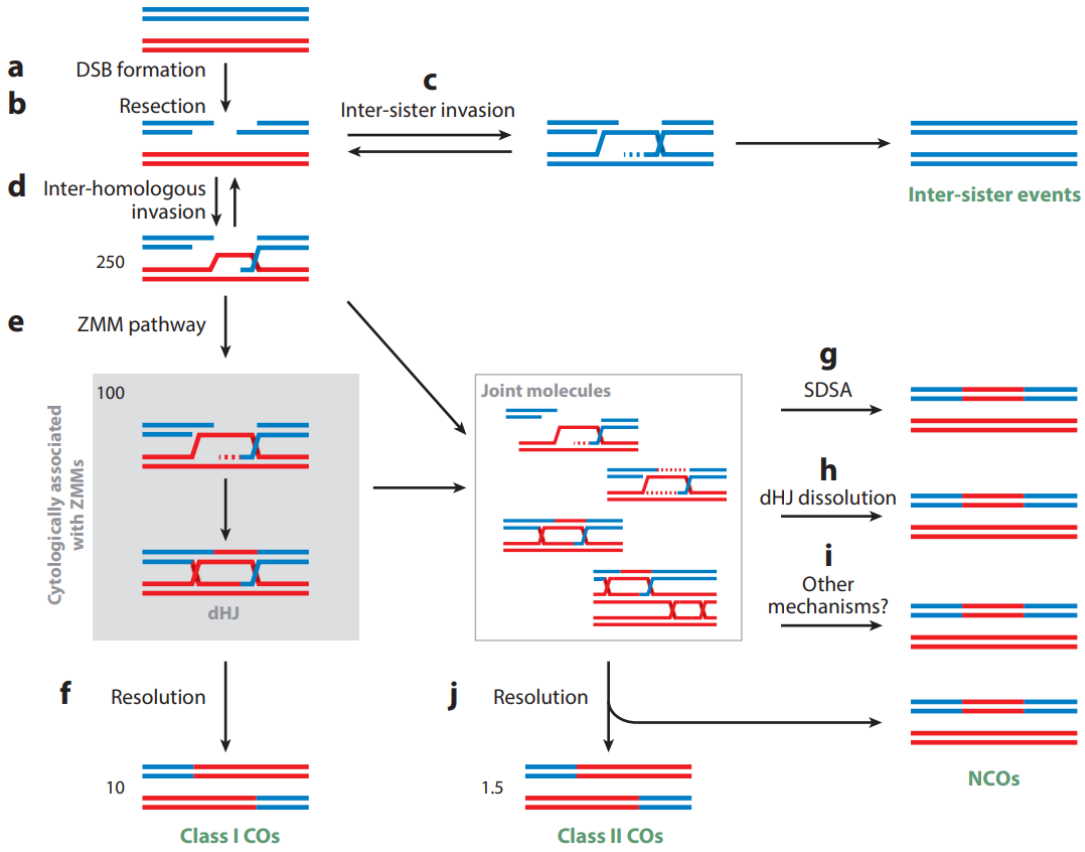


Figure 7. (Adapted from Mercier et al., 2015) The mechanisms of meiotic recombination. Beginning with the formation of double-strand breaks (DSBs) catalyzed by Spo11 proteins (a), single strands are produced by resection leading to a filament that will do the homology search (b). Then, a single strand invasion can occur on the sister chromatid (c) or on a homologous chromatid (d). The inter-homologous invasion produces D-loops that can further enter the ZMM pathway (e), leading to a double-Holliday junction (dHJ) that upon maturation can generate class I crossovers (COs) (f). On the other hand, crossovers independent of ZMM proteins are defined as class II COs (j). The recombination intermediates including D-loops, dHJ and other joint molecules can undergo different mechanisms to become noncrossovers (NCOs), such as synthesis dependent strand annealing (SDSA) (g), dHJ dissolution (h), and others (i).

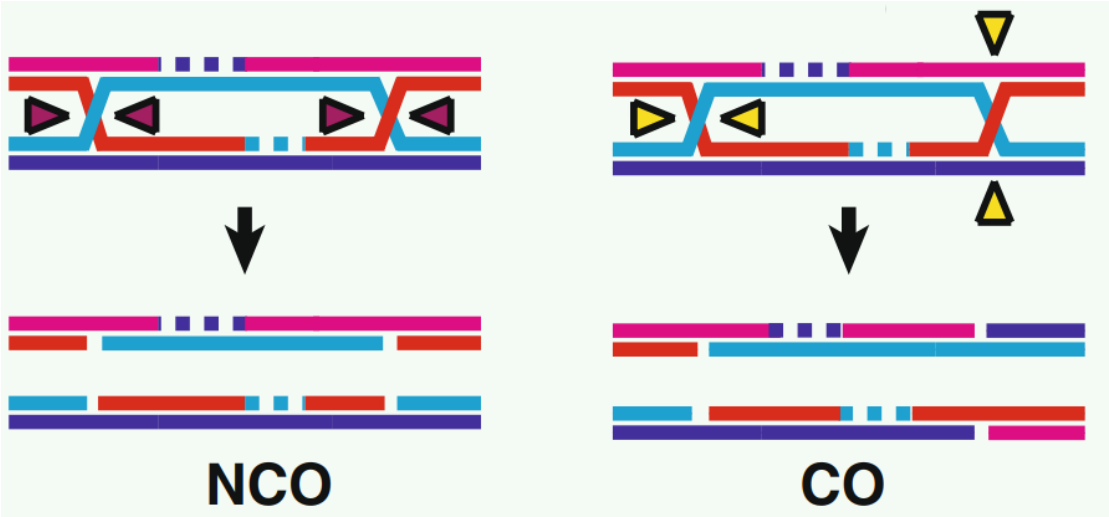


Figure 8. (Adapted from Schwartz & Heyer, 2011) The resolution of a double Holliday junction driven by topologically distinct combinations of endonuclease cleavages, leading respectively to a noncrossover or a crossover.

Required for class I COs, the ZMM-dependant pathway involves a number of proteins (Zip1, Zip2, Zip3, Zip4, Mer3, Msh4, Msh5, and Spo16) that were first identified in *S. cerevisiae* (Börner et al. 2004; Shinohara et al., 2008). Specifically, the phosphorylation of Zip1 provides a patch of negative charges on DSB sites that seems to help recruit other ZMM proteins involved in later DSB repair (Chen et al., 2015). Zip2, Zip4 and Spo16 work together as the ZZS complex that prefers to bind branched DNA, such as D-loops and dHJs, coordinated by Zip4 and Spo16, and acts as a hub, assisted by Zip2, to connect the chromosome axis components and crossover machinery (De Muyt et al., 2018). Zip3 is an E3 ligase that has a C3HC4 zinc finger RING domain which facilitates SUMOylation, one of the multiple post-translational modifications implicated in DNA repair (Cheng et al., 2006; Psakhye & Jentsch, 2012). This modification, conferred by Zip3, possibly serves as a consolidation of crossover machinery and other proteins which are required for DSB repair (De Muyt et al., 2018). Mer3 is a helicase which can unwind D-loops and dHJs, but it also interacts with the MutLβ complex, Mlh1-Mlh2 in yeast, to stop the extension (Duroc et al., 2017). Msh4 and Msh5 form the

MutSγ heterodimer that binds and stabilizes the recombination intermediates that are not disassembled by anti-recombinases (Jessop et al., 2006). Two additional proteins, MLH1 and MLH3, though not considered as ZMM proteins, are further required in the pathway of class I crossovers. These two proteins form the MutSγ heterodimer that has an endonuclease activity that is considered to resolve recombination intermediates into crossovers, and they are also the last markers which can be detected at late prophase I for class I COs (Gray & Cohen, 2016).

Class I COs form the majority of crossovers in many species, accounting for more than 75% of the total crossovers (Mercier et al., 2015). Mutations of genes involved in the formation of class I COs result in a significant reduction of crossovers. In yeast, single and double *zmm* mutants had only 15% of the number of crossovers of wild-type plants at 33°C (Börner et al. 2004). In *Arabidopsis*, a number of mutants of ZMM proteins, *Atshoc1*, *Athei10*, *Atzip4*, *Atmsh4* and *Atmsh5,* led to a significant reduction of chiasma frequency all the way down to 15% of the wild-type frequency, respectively (Chelysheva et al., 2007; Chelysheva et al., 2012; Higgins et al., 2004; Higgins et al., 2008; Macaisne et al., 2008). Besides, the *Atmer3* mutant resulted in 2.25 chiasma per cell on average, compared to wild-type plants with 9.2 averaged chiasma per cell, indicating a 75% reduction in crossover frequency by this mutation. Furthermore, two double mutants, *Atmer3/Atshoc1* and *Atmer3/Atmsh4*, respectively, led to 1.41 and 1.35 mean chiasma, corresponding to having only about 15% of residual crossovers (Macaisne et al., 2011). In rice, the mutations of *Osmer3*, *Oszip4* and *Oshei10* result in residual crossover numbers ranging from 27% to 31% of the number in wild-type plants. In addition, double mutants of *Osmer3*/*Oszip4* and *Osmer3*/*Oshei10* only maintained about (10 - 15%) of crossovers compared to the wild type genotypes (Shen et al., 2012; Wang et al., 2009; Wang et al., 2012). For both the *mlh3* mutant of *Arabidopsis* and rice, the crossover reductions are about

61% (*Atmlh3*) and 24% (*Osmlh3*), less severe than in the *zmm* mutants (Jackson et al., 2006; Mao et al., 2021). This result suggests that there are other proteins involved in the ZMM pathway to resolve DSB intermediates into class I COs.

Compared to the ZMM-pathway mechanism, the non ZMM-pathway mechanism is even less elucidated. In yeast, MUS81, Yen1 and SLX1 have an endonuclease activity, contributing to the formation of class II COs (De Muyt et al., 2012; Zakharyevich et al., 2012). In plants, the *mus81* mutant produces only about 10% fewer crossovers than the *Arabidopsis* wt, and the MUS81 foci per meiocyte in barley occupy ~12% of crossovers (Berchowitz et al., 2007; Desjardins et al., 2020). Interestingly, the non-interference characteristics carried by class II COs are exclusively for the distribution of class II COs. Anderson et al. (2014) utilized light and electron microscopy to identify class I and II COs, and they found that the interference between COs of the two classes exists.

In most organisms, the number of DSBs is far greater than the number of COs (Gray & Cohen, 2016), suggesting that the majority of crossover intermediates become NCOs that repair by copying the sequence information from the homologous chromosome without reciprocally exchanging large fragments between homologs (Mercier et al., 2015). For instance, more than 90% of DSBs lead to the formation of NCOs in *Arabidopsis* and maize (Franklin et al., 1999; Xue et al., 2018). The propensity of DSB repair to produce NCOs was also found in recombination intermediates in the ZMM-dependent pathway. Indeed, different studies showed that the ZMM foci are more numerous than COs, suggesting that these recombination intermediates are dynamic and can be dismantled even though they are protected by ZMM proteins. In *S. macrospora*, ~60 out of ~80 Msh4 foci disappear from late zygotene to mid pachytene (De Muyt et al., 2014). In *Arabidopsis*, the average number of

Msh5 foci per nucleus has a significant reduction from 76.1 to 15.5 when going from early zygotene to early pachytene (Higgins et al., 2008).

D-loops can either form dHJs for producing COs as indicated in previous paragraphs, or be disassembled. The dHJ themselves can be dissolved, generating NCOs using the conserved protein complexes Sgs1-Top3-Rmi1 (STR) and BLM-TOPIIIα-RMI1 (BLAP75)/RMI2 (BLAP18) (BTR) in yeast and human, respectively. In addition, the disassembled D-loop intermediates can produce NCOs following synthesis-dependent strand annealing (SDSA). In yeast, the STR complex facilitates the normal formation of recombination intermediates, and then promotes the NCO formations. If each of the three genes is absent, these intermediates will be repaired by the ZMM-independent pathway (Mus81-Mms4 and Yen1) for producing class II crossovers or NCOs (De Muyt et al., 2012; Kaur et al., 2015). In *Arabidopsis*, three homologous genes of the STR complex also limit the abnormal progress of meiotic recombination. The *AtRmi1* and *AtTop3α* mutants showed fragmented DNA in the late prophase or early meiosis termination, and the *AtRmi1* is crucial for the DSB repair (Chelysheva et al., 2008; Hartung et al., 2008). Furthermore, both *RECQ4*, the *Arabidopsis* Sgs1 homolog, and *FANCM* are involved in NCO pathways that limit CO formation, and the mutation of each of these two genes lead to the formation of additional COs belonging to the non-ZMM pathway (Crismani et al., 2012; Séguéla-Arnaud et al., 2015).

2.3 Where to place crossovers - The regulation of the number and distribution of crossovers

2.3.1 The distribution and number of double-strands breaks and crossovers

DSBs are not randomly distributed in the genome (Borde & de Massy, 2013). DSB sites depend on SPO11 accessibility, chromatin state, and binding of transcription factors or specific DNA-binding proteins. In *S. cerevisiae* and *S. pombe*, DSBs preferentially locate to nucleosome-depleted regions (NDRs), and these are highly correlated with promoters of genes (*S. cerevisiae*) and with large intergenic regions (> 3 kb) (*S. pombe*). Unlike *S. pombe*, the DSB sites of *S. cerevisiae* exhibit an enrichment of H3K4me3 deposited by SET1 complex, but an analysis focusing on the NDRs in promoter regions showed that the H3K4me3 status lacks predictive power when comparing DSB frequencies between different promoters (Borde et al., 2009; Tischfield & Keeney 2012). In mammals, a crucial factor that determines the position of DSB hotspots is the DNA motif recognized by the PRDM9 protein. PRDM9 has a methyltransferase domain and a specific DNA-binding domain with C2H2 zinc fingers; it recognizes a DNA-specific sequence and drives H3K4me3 formation (de Massy et al., 2013). The analysis in humans and mice showed that most DSB hotspots contain the motif recognized by PRDM9 and have the H3K4me3 mark (Brick et al., 2012; Pratto et al., 2014).

As opposed to the case of mammals, the PRDM9 protein isn't present in plants (Zhang & Ma, 2012) and so it is possible that there are no specific sequences responsible for CO localisation. By sequencing oligonucleotides covalently bound by SPO11-1 in *Arabidopsis*, Choi et al. (2018) were able to map DSB positions and found that the DSB level rises in open chromatin with AT-rich sequences including gene promoters, terminators and introns. In addition, H3K4me3 is enriched in 5' ends of genes close to DSB hotspots.

Interestingly, the DSB hotspots overlap with DNA transposons, Helitrons and Pogo/Tc1/Mariner, located in pericentromeric regions of chromosomes. On the contrary, the occurrence of DSB is suppressed in regions which are enriched in retrotransposons (Gypsy LTR). In maize, DSB sites are found in all chromosome regions even in centromeres. According to the genomic components, ~73.1 % of the DSB hotspots are deposited in repetitive sequences, mainly *Gypsy* retrotransposons. Moreover, a 20-bp GC-rich DNA motif was identified in 72% of genic DSB hotspots but not in nongenic DSB hotspots, and genic DSBs were the primary source for the CO formation (He et al., 2017).

As mentioned in the previous section, only a small fraction of DSBs become COs. The CO number is strictly controlled among different organisms. Based on the diverse data sources, ~80% of chromosomes from more than 35 species have fewer than 1.5 COs per meiosis (corresponding to a genetic length of 150 cM), and this behavior is irrespective of the chromosome size (Figure 9). For example, the physical sizes of chromosome 1 from *Arabidopsis*, tomato and barley are 35, 90 and 622 Mb, respectively, and these chromosomes show similar genetic sizes (*Arabidopsis*: 111 cM, tomato: 117 cM, barley: 133 cM). If the few COs produced per meiosis were randomly distributed, one would end up with chromosomes (actually bivalents) without any crossovers. Nevertheless, a phenomenon named CO assurance ensures that each chromosome pair will have at least one CO, ensuring that chromosomes segregate properly in meiosis. CO interference contributes to the regulation of CO numbers. In yeast, the SUMOylated form of two proteins, TOPOII (Topoisomerase II) and Red1 (an axis component) are required for CO interference. Three *top2* strains reduced by ~30% the inter-CO distance, leading to correspondingly more elevated CO numbers (Zhang et al., 2014). In addition, CO numbers are maintained when DSB numbers are modified in mouse, *C. elegans* and *S. cerevisiae*, and this phenomenon is called CO homeostasis (Wang & Copenhaver, 2018). However, in the case of maize, the

mean chiasma number was found to correlate linearly with the mean number of RAD51 foci when considering different maize inbred lines, suggesting this homeostatic control is not sufficient in that species (Sidhu et al., 2015).

Similarly to DSBs, COs are unevenly distributed in the genome. In general, CO occurrence is associated with promoter and open chromatin regions, and COs are highly suppressed in heterochromatic regions such as centromeres. In *Arabidopsis*, COs preferentially locate to (nucleosome-depleted) subtelomeric and pericentromeric but not centromeric regions. Based on SPO11-1-oligo-enriched regions, more than half of COs identified in Rowan et al. (2019) are covered by DSB hotspots defined in Choi et al. (2018). Moreover, COs of *Arabidopsis* from the Col/Ler cross are associated with (A/T), CTT/GAA, CT and CCN repeats according to a fine-scale analysis (Rowan et al., 2019). In rice, ~5% of the genome has more than 80% of the historical recombination events. The CO hotspots are enriched in simple sequence repeats and DNA transposon classes including *PIF*, *Harbinger* and *Stowaway*, but lack retrotransposon classes (Marand et al., 2019). In maize and wheat, plant species with particularly large genomes and an abundance of repeat sequences, exhibit a CO landscape with still higher contrast. Indeed, in terms of physical length, COs only occur in ~7% of the maize whole genome, and 19% of the wheat chromosome 3B obtain ~82% of COs, the large interstitial and centromeric regions suppressing COs in these species (Choulet et al., 2014; Darrier et al., 2017; Kianian et al., 2018). In addition, COs are depleted in regions close to TEs in maize, and are less frequent in the retrotransposon regions of wheat. Yet, COs have been shown to be associated with two motifs (A-stretch and CCG) and two DNA transposons (*TIR-Mariner* and *CACTA*) in wheat.

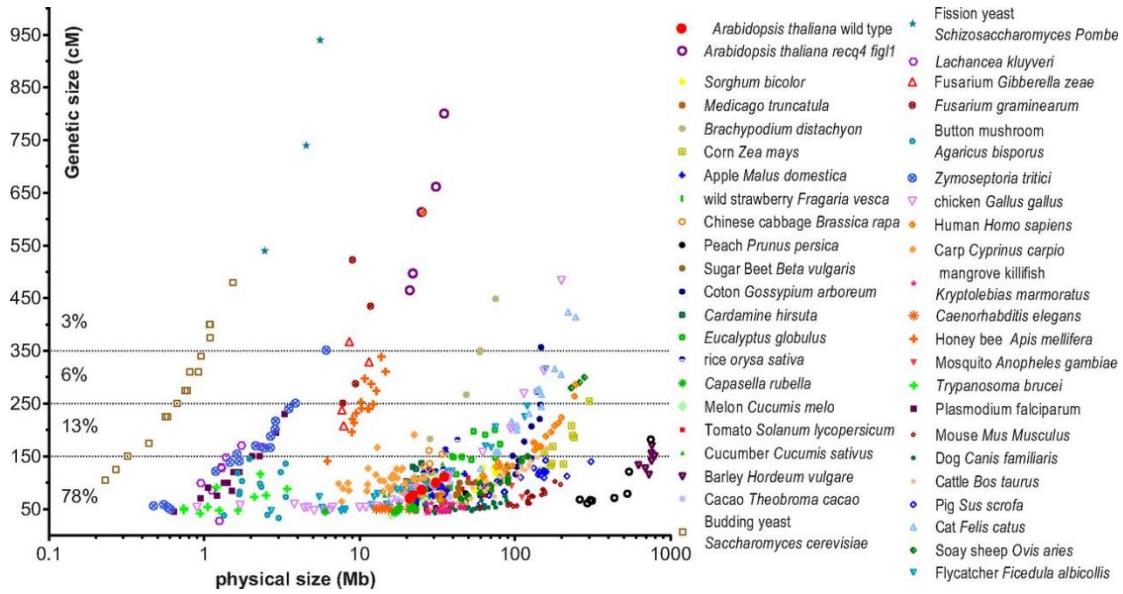


Figure 9. (Adapted from Fernandes et al., 2018) Crossover numbers per meiosis across a large number of eukaryotic organisms. The x-axis and y-axis are physical (log scale) and genetic sizes (linear scale), respectively. Each dot represents a chromosome, the genetic length is based on F_2 populations (and is thus the average of male and female meiosis). Sex chromosomes were not included in this figure. All dots can be classified into the four intervals of CO numbers, leading to the percentages shown on the left for the chromosomes falling in the corresponding intervals.

2.3.2 Crossover interference and its modeling

The phenomenon of crossover interference, whereby a crossover occurrence at one position seems to inhibit the occurrences of other crossovers nearby on the same chromosome, results in there being fewer crossovers than DSBs. In yeast and *Sordaria*, the synaptonemal complex (SC) formation is concomitant with CO designation, suggesting that CO interference guarantees the organized formation of the SC required for CO formation (Fung et al., 2004; Zhang et al. 2014). Another study based on the transverse filaments connecting homologs in *Arabidopsis* bivalents (via the ZYP1 protein) also indicates that the SC is associated with CO interference (Capilla-Pérez et al., 2021). In that study, the authors showed that the double mutants of ZYP1 (*zyp1a zyp1b*) have increased number of COs but no synapsis, and neither CO interference nor heterochiasmy were detected, suggesting that the SC is involved in the regulation of CO interference and heterochiasmy. On the contrary, SC formation is independent of DSBs in *C. elegans* and Drosophila (Rog & Dernburg 2013; Takeo et al., 2011; Tanneti et al., 2011).

The phenomenon of interference was discovered over a century ago by Sturtevant in *Drosophila* (Sturtevant, 1913). Since then, different mapping functions for estimating genetic maps based on recombination were established, Haldane's function without CO interference and Kosambi's function with CO interference (Haldane, 1919; Kosambi, 1943). To date, it has been found that CO formation arises through two pathways controlled respectively by ZMM proteins and (among others) Mus81. COs produced by the ZMM pathway, normally representing 75% or more of all crossovers in many organisms, are subject to CO interference while the COs produced by the other pathway seem to be free of interference.

Modeling CO interference relied on data from genetic and cytological maps. The associated frameworks fall into two classes: they use either statistical or

physical modeling approaches. For the statistical based-frameworks, a convenient approach consists in adding dependencies either between the number of COs (forcing deviations from the Poisson model) or between the distances of COs so that close-by COs are depleted, leading to the main characteristic found experimentally from CO interference. One of the most popular such models was introduced by McPeek and Speed (1995) and is referred to as the “Gamma model”. It has been shown in many systems that such models perform better than the Haldane model (having no CO interference) when analyzing recombination data. Falque et al. (2009) applied the Gamma model and the Haldane model to analyze maize CO data, and inferred quantitatively the contributions of the two CO pathways (interfering and non-interfering).

Among the models based on physical frameworks, one of the oldest is based on assuming that there is a polymerization process along synaptonemal complexes that leads to a coarsening of objects that can be thought of DSBs leading to COs displaying interference (King & Mortimer, 1990). Another physical model is referred to as the “beam film model”. This model uses a mechanical analogy and introduces precursors that “mature” as would crack in a system subject to mechanical stresses. Specifically the reference physical system consists of two materials with different coefficients of thermal expansion. With the increasing stress, some of the precursors mature, again leading to “crossovers” subject to CO interference due to the fact that the maturation process releases stress locally, so that a maturation event will inhibit the maturation of other cracks nearby (Kleckner et al., 2004). A much more recent physical modeling approach was taken by Morgan et al (2021), also based on a maturation process associated with coarsening dynamics. Those authors performed modeling of maturation of *HEI10* foci whose intensity changes with time as observed using super-resolution microscopy. Their model lets HEI10 diffuse along the SC and accumulate in designated

sites assumed to be DSBs. They showed that their physical model explained the CO patterning experimentally observed, with clear interference effects. In brief, these models are addressing the way multiple COs interact during one meiosis, but generally do not consider the question of recombination landscapes, and as a result the landscapes are very smooth (Morgan et al., 2021; Fozard et al., 2022). Of course CO interference in a model will affect the CO landscape but this effect is rarely considered as the factors driving small scale variations in recombination landscapes depend on genomic and epigenomic features that are not part of such models focused on CO interference. In the next section, I will introduce different factors that have an effect on the number and distribution of crossovers, where by distribution I include the landscape aspects.

2.3.3 The factors for the regulation of crossover number and distribution in plants

2.3.3.1 Methylation, histone modification and chromatin remodeling

As mentioned before, COs are highly suppressed in plant heterochromatin. Plant heterochromatin is maintained by DNA methylation and H3K9me2, a histone modification mark, for the regulation of diverse processes such as RNA Pol II transcription, chromatin condensation and chromatin-chromatin interactions (Fransz et al., 2002; Soppe et al., 2002; Zhang et al., 2006; Feng et al., 2014). DNA methylation occurs at cytosine bases within three contexts, namely CG, CHG and CHH. The *Arabidopsis* genome-wide methylation patterns shows that more symmetric methylation was identified (~24% of CG and ~6.7% of CHG) than asymmetric methylation (~1.7% of CHH) (Law & Jacobsen, 2010). In plants, the occurrence of CO is negatively correlated with DNA methylation levels. In *Arabidopsis*, the CO hotspots have low levels of DNA methylation in three contexts, but the CO deserts have the high level of

CG methylation in structural variations (SVs), TEs, and even some regions with protein-coding genes (Choi et al., 2013; Rowan et al., 2019). In maize, the CG methylation level and CO frequency are also negatively correlated on the broad scale (Rodgers-Melnick et al., 2015).

Three DNA methylation contexts are maintained by different systems. Methyltransferase1 (Met1) and the SWI/SNF chromatin remodeling protein "Decreased DNA Methylation1" (DDM1) work together to maintain CG methylation (Vongs et al. 1993; Saze et al. 2003; Stroud et al. 2013). On the other hand, Chromomethylase2 (CMT2), Chromomethylase3 (CMT3) and "Domains Rearranged Methylase2" (DRM2) individually maintain non-CG methylation (Cao et al., 2003; Stroud et al. 2013, 2014; Zemach et al., 2013). In *Arabidopsis*, both *met1* and *ddm1* mutants significantly reduce the methylation level in pericentromeric regions, but surprisingly that does not lead to increased recombination rate in that region. Instead, the recombination rate in the euchromatic region is elevated. Noting that Mirouze et al (2012) even discovered that the total CO number in *met1* mutants and wild type are very close, one concludes that these *met1* CG methylation mutants only redistribute COs in the *Arabidopsis* genome (Melamed-Bessudo & Levy 2012; Mirouze et al. 2012; Yelina et al. 2012). On the contrary, Underwood et al. (2018) used the mutation of *CMT3* and H3K9 methyltransferase genes *KYP/SUVH4 SUVH5 SUVH6* to discover that a significantly reduced non-CG methylation level is associated with an elevated recombination rate in pericentromeric regions. Furthermore, the reduction of CG and non-CG methylation leads to an increased DSB occurrence, suggesting that both symmetric and asymmetric methylation restrict DSB formation but only asymmetric methylation and/or H3K9me2 inhibits CO formation (Choi et al., 2018; Underwood et al., 2018). Machine learning approaches also suggested that DNA methylation and nucleosome occupancy are both important for CO sites in maize and *Arabidopsis* (Wang et al., 2022).

2.3.3.2 Heterozygosity from chromosomal scale to small scale

The mismatch repair (MMR) system, MutS and MutL (or their homologs), can be a barrier for recombination when in presence of diverged sequences (Dluzewska et al., 2018). In tomato, a BC_1 population from the interspecies hybrid between the cultivated (*L. esculentum*) and wild tomato (*S. lycopersicoides*) has ~27% reduction of genetic length for all chromosomes compared to two populations (F_1 and BC_1) from the *L. esculentum* x *L. pennellii* cross (Chetelat et al., 2000).

Let us now consider small scales. The a1-sh2 interval of the maize genome is a 140-kb recombination hotspot that contains four genes and abundant SNP and InDel polymorphisms (Yao et al., 2002). The comparison among haplotypes showed that the recombination rate associated negatively with sequence polymorphisms for subintervals and their adjacent subintervals. Nevertheless, this correlation cannot fully explain the relationship of nonadjacent subintervals (Yao & Schnable, 2005). That study could be problematic because the same effect was assigned for SNPs and InDels. In *Arabidopsis*, the pollen-typing method was utilized to identify ~1,000 crossovers from the Col/Ler F_1 plants within the RESISTANCE TO ALBUGO CANDIDA1 (RAC1) R gene hotspot; this intragenic hotspot also showed a negative relationship between recombination rate and SNP frequency (Choi et al., 2016).

Among the two CO pathways, it seems that mainly the non-interfering CO pathway is sensitive to heterozygosity. The *fancm* and *fancm zip4* mutants produces COs from both pathways and from the non-interfering pathway only, respectively. In *Arabidopsis*, when considering the 420 interval (Chr3: 0.2 Mb - 5.3 Mb) in a homozygous background, both mutants have significantly higher recombination rate than the wild-type line. On the contrary, compared to the wild-type line in the same interval with heterozygous status, the *fancm*

mutant showed a comparable recombination rate as the wild-type line, and the *fancm zip4* mutant had a strikingly reduced recombination rate (Ziolkowski et al., 2015). However, the class II COs can still be repaired in heterozygous regions in the *figl1* (FIDGETIN-Like-1, another meiotic anti-CO factor) mutant (Girard et al., 2015). Moreover, the double (*recq4 figl1*) and triple mutants (*recq4 figl1 fancm*) of meiotic anti-CO factors have not only substantially increased recombination rate but also sensitivity to heterozygosity (Fernandes et al., 2018). The *recq4 figl1* mutant leads to a strong negative correlation between recombination rate and SNP density, which is not present in the wild-type lines. From centromere to telomere, the recombination rate of all mutants just increases moderately in SNP dense pericentromeric regions but rises significantly in arms, then reaches the maximum at regions close to telomeres (Fernandes et al., 2018). This result supports that the class II COs are more sensitive to SNP density, and they tend to locate in telomeres instead of pericentromeric regions.

Interestingly, even though different studies showed that heterozygosity somewhat suppresses crossover rate, it was found that crossovers are increased in a heterozygous segment juxtaposed by two homozygous segments (Ziolkowski et al., 2015). Based on 6 F_2 populations, it was shown that SNP density and recombination rate have a non-monotonic relationship, specifically regions with intermediate SNP density have more COs than regions with too many or too few SNPs (Blackwell et al.,2020).

2.3.3.3 Heterochiasmy

Heterochiasmy refers to male and female meiosis having significantly different recombination rates. In *Arabidopsis*, two BC_1 populations with more than 3,000 individuals in total, derived from the cross between Col and Ler, were used to identify 13,535 crossovers (Giraut et al., 2011). The genome-wide genetic length of male meiosis is 575 cM, 0.73 times larger than the one of

female meiosis. The most contrasting difference for crossover frequency between the male and female meiosis is located in the telomeric intervals, where the male recombination rates are quite high but the female recombination rates are very low. Even if the ends of chromosomes are removed from the analysis, the male recombination rate of the remaining regions is still significantly higher than the female one, suggesting that the male meiosis in *Arabidopsis* tends to recruit more COs than the female meiosis (Giraut et al., 2011).

In maize, male and female meiosis have similar trends for CO number and distribution, but there are differences for the CO rates and chromatin features at local scales (Kianian et al., 2018). For the CO sites close to genes, male derived COs tend to associate with the genes related to phosphorylation-related processes, and female COs are more related to genes obtaining oxidoreductase activity and cofactor binding. In addition, for CO located in promoters, more male COs were deposited at ~ 400 bp upstream from TSS, while female CO peaks were more often identified close to the TSS. For the H3K4me3 levels of CO locations, the male H3K4me3 peaks are located farther (~250 bp upstream) from the CO sites than the female peaks.

Quantitative Plant Biology

cambridge.org/qpb

Quantitative modelling of fine-scale variations in the *Arabidopsis thaliana* crossover landscape

Yu-Ming Hsu[1,2,3], Matthieu Falque[3] and Olivier C. Martin[1,2,3,*]

[1]Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France; [2]Université de Paris, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91405, Orsay, France; [3]Université Paris-Saclay, INRAE, CNRS, AgroParisTech, GQE - Le Moulon, 91190 Gif-sur-Yvette, France

Original Research Article

Cite this article: Y.-M. Hsu et al. Quantitative modelling of fine-scale variations in the *Arabidopsis thaliana* crossover landscape. *Quantitative Plant Biology*, **3**:e3, 1–11
https://dx.doi.org/10.1017/qpb.2021.17

Received: 21 October 2021
Revised: 10 December 2021
Accepted: 15 December 2021

Keywords:
recombination rate; chromatin state; epigenetic features; sequence divergence.

Author for correspondence:
O. C. Martin
E-mail: olivier.c.martin@inrae.fr

Abstract

In, essentially, all species where meiotic crossovers (COs) have been studied, they occur preferentially in open chromatin, typically near gene promoters and to a lesser extent, at the end of genes. Here, in the case of *Arabidopsis thaliana*, we unveil further trends arising when one considers contextual information, namely summarised epigenetic status, gene or intergenic region size, and degree of divergence between homologs. For instance, we find that intergenic recombination rate is reduced if those regions are less than 1.5 kb in size. Furthermore, we propose that the presence of single nucleotide polymorphisms enhances the rate of CO formation compared to when homologous sequences are identical, in agreement with previous works comparing rates in adjacent homozygous and heterozygous blocks. Lastly, by integrating these different effects, we produce a quantitative and predictive model of the recombination landscape that reproduces much of the experimental variation.

1. Introduction

Crossovers (COs) formed during meiosis drive the shuffling of allelic combinations when going from one generation to the next. They thereby play a central role in genetics and evolution and they are also key in all forms of breeding. Pericentromeric regions tend to be refractory to COs (Bauer et al., 2013; Choulet et al., 2014). Although these regions have a high density of transposable elements, in crops they nevertheless contain a sizable number of genes. Attracting COs into these regions could have benefits for genetic studies (e.g., to identify gene functions) and for selection of new combinations of alleles of relevance for breeding.

CO formation processes (Mercier et al., 2015; Villeneuve & Hillers, 2001) start with the active formation of double strand breaks (Keeney & Neale, 2006) and end with DNA repair, leading to either COs or non-COs (Hunter, 2015). They are tightly regulated, in particular, they ensure at least one CO per bivalent (Jones & Franklin, 2006; Zickler & Kleckner, 2016), but not many more in spite of huge variations in genome size (Fernandes et al., 2018). Furthermore, CO distribution tends to be very heterogeneous along chromosomes, indicating that there are determinants of CO formation at finer scales. Typically, pericentromeres and more generally, regions rich in heterochromatin are depleted in COs. In contrast, regions of open chromatin such as gene promoters are enriched in COs. In several species, it has been possible to measure the distribution of double strand breaks (precursors of both COs and non-COs), revealing a very high level of heterogeneity genome-wide (Khil et al., 2012; Pan et al., 2011; Pratto et al., 2014). It is generally assumed that such heterogeneities, detected all the way down to the scale of a few kb, arise also for CO distributions, but unfortunately, the resolution of CO maps in plants has been so far insufficient to fully confirm this expectation. Indeed, the best dataset in plants averages about one CO every 3.5 kb (Rowan et al., 2019).

Our objective is to shed light on genomic and epigenomic features that shape recombination rate on fine scales in *Arabidopsis thaliana*, a species chosen because it has more extensive CO datasets than other plants. Here, we exploit a recent high-resolution dataset detecting 17,077 COs in a large *A. thaliana* F_2 population (Rowan et al., 2019). The quantitative analysis of these COs provides new insights. For instance, recombination rate depends on the *size* of an intergenic

© The Author(s), 2022. Published by Cambridge University Press in association with The John Innes Centre. This is an Open Access article, distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives licence (https://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is unaltered and is properly cited. The written permission of Cambridge University Press must be obtained for commercial re-use or in order to create a derivative work.

John Innes Centre
Unlocking Nature's Diversity
CAMBRIDGE UNIVERSITY PRESS

region, there being a suppression for regions whose size is less than about 1.5 kb. Furthermore, it is possible that COs are partly suppressed by lack of single nucleotide polymorphisms (SNPs), a result that would explain the 'heterozygous block effect' found previously (Ziolkowski et al., 2015), whereby the insertion of a heterozygous block into an otherwise homozygous region enhances recombination rate therein. These different insights allow us to build a quantitative model that integrates genomic information, local epigenetic status and contextual effects. This model has low complexity, the inclusion of its different parameters is justified by AIC and BIC statistical tests, it has good predictive power and reproduces much of the recombination rate variation in *A. thaliana*, pointing to the importance of different contextual effects modulating local CO rate.

2. Materials and Methods

2.1. CO datasets

COs were inferred to lie within intervals delimited by SNPs, anchoring transitions between homozygous and heterozygous regions of F_2 individuals (Rowan et al., 2019). When measuring recombination rate in a given bin, we count one CO for each CO interval lying completely within that region, and otherwise we apply the simple *pro-rata* rule. However, for reasons of tractability, when we use the maximum likelihood method, we instead simply assign the CO to the middle of its interval (see below). We downloaded the dataset of CO intervals of Rowan et al. (2019) based on 2,182 F_2 individuals from a cross between Col-0 and Ler. We also used the data of five F_2 populations based on crossing Col-0 with five other accessions (Blackwell et al., 2020). The associated files were kindly provided by Ian Henderson, University of Cambridge, Cambridge, UK, and are included as Supplementary Material (Supplementary File S1). For the whole study, the experimental recombination rate r (in cM/Mb) was calculated using the formula: $r = 100 \times n_{\mathrm{co}} / \left(n_{\mathrm{plant}} \times 2 \times L_{\mathrm{Mb}}\right)$, where n_{co} is the number of COs contained in the relevant bin or region, n_{plant} is the number of F_2 plants and L_{Mb} is the length of the bin or region in Mb.

2.2. Genomic annotation of Col-0 and structural variations between Col-0 and Ler genomes

For Col-0 genomic features, we utilised TAIR10 annotation specifying coding genes and super families of transposable elements. We compared the TAIR10 reference Col-0 genome and the Ler assembled genome to detect syntenic regions and structural variations (SVs) (Berardini et al., 2015; Jiao & Schneeberger, 2020). SVs were identified using (freely available) MuMmer4 and SyRI software (Goel et al., 2019). The parameters used in the 'nucmer' function of MuMmer4 were set via '-l 40 -g 90 -b 100 -c 200'. All genomic and epigenomic features were computed after masking out the regions containing the SVs defined by SyRI.

2.3. Col-0 epigenomic features and segmentation of chromosomes into chromatin states

BigWig, bedGraph and bed files of H3K4me1, H3K4me3, H3K9me2, H3K27me3, ATAC and DNase measurements on Col-0 were downloaded from the NCBI and ArrayExpress databases (cf. Supplementary Table S1). Segmentation of the chromosomes into nine chromatin states was obtained from the study of Sequeira-Mendes et al. (2014) which again is specific to Col-0.

2.4. Identifying SNPs in the 5 F2 populations

The five F_2 populations (Blackwell et al., 2020) had Col-0 as shared parent, the other parent was Ler, Ws, Ct, Bur or Clc. Their sequences were downloaded from the ArrayExpress database (accession identifiers: E-MTAB-5476, E-MTAB-6577, E-MTAB-8099, E-MTAB-8252, E-MTAB-8715 and E-MTAB-9369). For aligning the reads to the TAIR10 reference genome (Berardini et al., 2015), we used the 'mem' algorithm of Burrows-Wheeler Alignment (BWA-MEM; v0.7.17) (Li, 2013), then samtools (v1.10) (Li, 2011) and bcftools (v1.12) for SNPs calling. Finally, we applied filters to keep SNPs with (a) a quality score ≥100, (b) mapping quality score ≥20, (c) depth below 2.5 mean depth of the corresponding F_2 population to eliminate anomalously high coverages indicative of multi-mappings, (d) positions that only contained uniquely mapped reads and (5) maximum allele frequency less than 0.9.

2.5. The quantitative model based on epigenetic states and genomic features

Sequeira-Mendes et al. (2014) identified nine distinct chromatin states in Col-0 segmenting the whole genome. We modified their segmentation as follows. First, noting that heterochromatic regions often contained stretches of alternating states 8 and 9, we relabelled segments of state 8 as state 9 when they were sandwiched between two state 9 segments. This relabelling affected almost exclusively segments in the pericentromeric regions and provided a proxy for heterochromatin. We verified that recombination rate was highly suppressed in such relabelled segments while non-relabelled state 8 segments (lying almost exclusively in the arms) did not lead to CO suppression. Second, we added a new state corresponding to having an SV or insufficient synteny between the two parental genomes of interest.

Given these 10 states and their segmentation of the genome, our model introduces an adjustable 'base' recombination rate for each state and then applies 3 multiplicative modulation effects associated with intergenic region size, density of SNP between homologs, and chromosome number. The modulation by the intergenic region size is straightforward if one considers a genomic segment lying entirely between two genes; if it does not satisfy that condition, we break it into underlying pieces so that each piece is either entirely within an intergenic region or entirely within a genic region; the modulation is then applied to each piece separately.

The 15 parameters of this quantitative model were identified by fitting to the experimental data using the maximum likelihood method as the measure of goodness of fit. Specifically, for a given bin, let p be the probability of introducing a CO therein during meiosis. Since the F_2 population is the result of twice as many meioses as there are plants, the likelihood of observing n_{CO} COs among the n_{plant} plants is given by the binomial distribution: $L = \mathrm{choose}(2\ n_{\mathrm{plant}},\ n_{\mathrm{CO}})\ p^{n\mathrm{CO}}\ (1 - \mathrm{p})^{2n\mathrm{plant}-n\mathrm{CO}}$, where choose() denotes the binomial coefficient. The parameters of the model were thus fitted by maximising the log likelihoods summed over all bins. To incorporate the fact that CO numbers are tightly regulated by the obligatory CO and by CO 'interference', in every iteration to fit this model, we rescaled predicted rates to ensure that

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

the predicted genetic length of each chromosome is the same as the experimental one.

Having such a maximum likelihood method allows one to compare the statistical relevance of different nested models. For instance, to determine whether the data justify including the intergenic region size effect, we can use the likelihood ratio test on the models without and with that effect. More generally, if *L0* is the likelihood of the simpler model and *L1*, the likelihood of the more complex one (having k additional parameters), then $-2\ln(L0/L1)$ follows a chi-square distribution with k degrees of freedom under the hypothesis that *L0* is the correct model. This framework allows us to reject that last hypothesis if the likelihood ratio is too small and to quote an associated p-value. Along similar lines, the AIC (Akaike information criterion) and BIC (Bayesian information criterion) criteria allow one to test whether such additional parameters are justified. Those two criteria differ in the way they penalise the number of parameters, but in any case, the AIC or BIC criterion allow one to select the best model via its minimisation of the corresponding criterion.

2.6. The software of statistical analysis and visualisation

All statistical analyses were based on R 3.63. For fitting model parameters to data, we used the 'optim' function with the method 'L-BFGS-B'. All visualisations were carried out using the 'tidyverse' package (Wickham et al., 2019). All codes are available as a gzip file (Supplementary Material), but can also be taken from the github site https://github.com/ymhsu/chromatin_state_model.

3. Results

3.1. Standard modelling of CO rate based on genomic and epigenomic variables is unsatisfactory

Based on 17,077 COs from an F_2 population (Rowan et al., 2019), we related recombination rate to the local density of various genomic and epigenomic features. As shown in Figure 1, the individual relations found are typically non-monotonic with correlations of one sign within chromosome arms and of the opposite sign within pericentromeric regions. Such a characteristic makes it difficult to assign a role to any individual feature. This result holds whether using feature data obtained from somatic tissues or from germinal tissues (cf. Supplementary Figure S1).

To combine all these features into a model, the standard approach is to consider an additive framework and then possibly generalise it by including interaction terms. The additive model corresponds to predicting recombination rate within a bin of the genome using the following formula:

$$r = a_0 + a_1 \times f_1 + a_2 \times f_2 + \cdots + a_n \times f_n, \tag{1}$$

where f_i is the density of the ith feature in the bin. In this spirit, we incorporate all nine feature densities of Figure 1 that is genes, TEs, the number of transcription starting sites, H3K4me1, H3K4me3, H3K9me2, H3K27me3, ATAC and DNase. In Supplementary Table S2, we provide the fitted values $a_0, a_1, \ldots, a_g$ when using different bin sizes. Somewhat surprisingly, the coefficient in equation (1) for gene coverage density is negative, making the interpretation of the model problematic and suggesting that the additivity assumption is not supported by the data. Finally, to have a measure of goodness of fit, we use the fraction of the recombination rate variation that is 'explained' by the model, defined as:

$$R^2 = 1 - \mathrm{mean}\left[(y - \hat{y})^2\right] / \mathrm{var}(y), \tag{2}$$

where y is the experimental and $\hat{y}$ is the predicted value of recombination rate in the different bins along the genome. R^2 as well as the coefficients in equation (1) depend on the bin size; for our 'reference' bin size of 100 kb, the model calibration gives $R^2 = 0.36$.

To allow for deviations from additivity we follow the standard practice of including interaction terms in the form of pairwise products of feature density values, leading to the formula:

$$r = a_0 + a_1 \times f_1 + a_2 \times f_2 + \cdots + a_g \times f_g + f_1 \times (b_2 \times f_2 + \cdots + b_g \times f_g) + f_2 \times (c_3 \times f_3 + \cdots + c_g \times f_g) + \cdots \tag{3}$$

This leads to 46 adjustable parameters *versus* 10 in the additive model. This more complex model explains a fraction $R^2 = 0.35, 0.43, 0.51$ and 0.66 of the total recombination rate variances when bin size is 50, 100, 200 and 500 kb. Although this is better than the additive model, the interactions do not lead to biological interpretations. Furthermore, the predictions are sometimes negative, and we also find that the fitted parameters vary substantially with bin size. Thus, this model with interactions is not satisfactory and it does not provide insights into the biological determinisms of recombination rate.

3.2. Aggregating genomic and epigenomic features using a chromatin state classifier

Given the drawbacks of the previous modelling framework, we performed aggregation using an automatic classifier approach (Sequeira-Mendes et al., 2014), assigning a 'chromatin state' to a local region according to a (non-linear) combination of such features. The methodology is general but those authors implemented it in the case of Col-0, producing 9 chromatin states based on the combination of 16 genomic or epigenomic features, namely H3K4me1, H3K4me2, H3K4me3, H3K9me2, H3K27me1, H3K27me3, H2Bub, H3K36me3, H3K9ac, H3K14ac, H4K5ac, CG methylation, H3 content, H2A.Z, H3.1 and H3.3. Their states 8 and 9 correspond to AT-rich and GC-rich heterochromatic regions, respectively, with state 9 being strongly enriched in the pericentromeric regions. Their seven other states are typically euchromatic. They found that state 1 (respectively state 6) typically colocalises with transcription start sites (TSS) [respectively, transcription termination sites (TTS)]. States 3 and 7 are the most abundant states in gene bodies, with the former one tending to be present with state 1 at the 5′ end of genic regions and the latter one arising more frequently in larger transcriptional units. States 2 and 4 typically lie within intergenic regions and they tend to be proximal and distal to the gene's promoter, respectively. Like states 2 and 4, state 5 is generally within intergenic regions, but it also arises frequently in silenced genes with high levels of H3K27me3. See also top of Figure 2 for a graphical representation of these trends.

Because COs form between homologs, we also need to aggregate information about the local synteny between Col-0 and Ler, the two parents of the F_2 population (Rowan et al., 2019) used to estimate the recombination landscape. We thus assign the state 'SV' to the non-syntenic regions. We then have a total of 10 different 'states' that we will study in the rest of this work, referring to them as 'chromatin states' even if that is not completely correct. The fraction

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

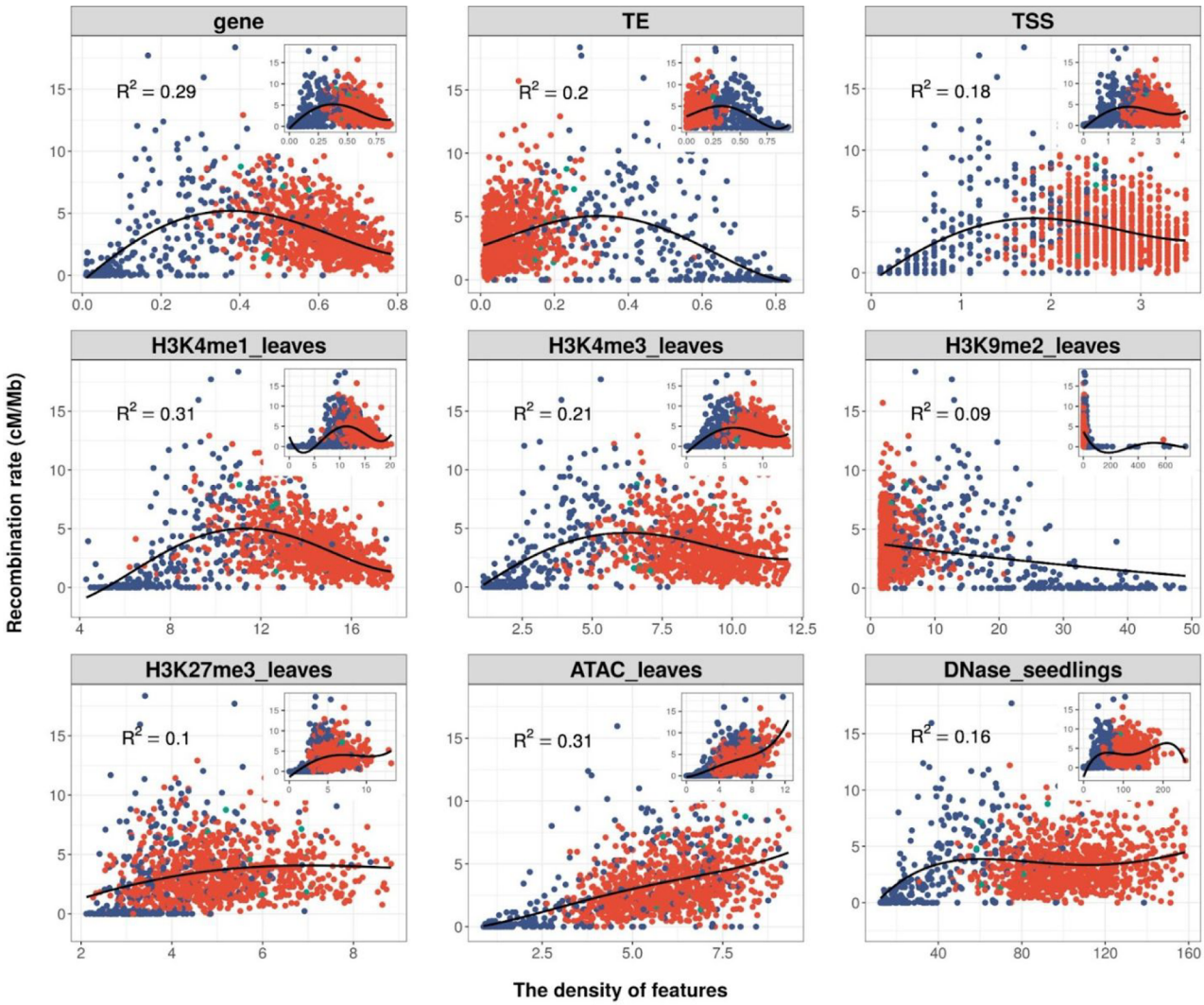


Fig. 1. The correlations between recombination rate and nine genomic or epigenomic features taken from somatic tissues (cf. titles). Each dot represents the values for a 100-kb bin. The *x*-axis shows the density of each feature, and the *y*-axis is the recombination rate based on a total of 17,077 crossovers from the Col-0-Ler F_2 population. Dots in red, blue or green are for bins located in arms, pericentromeric regions or the transition regions between arms and pericentromeric regions, respectively. The black curves are fits to polynomials of degree 4 (function lm(y ~ poly(x,4)) of the statistical package R). R^2 corresponds to the fraction of explained variance when using the polynomial as predictor (equation (2)). To ensure that the points fill most of the space, the scale in the main part of each panel is a zoom to display only 95% of the data, cutting the 2.5% extremities on both sides of the *x*-axes in all these plots. Insets show the data in the whole range.

of the genome covered by any of these chromatin states varies between 5.8 and 13.6%, with state 4 (intergenic, distal) being the most represented and state 8 (heterochromatic, AT rich) the least (cf. Figure 2a, top).

To transform the trends found by Sequeira-Mendes et al. (2014) into quantitative patterns, we have generated the frequency profiles for each chromatin state as a function of position within gene bodies and their flanking regions. For that task, we used the 25,708 genes extracted from syntenic regions and also considered their extensions on both sides, going out to 3 kb upstream of the TSS and downstream of the TTS. The computed profiles (Figure 2b, top) reveal that there is a clear gradient in the chromatin state content along the gene bodies and also along their flanking regions. For instance, the frequency of state 1 has a very sharp rise as one enters the gene on the 5′ side while the frequency of state 7 has a steep fall as one exits the gene on the 3′ side. We performed the analogous computations for intergenic regions and find that the frequency profiles there (cf. Figure 2c, top) have much less variation than in gene bodies.

3.3. A simple quantitative model of recombination rate based on discrete chromatin states and SVs

In contrast to the quantitative variables used in equation (1), the state classifier approach identifies discrete states. These can be used as factors (*qualitative variables*) in a model of recombination rate by making the perhaps simplistic assumption that each state has its own specific recombination rate. This framework both allows for a direct biological interpretation and is mathematically particularly simple. Comparing the genomic fraction of each chromatin state to the observed CO fraction for that state (top and bottom of Figure 2a) determines the 10 average recombination rates: 3.08, 4.78, 2.16, 6.37, 5.14, 3.48, 1.5, 3.35, 0.7 and 0.57 cM/Mb. Hereafter, these values are referred to as the 'experimentally measured

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

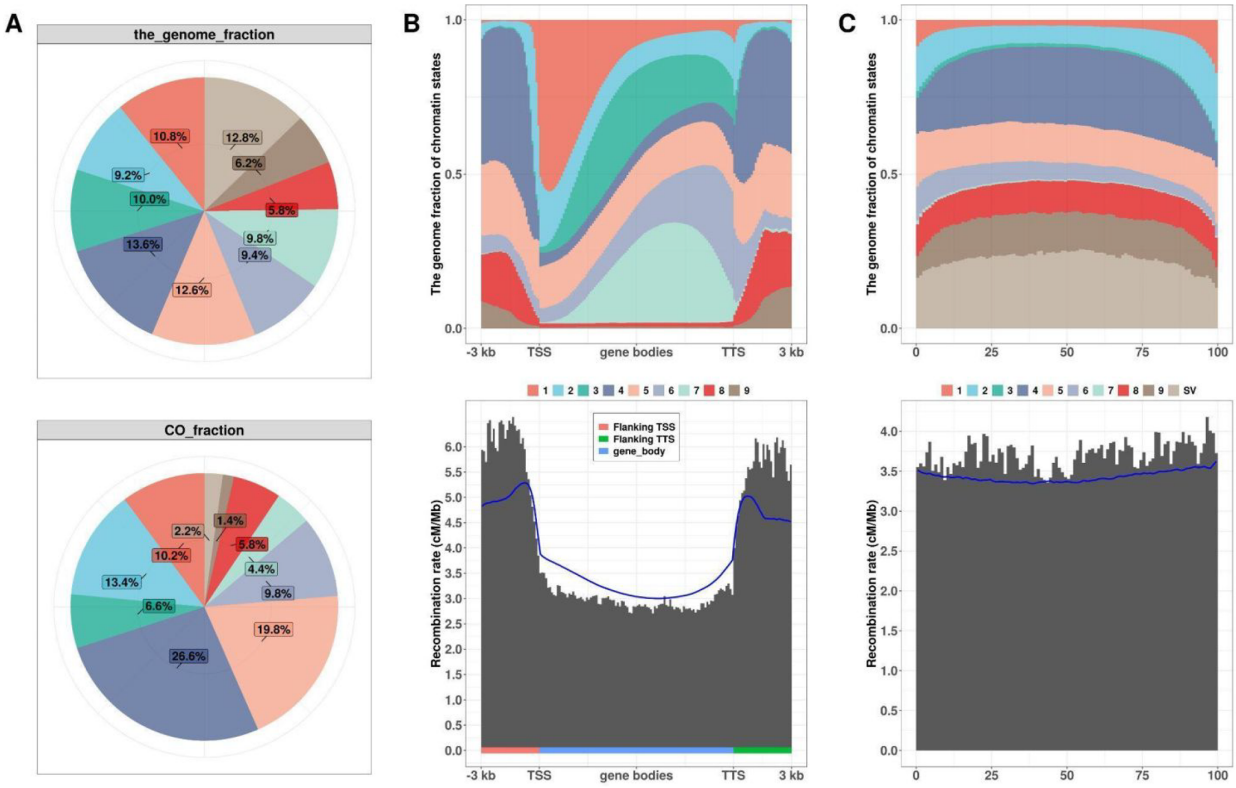


Fig. 2. Relations between our 10 chromatin states, genes, intergenic regions and recombination rate. (a) The top pie chart shows the genome-wide occupation percentages of each of the 10 states. 'SV' refers to low synteny regions or structural variations between Col-0 and Ler. The characteristics of the nine other states are: state 1 (intragenic, transcription starting site (TSS)), state 2 (intergenic, proximal promoter), state 3 (intragenic, coding sequence), state 4 (intergenic, distal promoter), state 5 (intergenic, H3K27me3 rich), state 6 (intergenic, transcription termination site (TTS)), state 7 (intragenic, long genes), state 8 (heterochromatic, AT rich) and state 9 (heterochromatic, GC rich). The lower pie chart shows the percentage of crossover occurrences identified in the 10 states. (b) Two plots, giving respectively the profiles of cumulated fractions of occurrences of the 10 different states (top) and the recombination rate pattern (bottom) in cM per Mb, along gene bodies and their 3-kb flanking regions. In the absence of SV, the entire 3-kb flanking region was used, otherwise it was truncated. The gene body goes from the TSS to the TTS as given in TAIR 10. Only non-transposable element coding genes satisfying the synteny filter have been included in the analysis. For the gene body region, the *x*-axis represents *relative* position, that is the distance from the TSS divided by the distance between TTS and TSS. That procedure allows one to pool genes of different sizes. For the flanking regions, *x*-axis represents position relative to the TSS or TTS in kb. The blue curve at the bottom is the predicted recombination rate when using the chromatin state profiles at the top together with the genome-wide recombination rates derived from (a). (c) Two plots as in (b) but now for the intergenic regions. Again, the blue curve is the predicted recombination rate when using the chromatin state profiles at the top together with the genome-wide recombination rates derived from (a). The legend in the middle of (b) and (c) indicates the corresponding chromatin state of each color used in plotting the chromatin-state profiles.

state-specific recombination rates'. They are to be compared to the genome-wide average recombination rate of 3.3 cM/Mb. As expected, recombination is strongly suppressed in states 9 (pericentromeric heterochromatin) and SV.

In Supplementary Figure S2, we compare experimental recombination rates to those predicted by this minimal 'model'. For instance, when segmenting the genome into bins of size 100 kb, the fraction of the variance in the experimental recombination rates that is explained by the model is R^2 = 0.24. This value is lower than that of the additive model using equation (1) (cf. Supplementary Table S2) but note that when using the experimentally measured state-specific recombination rates there are *no adjustable parameters*. Furthermore, this 'model' based on chromatin states overcomes the defect of predicting negative recombination rates when gene density is high.

3.4. The model with discrete chromatin states predicts fine-scale recombination patterns

Figure 2b (bottom) shows the recombination rate pattern along genes and their 3-kb flanking regions (same syntenic genes and binning methodology as for the top of that figure). Regions just upstream of the TSS are richer in COs than regions downstream of the TTS which themselves are richer than gene bodies. Interestingly, these recombination patterns are quite well-*predicted* by the proportions of each chromatin state (top of Figure 2b) using the experimentally measured state-specific recombination rates as displayed by the continuous blue curve in Figure 2b (bottom). This implies that the determinants of recombination rate are at least partly encoded into our 10 states.

We performed the analogous analysis on intergenic regions as shown in Figure 2c (bottom). Again, the experimental behaviour is well-predicted by our model that assigns one recombination rate to each chromatin state (cf. blue curve).

3.5. Recombination rate is suppressed in small intergenic regions

The profiles and patterns in Figure 2b,c pool gene bodies or intergenic regions, ignoring their sizes. To further test the model, we have considered the possibility that recombination rate patterns might vary as a function of the size of the region. For instance, the content in exons and introns is quite different for small and large genes and so this could potentially affect recombination rates.

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

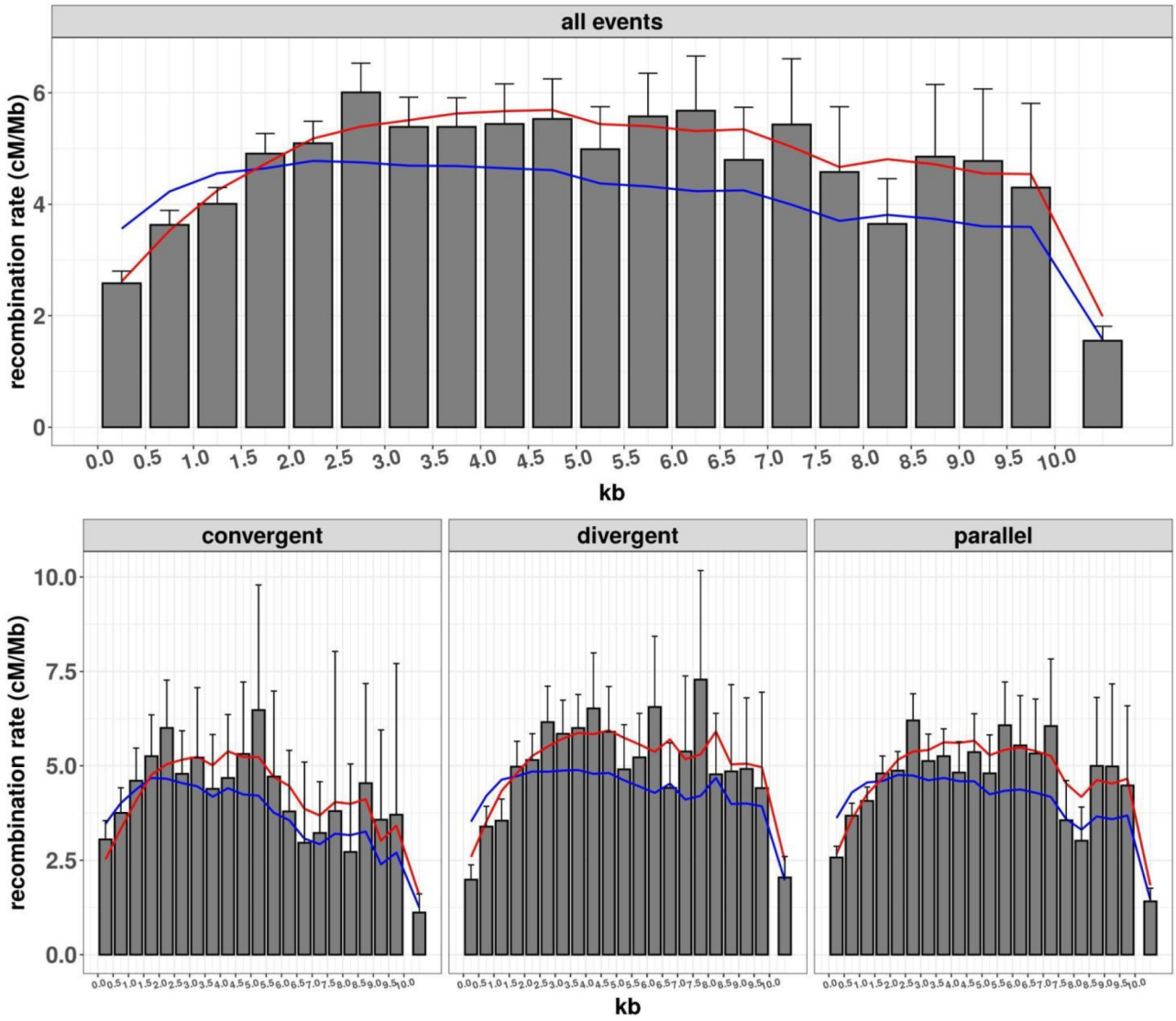


Fig. 3. The relationship between the size of intergenic regions and their average recombination rate. These bar charts were constructed using all intergenic regions, but in the bottom, the regions were divided into three categories according to the transcription orientations of the two flanking genes, corresponding to convergent, divergent and parallel transcriptions. In all cases, the *x*-axis gives the size of the intergenic regions in kb, and the *y*-axis gives the corresponding averaged recombination rate (cM/Mb). Binning of the intergenic region sizes was applied every 500 bases up to a total size of 10 kb. For example, the leftmost bin covers intergenic regions of size 0–0.5 kb. However, we also include a rightmost bar on each chart to cover intergenic regions of sizes larger than 10 kb. Error bars are errors on the mean computed by the jackknife method (only the top segments are displayed). In both top and bottom figures, the blue curves give the predicted recombination rates using the genome-wide recombination rates of the 10 chromatin states as obtained from Figure 2a. The red curves show the predicted recombination rates when one includes the modulation based on the size of the intergenic regions as specified in equation (4).

To study the possible influence of gene body size, we divided the genes into size quantiles and recalculated the corresponding state occurrence profiles and recombination rate patterns. As illustrated in Supplementary Figure S3, gene body size strongly affects chromatin state content. Furthermore, recombination rate patterns become more contrasted as gene size increases, with a concomitant decrease in the average recombination rate. Nevertheless, the model of 10 chromatin states correctly predicts these trends as shown by the blue curves.

The analogous study for intergenic region size is summarised in Supplementary Figures S4–S6, treating separately the three possible orientations of the genes flanking the intergenic region: divergent, convergent and parallel. In contrast to the gene body case, the 10 chromatin state models' predictions (blue curves) are not so good: the model significantly over-estimates the recombination rates when the size of the intergenic region is small.

To quantify this result, consider how the *average* recombination rate within intergenic regions depends on region size. In Figure 3, we display this dependence, for all intergenic regions pooled (top) or separated according to the orientation of their flanking genes (bottom). There is a clear suppression of recombination rate when the size of the intergenic regions is less than 1.5 kb, while beyond 2.5 kb the curves are rather flat, with perhaps a trend to decrease beyond 10 kb. Figure 3 also displays the recombination rates *predicted* when using the 10 states chromatin models. Clearly, the predictions over-estimate the recombination rate when the size of intergenic regions is small, in agreement with the trends seen in Supplementary Figure S4–S6.

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

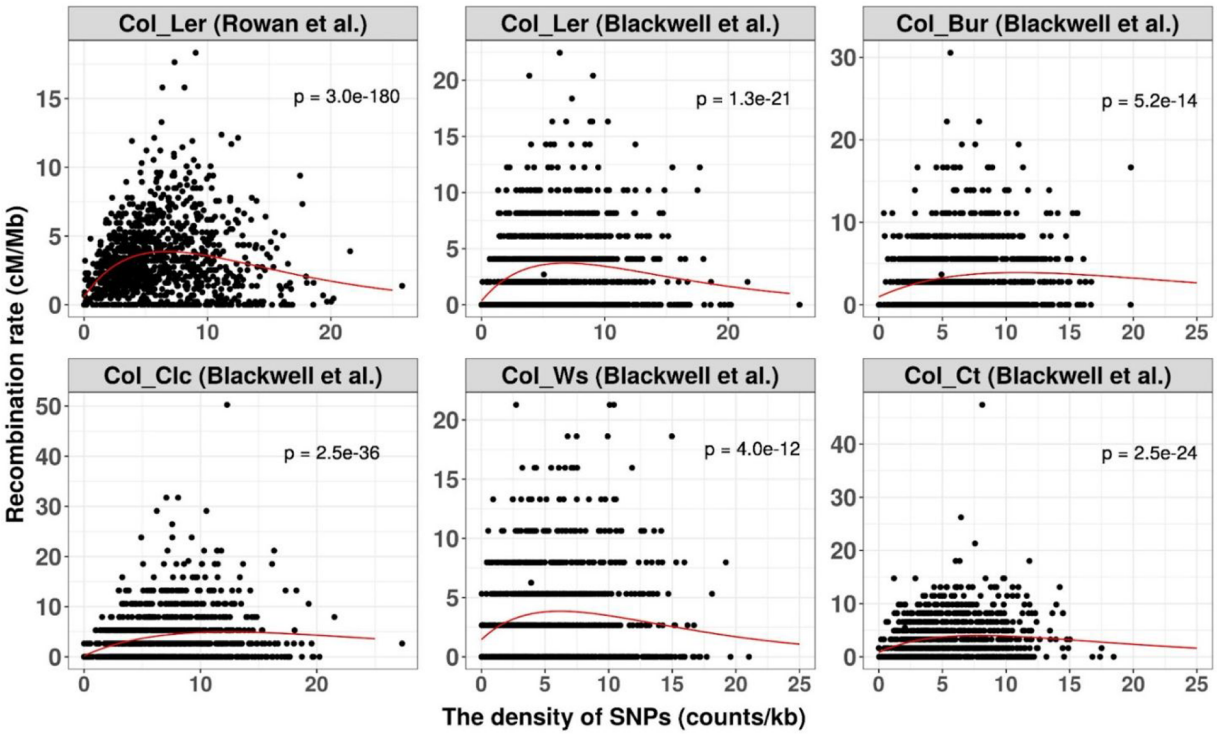


Fig. 4. The relationship between recombination rate and single nucleotide polymorphism (SNP) density. The Col-0 genome was decomposed into bins of 100 kb. For each cross starting with that of Rowan et al. (2019), SNPs and crossovers (COs) were inferred from reads produced using the F_2 populations by mapping to the Col-0 genome. SNP density and recombination rates were then determined for each bin and displayed as a scatter plot. The five additional crosses are from Blackwell et al. (2020). The continuous red curves are fits when using the function $(a + b\,x)\,\exp(-cx)$ so as to maximise the log likelihood. To filter out the high SNP density regions that are expected to causally repress recombination, we restricted the analysis to SNP densities in the first two quantiles. All crosses show a reduced recombination rate at low SNP density and the likelihood ratio test allows us to reject the hypothesis H0 that '$b = 0$', corresponding to no such suppressive effect (p-values shown for each cross and computed using the chi-square distribution with one degree of freedom).

These results motivated us to improve the model by including a modulation effect taking into account the sizes of intergenic regions. We parameterise this modulation by multiplying the recombination rate r_i of a segment in state i by the factor

$$1/\left(\beta_1 + \beta_2\,\exp\left(-\beta_3 \ell\right)\right), \tag{4}$$

whenever the segment lies within an intergenic region of size ℓ kb. The detailed form of this modulation function is not so important, but it should go smoothly from its minimum at $\ell = 0$ to its maximum at large ℓ. The quantities β_1, β_2 and β_3 are free parameters that we can adjust to minimise the deviation between observed and predicted recombination rates over all intergenic regions. The red curves in Figure 3 show the corresponding improved predictions when including this modulation effect.

3.6. Recombination rate is suppressed in regions of low SNP density

A high divergence between homologs suppresses recombination rate, a trend that is visible in the top left of Figure 4, where SNP density is used as a proxy for divergence between homologs. However, we see that *low* SNP density is also associated with reduced recombination. To confirm that this is not an artefact of the Rowan et al. (2019) dataset, we examined five other crosses published by Blackwell et al. (2020) who had found the same effect. The minor differences between our panels and those in their paper come from using different choices in the analysis pipelines: including or not the pericentromeric regions, using a bin size of 100 kb *versus* 1 Mb, applying different filtering criteria to the remapped reads to define SNPs, and forbidding or not the fitting function to have negative values. The important point is that the two independent analyses reach the same conclusion: low SNP density is associated with lower recombination rate (cf. Figure 4).

3.7. Low SNP density may be a causal factor of recombination rate suppression

In natural populations undergoing panmictic reproduction and subject to spontaneous mutations, drift generates linkage disequilibrium depending on recombination rate. Indeed, if a region of the genome has lower than average recombination rate, it will sustain larger haplotypic blocs and so its SNP density will be below average, producing the kind of correlation found in Figure 4. However, *A. thaliana* is a selfer, so linkage disequilibrium and thus the pattern of accumulation of mutations will not be affected by recombination. Specifically, if we consider the most recent common ancestor to Col-0 and Ler, it produced two separate lineages by successive generations of selfings, lineages in which mutations have accumulated *independently*. Under such dynamics, recombination cannot influence SNP density unless recombination itself generates mutations. This last possibility has long been downplayed because homologous recombination was considered to be nearly error-free (Guirouilh-Barbat et al., 2014), but it is now known that CO formation produces mutations in human (Arbeithuber et al., 2015; Halldorsson et al., 2019). In the absence of any such evidence in plants, we formalised as follows a test for the possibility that

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

SNP density influences recombination. We fit each scatter plot of Figure 4 to the function $(a + b\ x)\ \exp(-cx)$ that embodies a suppression effect at low SNP density. Then we compare the likelihood for that fit to the one obtained when the parameter b is set to 0 (corresponding to no suppression at low SNP density). The likelihood ratio test then allows us to reject or not the absence of this suppression effect. In all six populations, the *p*-value shows that the data strongly favours the presence of a suppression. A slightly modified formalisation is tested in the Supplementary Material (cf. Figure S7), reaching the same conclusion.

3.8. A state-based quantitative model with multiple effects modulating recombination rate has good predictive power

Our quantitative model builds on the framework of 10 discrete chromatin states by assigning to each an adjustable base recombination rate, but also by applying three context-dependent multiplicative modulating effects. The first effect is associated with intergenic region size $\boldsymbol{\ell}$: we parameterise the multiplicative modulation via the function $1/(\beta_1 + \beta_2 \exp(-\beta_3 \boldsymbol{\ell}))$, where $\boldsymbol{\ell}$ is the size of the intergenic region in kb. The second effect is associated with SNP density $\boldsymbol{\rho}$: we multiply the recombination rate by $(1 + \alpha_1 \boldsymbol{\rho})\ \exp(-\alpha_2 \boldsymbol{\rho})$. Lastly, at the whole chromosome level, it is known that CO numbers are tightly regulated with the result that genetic lengths do not vary linearly with genome size, especially in species that have chromosomes of very different physical lengths. This regulation presumably arises through both CO 'interference' (COs tend to be well separated) and the obligatory CO (there is at least one CO per bivalent), both of these acting on large rather than fine scales. As a result, the recombination rate of a specific genomic segment can be significantly higher if it belongs to a small chromosome than if it belongs to a large one. To incorporate this chromosome-wide effect, we rescale all predicted recombination rates within a chromosome to enforce its experimentally measured genetic length.

Overall our model has 15 adjustable parameters: the 10 base recombination rates and the 5 additional parameters for the modulation effects (the chromosome-specific rescalings do not require introducing any parameters or fits). To calibrate the resulting quantitative model, we apply the maximum likelihood approach which quantifies the deviation between the model's predicted rates and the experimental ones from Rowan et al. (2019) when using a binning along the genome (see Section 2 for details). In Supplementary Table S3, we provide the AIC and BIC values when the additional parameters are successively included. The minimum value is always reached for the full (highest complexity) model which is why we discuss only that case hereafter. The optimised parameters are provided in Supplementary Table S4 when calibrating over the whole genome using various bin sizes. In Supplementary Figure S8, we compare the predictions of recombination rate in our quantitative model to the experimental ones when using bins sizes ranging from 50 to 500 kb. One can also do the comparison at the level of the recombination landscapes: in Figure 5, we show the predicted and experimental landscapes for chromosome 1 when using bins of size 100 kb (cf. Supplementary Figure S9 for the other chromosomes). We see that the adjusted model reproduces much of the qualitative structure of the landscape. The inset in Figure 5 provides a zoom on a region in the right arm, allowing one to better see the small scale trends. Even for this bin size which is rather large compared to the typical distance between genes, the model and experimental landscapes are far from smooth. Furthermore, both in the inset and in the main part of the figure, we see that though there is quite a lot of concordance between the two curves for local minima and maxima, the model's landscape generally underestimates the observed variance. This is partly due to the experimental landscape being subject to the stochasticity of CO numbers, but it may also point to other determinants that could be missing in our analysis or data.

Finally, to test the *predictive power* of our modelling approach and ensure that it does not introduce overfitting, we also have calibrated the model on one chromosome and then used that calibration to predict recombination on the other chromosomes. Supplementary Table S5 gives the corresponding values of R^2. For comparison, we perform the same test in Supplementary Tables S6 and S7 when using the additive model (equation (1)) or its extension with interactions (equation (3)). Clearly, our model has significantly higher predictive power than those other models.

4. Discussion and conclusions

4.1. Aggregated chromatin states as predictors of recombination rate

The genome-wide distribution of COs is expected to follow largely from the degree to which the double strand break machinery can access the DNA. This will depend of course on the state of the chromatin and indeed many genomic and epigenomic features are empirically found to correlate with recombination rate. Qualitative modelling based on such features allows one to distinguish hot *versus* low recombination regions (Demirci et al., 2018) but quantitative modelling has been limited to frameworks like equations (1) and (3) (Blackwell et al., 2020; Rodgers-Melnick et al., 2015). Unfortunately, the dependence on a feature is typically non-monotonic as displayed in Figure 1. As a result, recombination rate modelling using these features as quantitative variables requires strong non-linearities and leads to an unmanageable combinatorial complexity (cf. the 46 parameters in equation (3)), not to mention problems for interpreting the resulting models and their low prediction power (cf. Supplementary Tables S6 and S7). To overcome this difficulty, we use a classifier approach to automatically aggregate 16 genomic and epigenomic features into discrete classes (Sequeira-Mendes et al., 2014). This defines the starting point of our modelling wherein each position of the genome is considered to be in one of 10 chromatin states. Using the genome-wide recombination rates in each of these 10 states, Figure 2b,c shows that recombination patterns around genes and in intergenic regions are rather well predicted. In particular, near the extremities of genes, this simple modelling leads to enhanced recombination rates, in agreement with experiment (Choi et al., 2013; Kianian et al., 2018; Marand et al., 2017).

4.2. Intergenic region size modulates recombination rate

The simple model using genome-wide recombination rates in each of the 10 states does not adequately predict the suppressed recombination rate in small intergenic regions (cf. Figure 3). This suppression effect could be the consequence of a local context affecting chromatin accessibility for biophysical reasons. A first such reason could be that small intergenic regions are partly hidden from the double strand break machinery by their flanking regions when these are in dense chromatin. A second such reason could be the way chromatin loops are organised in meiosis; if denser chromatin (e.g., containing gene bodies) is preferentially tethered to the base of those loops, it will pull along with it adjacent stretches of open

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

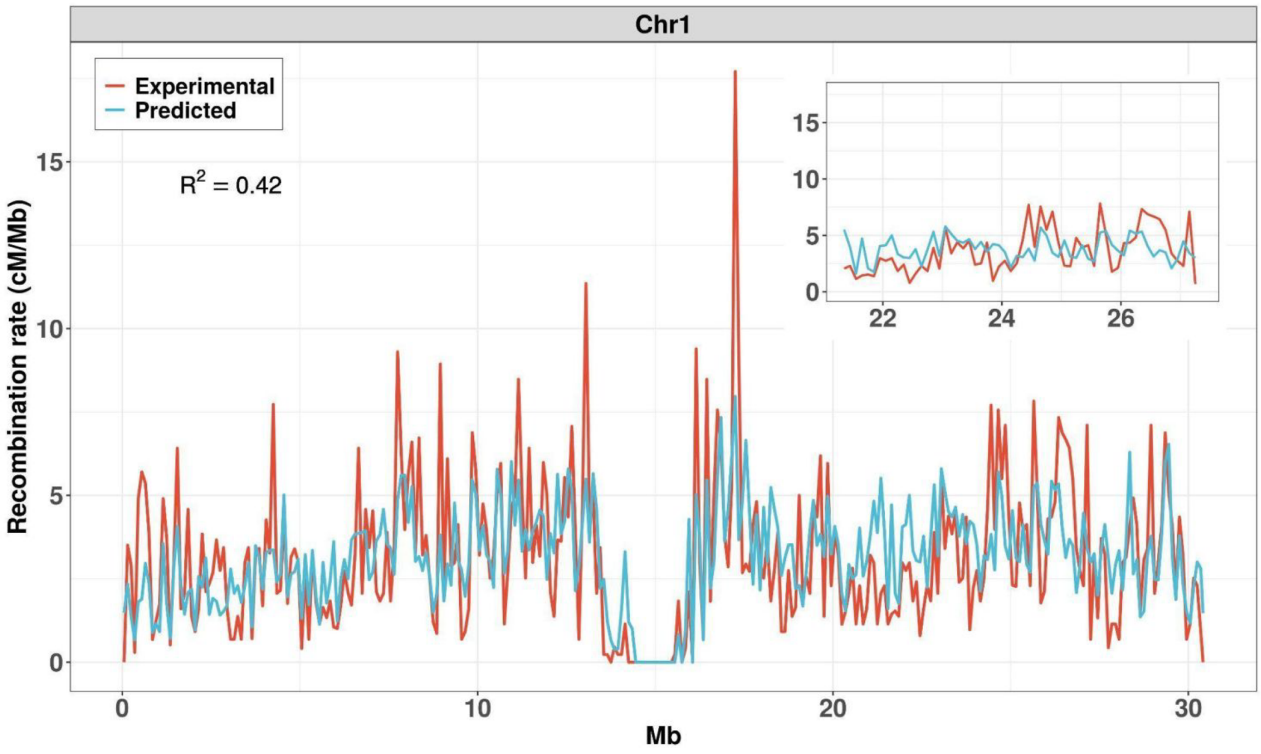


Fig. 5. Experimental and predicted recombination landscapes of chromosome 1. Landscapes using 100 kb bins obtained from the Rowan et al. (2019) dataset (red) and predicted from our calibrated model based on chromatin states (blue) with 15 parameters. Inset: a zoom in the right arm. For landscapes of all chromosomes, see Supplementary Figure S9.

chromatin, hiding these from the double strand break machinery (Tock & Henderson, 2018).

4.3. Lack of any sequence divergence may drive lower recombination rate

The empirical data in multiple crosses show that regions with very low divergence between homologs typically have low recombination rate (cf. Figure 4). That is expected in panmictic populations where recombination shapes linkage disequilibrium and thus SNP density. However, *A. thaliana* is a selfing species with a very low rate of outcrossing of about 2% (Hoffmann et al., 2003; Platt et al., 2010). That leads to low genetic divergence within given habitats which is further exacerbated by adaptive pressures, so recombination in the wild will hardly do any allelic shuffling. We thus argue that our observations from the data in this species might be explained if an absence of divergence between homologs causally suppresses COs. Clearly, such an effect makes sense from an evolutionary perspective: if a genomic region has no underlying sequence diversity, there is little point in producing COs there.

Interestingly, a reduction of recombination rate caused by near perfect sequence homology was demonstrated in three previous works on *A. thaliana*. The oldest such work, by Barth et al. (2001), found that on average homozygous homologs led to fewer COs than heterozygous ones. Second, Ziolkowski et al. (2015) considered a heterozygous block within an otherwise homozygous chromosome and found that CO frequency was enhanced in the heterozygous region. Third, Blackwell et al. (2020) showed that *msh2*, a mutant of mismatch repair, redistributed COs towards regions of lower SNP density, suggesting that, in wild type, CO formation is disadvantaged when sequence homology is perfect. The behaviours found in all these works can be interpreted as a large-scale manifestation of the causal SNP effect we hypothesise.

4.4. A quantitative model of recombination rate with good predictive power

Our full model integrates local genomic and epigenomic features but also context-dependent information. All of its 15 parameters have very direct interpretations and are statistically justified by the AIC and BIC tests (cf. Supplementary Table S3). This model has good predictive power as shown in Supplementary Tables S5–S7 and is able to reproduce much of the variation in rates arising in the recombination landscape (cf. Figure 5 and Supplementary Figure S9). Clearly not all of the variation is captured by our model. First, there is statistical noise inherent to the experimental landscape. Second, although the model predicts major peaks and troughs in the landscape, it tends to underestimate their amplitude. This may suggest a form of competition between sites for recruiting the machinery that produces double strand breaks. There are also other caveats to our modelling. The most obvious one is that because of lack of appropriate data, we had to use measurements of epigenetic marks in Col-0 only and from tissues such as leaf or root rather than from meiocytes. Fortunately, it seems that the epigenetic landscape is largely shared between somatic and germline tissues, the differences being restricted to a small fraction of the genome (Walker et al., 2018). We did a systematic investigation of this point using published data (cf. Supplementary Figure S1) and showed that the epigenomic patterns are surprisingly similar between somatic and germline tissues. Another limitation of our modelling is that it necessarily ignores any sex-dependent differences in recombination landscapes, focussing only on the female–male average. Similarly, we have not explicitly included

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

CO interference or the obligatory CO, we have just incorporated a proxy of their effects *via* chromosome-specific rescalings. Such a choice is in line with the expectation that CO interference and the obligatory CO shape recombination landscapes on large scales (Lloyd & Jenczewski, 2019; Morgan et al., 2021), leaving open the determinants at fine scales. Lastly, but perhaps very importantly, we take no account of the well-known fact that meiotic chromosomes are organised in loops tethered to an axis. This structural aspect of meiotic chromosomes may be important for modulating local recombination rates and it is tempting to conjecture that these loops may be responsible for the large peaks seen in the recombination landscape (cf. Figure 5 and Supplementary Figure S9). Unfortunately, very little is known about these loops, in particular concerning their size, position and variability across genetic backgrounds. Hopefully, these uncertainties will be lifted in the near future, given that standard chromosome conformation capture techniques applied to meiotic cells should provide the required information quite directly.

Acknowledgements

We thank T. Blein, M. Rousseau-Gueutin and P. Sourdille for discussions and we are particularly indebted to M. Grelon who provided multiple feedback on our drafts. We are also grateful to B. Rowan and I. Henderson for sharing their data.

Financial support. GQE – Le Moulon and IPS2 benefit from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). The work of Y-M. Hsu was supported by a PhD grant provided by the Ministry of Education (Taiwan) and Université Paris-Sud/Saclay.

Conflicts of interest. The authors declare no potential conflicts of interest.

Authorship contributions. M.F. and O.M. conceived the study. O.M. designed the modelling. Y.M.H. performed all the bioinformatics and code writing necessary for the subsequent analyses. All authors worked on the analyses and wrote the article.

Data availability statement. This work produced no new data but exploited previously published data (see Section 2, Supplementary Material and Supplementary Table S1). All pipelines and analysis codes are given as Supplementary files.

Supplementary Materials. To view supplementary material for this article, please visit http://doi.org/10.1017/qpb.2021.17.

References

Arbeithuber, B., **Betancourt, A. J.**, **Ebner, T.**, & **Tiemann-Boege, I.** (2015). Crossovers are associated with mutation and biased gene conversion at recombination hotspots. *Proceedings of the National Academy of Sciences of the United States of America*, **112**, 2109–2114. https://doi.org/10.1073/pnas.1416622112

Barth, S., **Melchinger, A. E.**, **Devezi-Savula, B.**, & **Lübberstedt, T.** (2001). Influence of genetic background and heterozygosity on meiotic recombination in *Arabidopsis thaliana*. *Genome*, **44**, 971–978. https://doi.org/10.1139/g01-094

Bauer, E., **Falque, M.**, **Walter, H.**, **Bauland, C.**, **Camisan, C.**, **Campo, L.**, **Meyer, N.**, **Ranc, N.**, **Rincent, R.**, **Schipprack, W.**, **Altmann, T.**, **Flament, P.**, **Melchinger, A. E.**, **Menz, M.**, **Moreno-González, J.**, **Ouzunova, M.**, **Revilla, P.**, **Charcosset, A.**, **Martin, O. C.**, & **Schön, C.-C.** (2013). Intraspecific variation of recombination rate in maize. *Genome Biology*, **14**, R103. https://doi.org/10.1186/gb-2013-14-9-r103

Berardini, T. Z., **Reiser, L.**, **Li, D.**, **Mezheritsky, Y.**, **Muller, R.**, **Strait, E.**, & **Huala, E.** (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. *Genesis*, **53**, 474–485. https://doi.org/10.1002/dvg.22877

Blackwell, A. R., **Dluzewska, J.**, **Szymanska-Lejman, M.**, **Desjardins, S.**, **Tock, A. J.**, **Kbiri, N.**, **Lambing, C.**, **Lawrence, E. J.**, **Bieluszewski, T.**, **Rowan, B.**, **Higgins, J. D.**, **Ziolkowski, P. A.**, & **Henderson, I. R.** (2020). MSH2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. *The EMBO Journal*, **39**, e104858. https://doi.org/10.15252/embj.2020104858

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. H.**, **McVean, G.**, & **Henderson, I. R.** (2013). Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics*, **45**, 1327–1336. https://doi.org/10.1038/ng.2766

Choulet, F., **Alberti, A.**, **Theil, S.**, **Glover, N.**, **Barbe, V.**, **Daron, J.**, **Pingault, L.**, **Sourdille, P.**, **Couloux, A.**, **Paux, E.**, **Leroy, P.**, **Mangenot, S.**, **Guilhot, N.**, **Le Gouis, J.**, **Balfourier, F.**, **Alaux, M.**, **Jamilloux, V.**, **Poulain, J.**, **Durand, C.**, … **Feuillet, C.** (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science*, **345**, 1249721. https://doi.org/10.1126/science.1249721

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. *The Plant Journal*, **95**, 686–699. https://doi.org/10.1111/tpj.13979

Fernandes, J. B., **Séguéla-Arnaud, M.**, **Larchevêque, C.**, **Lloyd, A. H.**, & **Mercier, R.** (2018). Unleashing meiotic crossovers in hybrid plants. *Proceedings of the National Academy of Sciences of the United States of America*, **115**, 2431–2436. https://doi.org/10.1073/pnas.1713078114

Goel, M., **Sun, H.**, **Jiao, W. B.**, & **Schneeberger, K.** (2019). SyRI: Finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biology*, **20**, 277. https://doi.org/10.1186/s13059-019-1911-0

Guirouilh-Barbat, J., **Lambert, S.**, **Bertrand, P.**, & **Lopez, B. S.** (2014). Is homologous recombination really an error-free process? *Frontiers in Genetics*, **5**, 175. https://doi.org/10.3389/fgene.2014.00175

Halldorsson, B. V., **Palsson, G.**, **Stefansson, O. A.**, **Jonsson, H.**, **Hardarson, M. T.**, **Eggertsson, H. P.**, **Gunnarsson, B.**, **Oddsson, A.**, **Halldorsson, G. H.**, **Zink, F.**, **Gudjonsson, S. A.**, **Frigge, M. L.**, **Thorleifsson, G.**, **Sigurdsson, A.**, **Stacey, S. N.**, **Sulem, P.**, **Masson, G.**, **Helgason, A.**, **Gudbjartsson, D. F.**, … **Stefansson, K.** (2019). Characterizing mutagenic effects of recombination through a sequence-level genetic map. *Science*, **363**, eaau1043. https://doi.org/10.1126/science.aau1043

Hoffmann, M. H., **Bremer, M.**, **Schneider, K.**, **Burger, F.**, **Stolle, E.**, & **Moritz, G. B.** (2003). Flower visitors in a natural population of *Arabidopsis thaliana*. *Plant Biology*, **5**, 491–494. https://doi.org/10.1055/s-2003-44784

HunterN. (2015). Meiotic recombination: The essence of heredity. *Cold Spring Harbor Perspectives in Biology*, **7**, a016618. https://doi.org/10.1101/cshperspect.a016618

Jiao, W. B., & **Schneeberger, K.** (2020). Chromosome-level assemblies of multiple Arabidopsis genomes reveal hotspots of rearrangements with altered evolutionary dynamics. *Nature Communications*, **11**, 989. https://doi.org/10.1038/s41467-020-14779-y

Jones, G. H., & **Franklin, F. C.** (2006). Meiotic crossing-over: Obligation and interference. *Cell*, **126**, 246–248. https://doi.org/10.1016/j.cell.2006.07.010

Keeney, S., & **Neale, M. J.** (2006). Initiation of meiotic recombination by formation of DNA double-strand breaks: Mechanism and regulation. *Biochemical Society Transactions*, **34**, 523–525. https://doi.org/10.1042/BST0340523

Khil, P. P., **Smagulova, F.**, **Brick, K. M.**, **Camerini-Otero, R. D.**, & **Petukhova, G. V.** (2012). Sensitive mapping of recombination hotspots using sequencing-based detection of ssDNA. *Genome Research*, **22**, 957–965. https://doi.org/10.1101/gr.130583.111

Kianian, P., **Wang, M.**, **Simons, K.**, **Ghavami, F.**, **He, Y.**, **Dukowic-Schulze, S.**, **Sundararajan, A.**, **Sun, Q.**, **Pillardy, J.**, **Mudge, J.**, **Chen, C.**, **Kianian, S. F.**, & **Pawlowski, W. P.** (2018). High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nature Communications*, **9**, 2370. https://doi.org/10.1038/s41467-018-04562-5

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics (Oxford, England), **27**(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

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

Lloyd, A., & **Jenczewski, E.** (2019). Modelling sex-specific crossover patterning in Arabidopsis. *Genetics*, **211**, 847–859. https://doi.org/10.1534/genetics.118.301838

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.**, **Buell, C. R.**, & **Jiang, J.** (2017). Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biology*, **18**, 203. https://doi.org/10.1186/s13059-017-1326-8

Mercier, R., **Mézard, C.**, **Jenczewski, E.**, **Macaisne, N.**, & **Grelon, M.** (2015). The molecular biology of meiosis in plants. *Annual Review of Plant Biology*, **66**, 297–327. https://doi.org/10.1146/annurev-arplant-050213-035923

Morgan, C., **Fozard, J. A.**, **Hartley, M.**, **Henderson, I. R.**, **Bomblies, K.**, & **Howard, M.** (2021). Diffusion-mediated HEI10 coarsening can explain meiotic crossover positioning in Arabidopsis. *Nature Communications*, **12**, 4674. https://doi.org/10.1038/s41467-021-24827-w

Pan, J., **Sasaki, M.**, **Kniewel, R.**, **Murakami, H.**, **Blitzblau, H. G.**, **Tischfield, S. E.**, **Zhu, X.**, **Neale, M. J.**, **Jasin, M.**, **Socci, N. D.**, **Hochwagen, A.**, & **Keeney, S.** (2011). A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation. *Cell*, **144**, 719–731. https://doi.org/10.1016/j.cell.2011.02.009

Platt, A., **Horton, M.**, **Huang, Y. S.**, **Li, Y.**, **Anastasio, A. E.**, **Mulyati, N. W.**, **Ågren, J.**, **Bossdorf, O.**, **Byers, D.**, **Donohue, K.**, **Dunning, M.**, **Holub, E. B.**, **Hudson, A.**, **Le Corre, V.**, **Loudet, O.**, **Roux, F.**, **Warthmann, N.**, **Weigel, D.**, **Rivero, L.**, … **Borevitz, J. O.** (2010). The scale of population structure in *Arabidopsis thaliana*. *PLOS Genetics*, **6**(2), e1000843. https://doi.org/10.1371/journal.pgen.1000843

Pratto, F., **Brick, K.**, **Khil, P.**, **Smagulova, F.**, **Petukhova, G. V.**, & **Camerini-Otero, R. D.** (2014). DNA recombination. Recombination initiation maps of individual human genomes. *Science*, **346**, 1256442. https://doi.org/10.1126/science.1256442

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. *Proceedings of the National Academy of Sciences*, **112**, 3823. https://doi.org/10.1073/pnas.1413864112

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. https://doi.org/10.1534/genetics.119.302406

Sequeira-Mendes, J., **Aragüez, I.**, **Peiró, R.**, **Mendez-Giraldez, R.**, **Zhang, X.**, **Jacobsen, S. E.**, **Bastolla, U.**, & **Gutierrez, C.** (2014). The functional topography of the Arabidopsis genome is organized in a reduced number of linear motifs of chromatin states. *The Plant Cell*, **26**, 2351–2366. https://doi.org/10.1105/tpc.114.124578

Tock, A. J., & **Henderson, I. R.** (2018). Hotspots for initiation of meiotic recombination. *Frontiers in Genetics*, **9**, 521. https://doi.org/10.3389/fgene.2018.00521

Villeneuve, A. M., & **Hillers, K. J.** (2001). Whence meiosis? *Cell*, **106**, 647–650. https://doi.org/10.1016/S0092-8674(01)00500-1

Walker, J., **Gao, H.**, **Zhang, J.**, **Aldridge, B.**, **Vickers, M.**, **Higgins, J. D.**, & **Feng, X.** (2018). Sexual-lineage-specific DNA methylation regulates meiosis in Arabidopsis. *Nature Genetics*, **50**, 130–137. https://doi.org/10.1038/s41588-017-0008-5

Wickham, H., **Averick, M.**, **Bryan, J.**, **Chang, W.**, **McGowan, L.**, **François, R.**, **Grolemund, G.**, **Hayes, A.**, **Henry, L.**, **Hester, J.**, **Kuhn, M.**, **Pedersen, T.**, **Miller, E.**, **Bache, S.**, **Müller, K.**, **Ooms, J.**, **Robinson, D.**, **Seidel, D.**, **Spinu, V.**, … **Yutani, H.** (2019). Welcome to the tidyverse. *Journal of Open Source Software*, **4**, 1686. https://doi.org/10.21105/joss.01686

Zickler, D., & **Kleckner, N.** (2016). A few of our favorite things: Pairing, the bouquet, crossover interference and evolution of meiosis. *Seminars in Cell & Developmental Biology*, **54**, 135–148. https://doi.org/10.1016/j.semcdb.2016.02.024

Ziolkowski, P. A., **Berchowitz, L. E.**, **Lambing, C.**, **Yelina, N. E.**, **Zhao, X.**, **Kelly, K. A.**, **Choi, K.**, **Ziolkowska, L.**, **June, V.**, **Sanchez-Moran, E.**, **Franklin, C.**, **Copenhaver, G. P.**, & **Henderson, I. R.** (2015). Juxtaposition of heterozygous and homozygous regions causes reciprocal crossover remodelling via interference during Arabidopsis meiosis. *ELife*, **4**, e03708. https://doi.org/10.7554/eLife.03708

https://doi.org/10.1017/qpb.2021.17 Published online by Cambridge University Press

Quantitative modeling of fine-scale variations in the *Arabidopsis thaliana* crossover landscape

Yu-Ming Hsu, Matthieu Falque and Olivier C. Martin

Supplementary Material

epigenomic feature	Sample accession or series accession number	tissue	reference
H3K4me1	GSM3674621	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM4668649	seedlings	Niu et al., 2021
	GSM4609829	root non- hair cells	missing
	GSM4785549	inflorescence	Liu et al., 2021
	E-MTAB-7370	unopened flower buds	Lambing et al., 2020
H3K4me3	GSM3674620	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM4154769	seedlings	Liu et al., 2020
	GSM2210857	roots	Yen et al., 2017
	GSM4785552	inflorescence	Liu et al., 2021
	GSE120664	sperm nuclei	Borg et al., 2020
H3K9me2	GSM4734580	leaves	Wang et al., 2021
	GSM3040062	10-day seedlings	Ma et al., 2018
	GSM4422529	mature embryos	Parent et al., 2021
	GSM4818168	flowers	Feng et al., 2020

	E-MTAB-7370	unopened flower buds	Lambing et al., 2020
H3K27me3	GSM3674617	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM3617717	seedlings	Shu et al., 2021
	GSM2210865	roots	Yen et al., 2017
	GSM4785573	inflorescences	Liu et al., 2021
	GSE120664	sperm nuclei	Borg et al., 2020
ATAC	GSM3674715	leaves	Lu et al., 2019; Crisp et al., 2020
	GSM2719200	stem cells	Sijacic et al., 2018
	GSM2719204	mesophyll cells	
	GSM3498708	flowers	Potok et al., 2019
	GSE155344	microspores	Borg et al., 2021
DNase	GSM1289358	seedlings	Sullivan et al., 2014; Sullivan et al., 2019
	GSM1289374	whole roots	
	GSM1289378	seed coats	
	GSM1289380	open flowers	
	GSM1289381	unopened flower	

Supplementary Table S1. Origin and description of datasets for the 6 epigenomic features used in this study.

	intercept (a_0)	gene (a_1)	TE (a_2)	TSS (a_3)	H3K4me1 (a_4)	H3K4me3 (a_5)	H3K9me2 (a_6)	H3K27me3 (a_7)	ATAC (a_8)	DNase (a_9)	R^2
50kb	1.56**	-3.6***	-1.67**	0.17	-0.04	0.05	-0.004***	0.11***	0.65***	0.006*	0.28
100kb	1.00	-5.02***	-1.14	0.26	-0.07	0.16*	-0.01***	0.14**	0.71***	-0.005	0.36
200kb	0.06	-4.44*	0.23	0.3	-0.09	0.16	-0.01**	0.14	0.75***	-0.0007	0.42
500kb	-1.01	-5.82	1.08	0.27	-0.08	0.24	-0.01	0.16	0.83***	0.003	0.50

Supplementary Table S2. Adjusted parameters and R^2 values for the additive model when using different bin sizes. The 9 successive features are those in Fig. 1 (ordered left to right and top to bottom). Parameter values were obtained using the lm() function in R. *, ** and *** correspond to parameters having *p*-values less than 0.05, 0.01 and 0.001 respectively for the hypothesis that the true value of the parameter vanishes. The first column gives the bin size used for each fit. Note that the statistical noise intrinsic to CO formation inevitably drives R^2 (last column, *cf*. Eq. 2 in Main) downward as bin size decreases.

bin size (kb)	AIC	BIC	R^2	Model considered
50	247310	247367.8	0.33	10 states
50	247214.7	247289.8	0.34	10 states + IR
50	246531.8	246618.4	0.39	10_states + IR + SNP
50	246459.3	246545.9	0.4	10_states + IR + SNP + rescaling
100	224047.6	224098.4	0.41	10 states
100	223974	224040	0.43	10 states + IR
100	223515.7	223592	0.48	10_states + IR + SNP
100	223444.3	223520.6	0.49	10_states + IR + SNP + rescaling
200	201007.7	201051.7	0.49	10 states
200	200953	201010.2	0.5	10 states + IR
200	200670.5	200736.4	0.54	10_states + IR + SNP
200	200590.1	200656	0.56	10_states + IR + SNP + rescaling
500	170023	170057.8	0.58	10 states
500	170017.7	170062.9	0.59	10 states + IR
500	169754	169806.2	0.64	10_states + IR + SNP
500	169681	169733.2	0.66	10_states + IR + SNP + rescaling

Supplementary Table S3. Model selection *via* AIC and BIC values. For each of the different bin sizes, we consider the sequence of models of increasing complexity, starting with the 10 parameters for the 10 states, adding to that the 3 parameters for the IR size effect, adding to that the 2 parameters for the SNP effect, and finally adding the rescaling (no additional parameters). The AIC and BIC approaches penalize the goodness of fit measure by an amount that depends on the number of parameters. Using a more complex model (with more parameters) is only justified if the associated criterion (AIC or BIC) is lower. The table shows that the data drives one to use the full model having 15 parameters and scaling.

name	50kb	100kb	200kb	500kb
r_state1	1.367	1.199	1.663	0.984
r_state2	1.908	1.998	2.457	1.965
r_state3	5.43E-09	5.95E-09	5.52E-09	4.95E-09
r_state4	1.822	1.832	2.54	1.926
r_state5	0.713	0.804	1.397	0.809
r_state6	0.328	5.95E-09	5.52E-09	4.95E-09
r_state7	5.43E-09	5.95E-09	5.52E-09	4.95E-09
r_state8	1.325	1.538	2.481	1.782
r_state9	0.007	0.002	5.52E-09	4.95E-09
r_SV	0.009	0.008	0.007	0.001
α_1	1.087	0.948	0.774	1.008
α_2	0.087	0.085	0.087	0.082
β_1	0.513	0.452	0.542	0.487
β_2	7.218	7.63	12.743	2.708
β_3	2.998	3.068	2.245	1.554
R^2	0.403	0.488	0.563	0.657

Supplementary Table S4. Parameter values after calibration of the quantitative model having 15 parameters when using bin sizes from 50 to 500 kb. In the column “name”, r_state1 to r_SV refer to the “base recombination rate” for each of the 10 chromatin states, α_1 and α_2 (respectively β_1, β_2. β_3) refer to the parameters in the SNP (respectively intergenic-region size) modulation effect, and finally R^2 refers to the fraction of the variance explained by the model (*cf*. Eq. 2 in Main).

	Chr1 (fit)	Chr2 (fit)	Chr3 (fit)	Chr4 (fit)	Chr5 (fit)
Chr1 (predict)	0.463	0.299	0.347	0.297	0.438
Chr2 (predict)	0.403	0.502	0.448	0.48	0.434
Chr3 (predict)	0.523	0.556	0.607	0.534	0.56
Chr4 (predict)	0.426	0.472	0.473	0.54	0.466
Chr5 (predict)	0.453	0.376	0.41	0.374	0.473

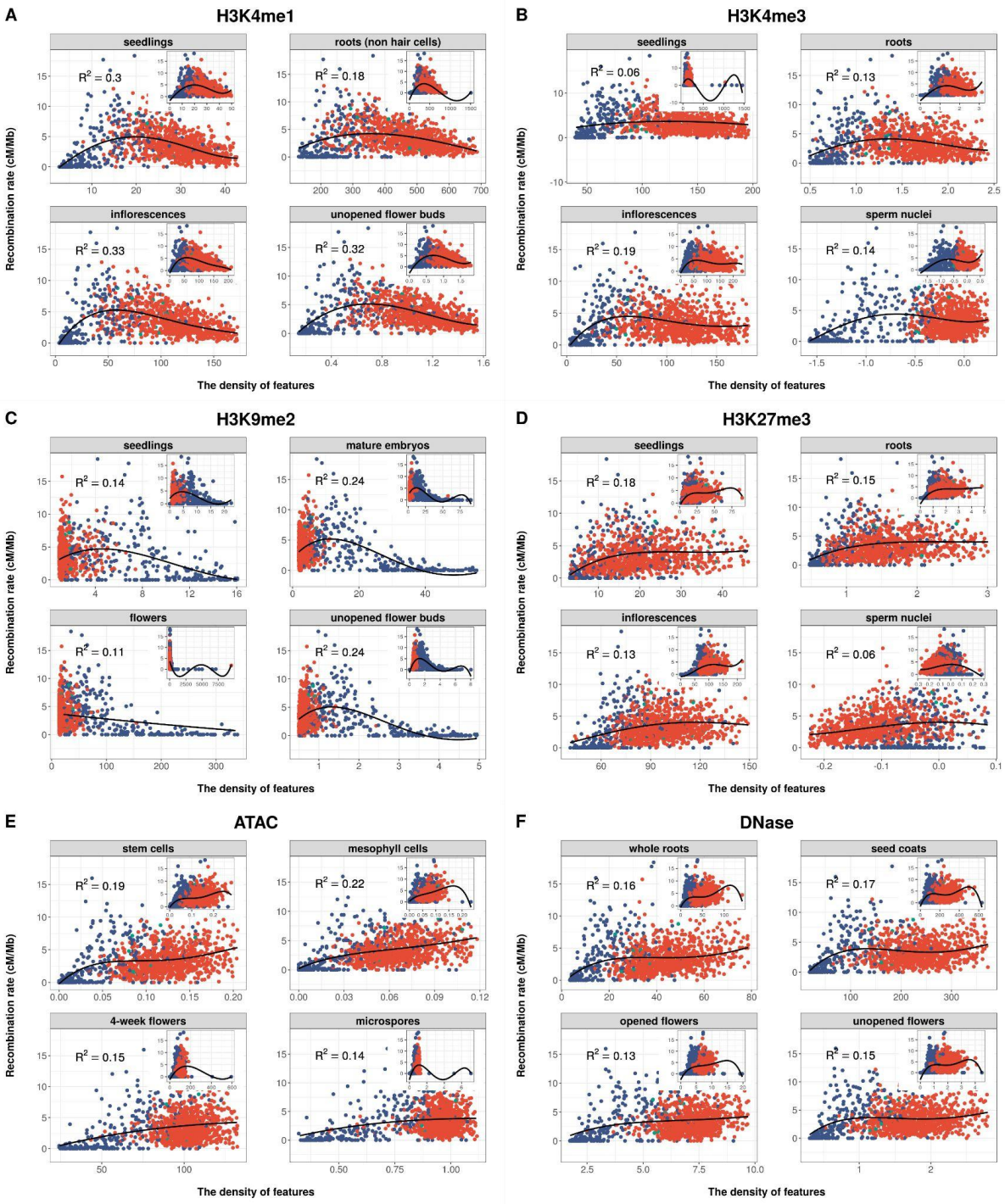
Supplementary Table S5. Predictive power of the model with 15 parameters. We provide the R^2 values when using one chromosome (that labeled by the considered column) to fit the 15 parameters and then apply that calibrated model to predict recombination landscapes of all 5 chromosomes. The genome has been segmented into bins of size 100 kb. Note that in each row the largest R^2 value must occur for the chromosome that has been used to do the fitting of parameters. Omitting the R^2 values produced by the calibrations (on the diagonal), the average R^2 of the predictions (remaining 20 values) is 0.427.

	Chr1 (fit)	Chr2 (fit)	Chr3 (fit)	Chr4 (fit)	Chr5 (fit)
Chr1 (predict)	0.348	0.222	0.22	0.171	0.292
Chr2 (predict)	0.211	0.409	0.263	0.344	0.339
Chr3 (predict)	0.138	0.35	0.455	0.353	0.383
Chr4 (predict)	0.218	0.347	0.34	0.383	0.328
Chr5 (predict)	0.281	0.218	0.274	0.215	0.346

Supplementary Table S6. Predictive power of the additive model (Eq. 1) with 10 parameters exploiting the genomic and epigenomic features of Fig. 1. We provide the R^2 values when using one chromosome (that labeled by the considered column) to fit the 10 parameters and then apply that calibrated model to predict recombination landscapes of all 5 chromosomes (same procedure as in Supplementary Table S5, again with bins of size 100 kb). Omitting the R^2 values produced by the calibrations (on the diagonal), the average R^2 of the predictions (remaining 20 values) is 0.275.

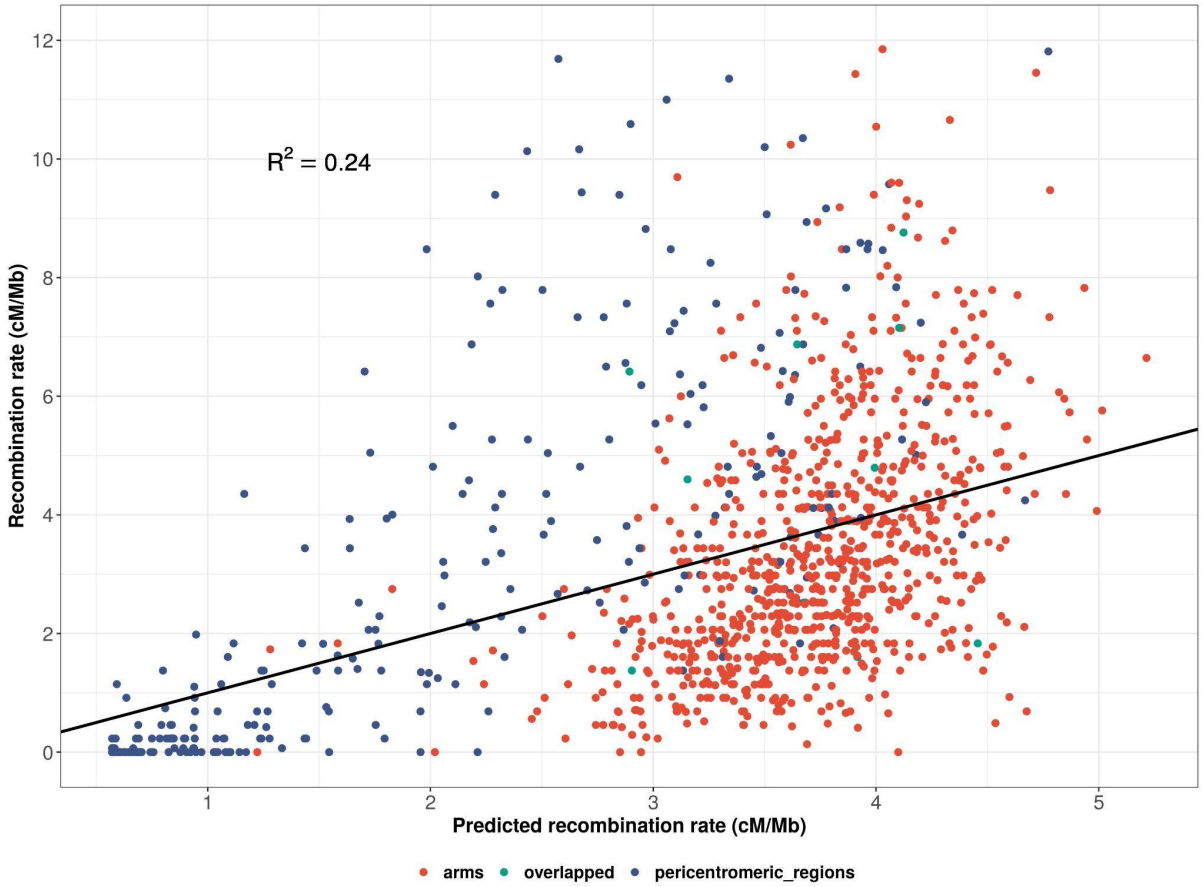
	Chr1_fit	Chr2_fit	Chr3_fit	Chr4_fit	Chr5_fit
Chr1_predict	0.447	-1.364	-0.493	-0.407	0.176
Chr2_predict	-0.299	0.579	-0.614	-39.286	-8.073
Chr3_predict	-0.307	-78.667	0.568	-39.829	-2.22
Chr4_predict	0.074	-17.86	0.001	0.545	-0.349
Chr5_predict	-0.3	-27.968	-1.393	-2.783	0.501

Supplementary Table S7. Predictive power of the model with interactions (Eq. 3) with 46 parameters exploiting the genomic and epigenomic features of Fig. 1. We provide the R^2 values when using one chromosome (that labeled by the considered column) to fit the 46 parameters and then apply that calibrated model to predict recombination landscapes of all 5 chromosomes (same procedure as in Supplementary Table S5, again with bins of size 100 kb). Note that the R^2 of most of the predictions are negative, showing that this model with interactions has no predictive power, presumably because it strongly overfits the data during calibration.

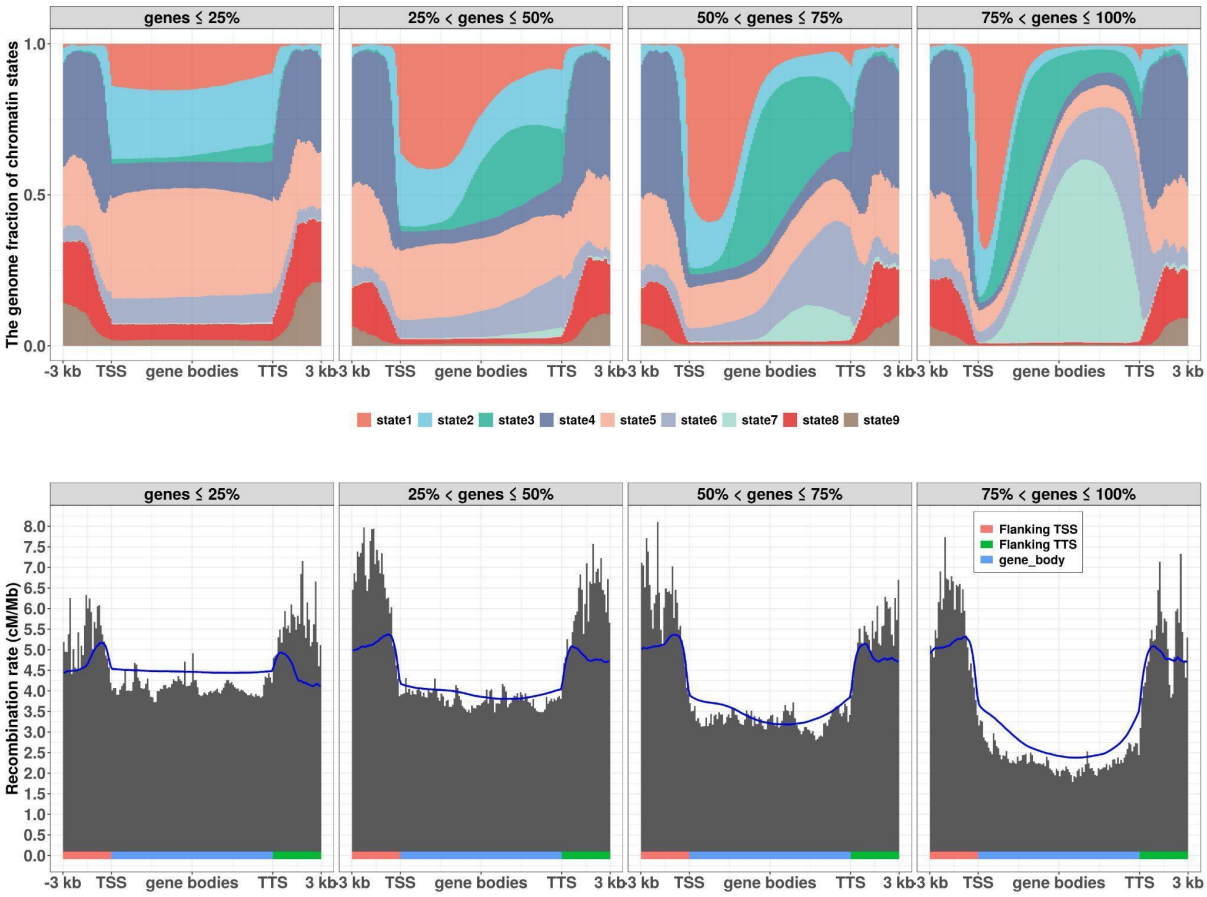


Supplementary Figure S1. The correlations between recombination rate and six epigenomic features when measured in somatic vs. germinal tissues. From (A) to (F), each sub figure combines four plots using data from two somatic and two germinal tissues for the same epigenomic feature. The subtitle on each plot indicates the corresponding tissue. Each dot represents the values for a 100-kb bin. The x-axis values correspond to the density of peaks or reads of each feature according to the format of raw data downloaded from NCBI or ArrayExpress databases. The y-axis gives the associated recombination rate based on a total of 17,077 crossovers from the Col-0-Ler F_2 population. As in Fig. 1 of Main, curves show the fits using a

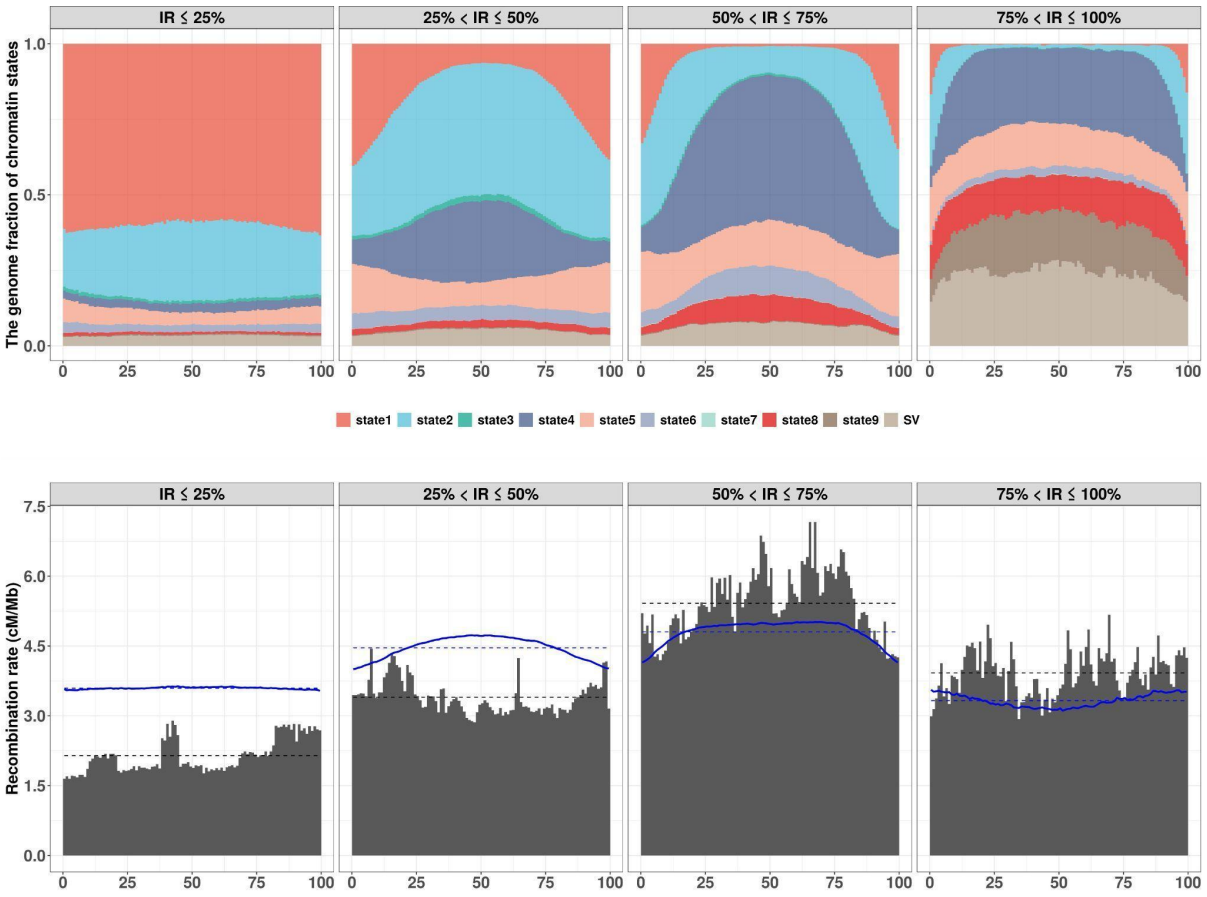
polynomial of degree 4 over the full data range from which the R^2 values are calculated. The main part of each panel corresponds to a zoom of the inset to show greater detail in the main part of the scatter plot.



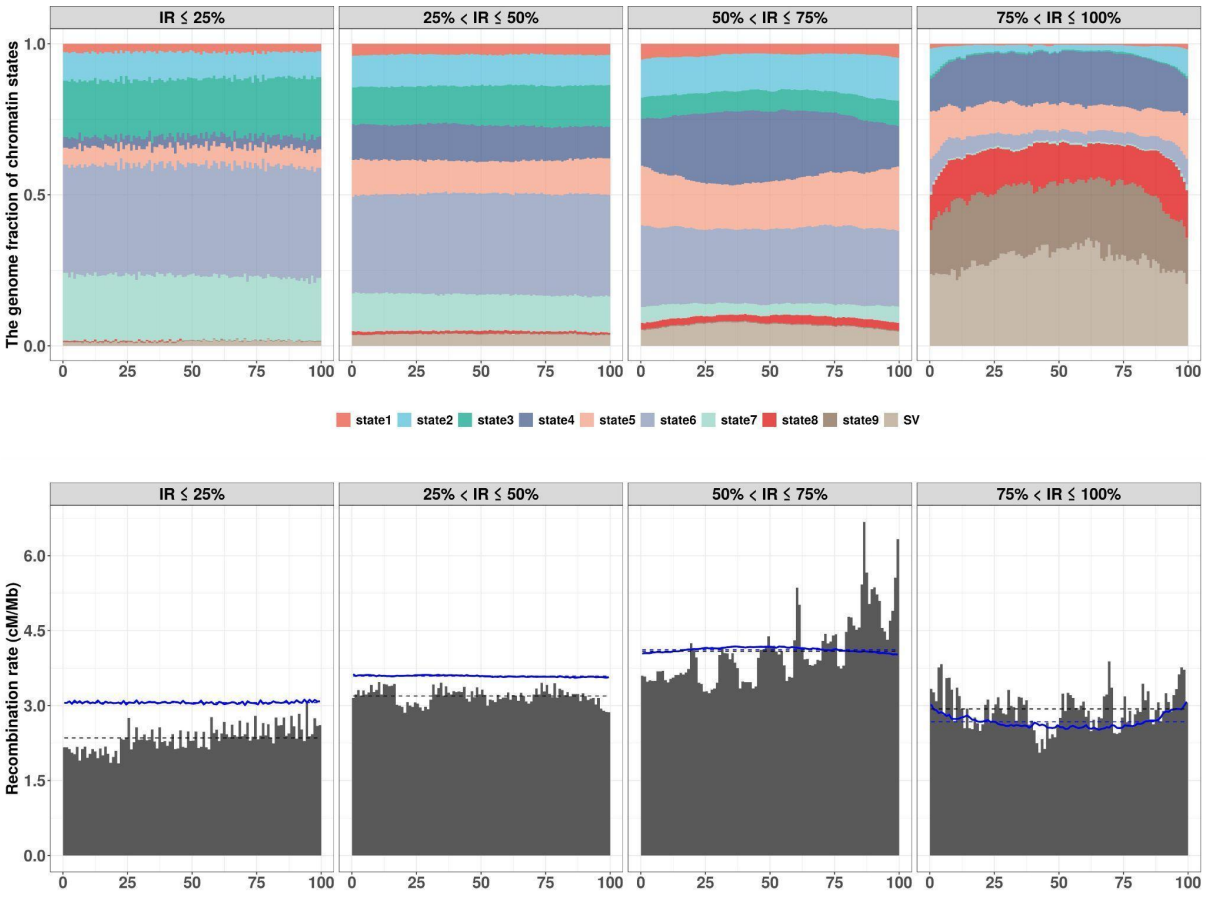
Supplementary Figure S2. Comparison of experimental and predicted recombination rates. Here the predictions are those of the 10 chromatin states model using the experimentally measured state-specific recombination rates (no adjustable parameters). Each data point is associated with a bin of 100 kb along the genome. The fraction of variance explained by the model (computed using the deviations from the predicted recombination rates) is $R^2 = 0.24$.



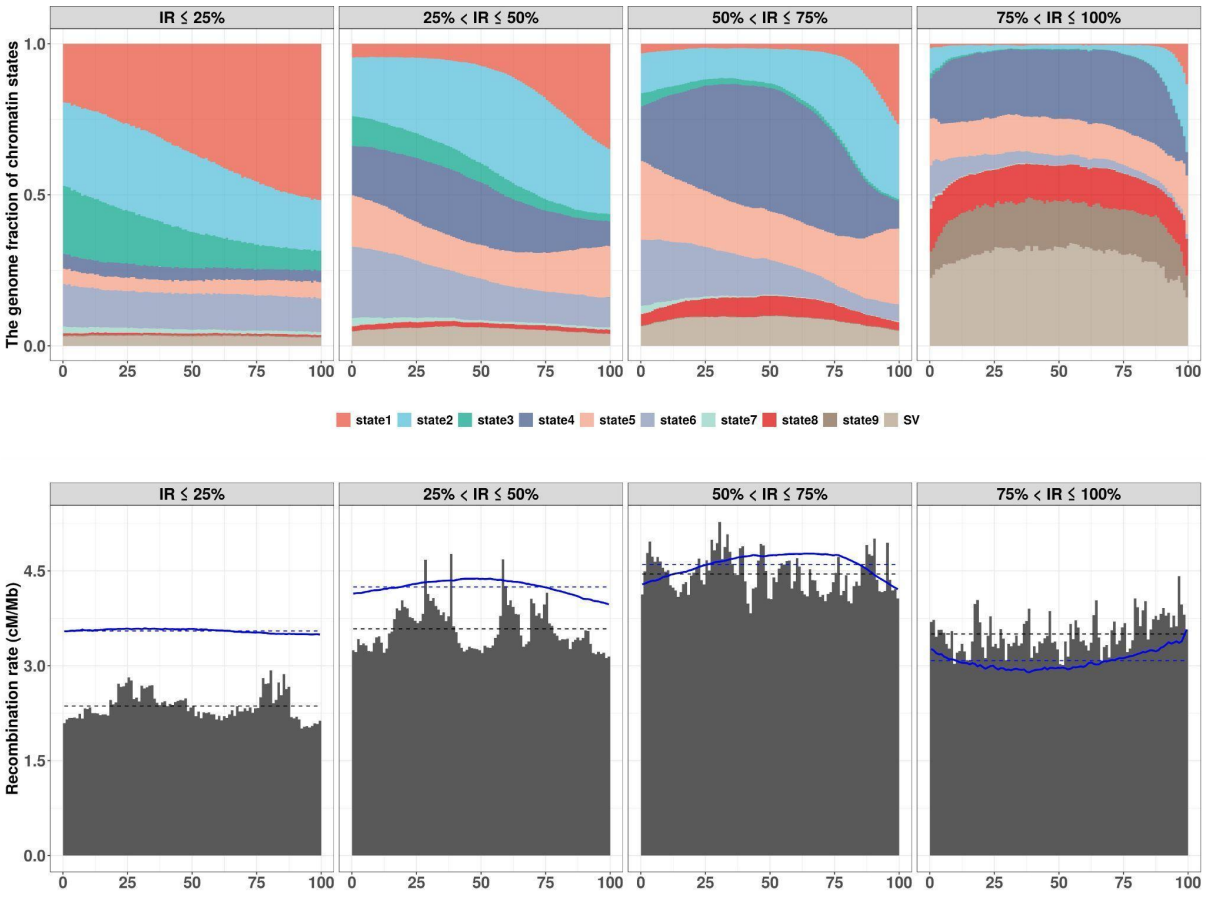
Supplementary Figure S3. Dependence of recombination patterns on gene body size. The profiles of chromatin states and the recombination rate patterns are determined separately in the four quantiles of gene body size. The procedures are the same as in Fig 2B, and the blue curve shows the prediction of the model with 10 chromatin states when using the experimentally measured state-specific recombination rates (no adjustable parameters). The predictions of the model follow the experimental values rather well.



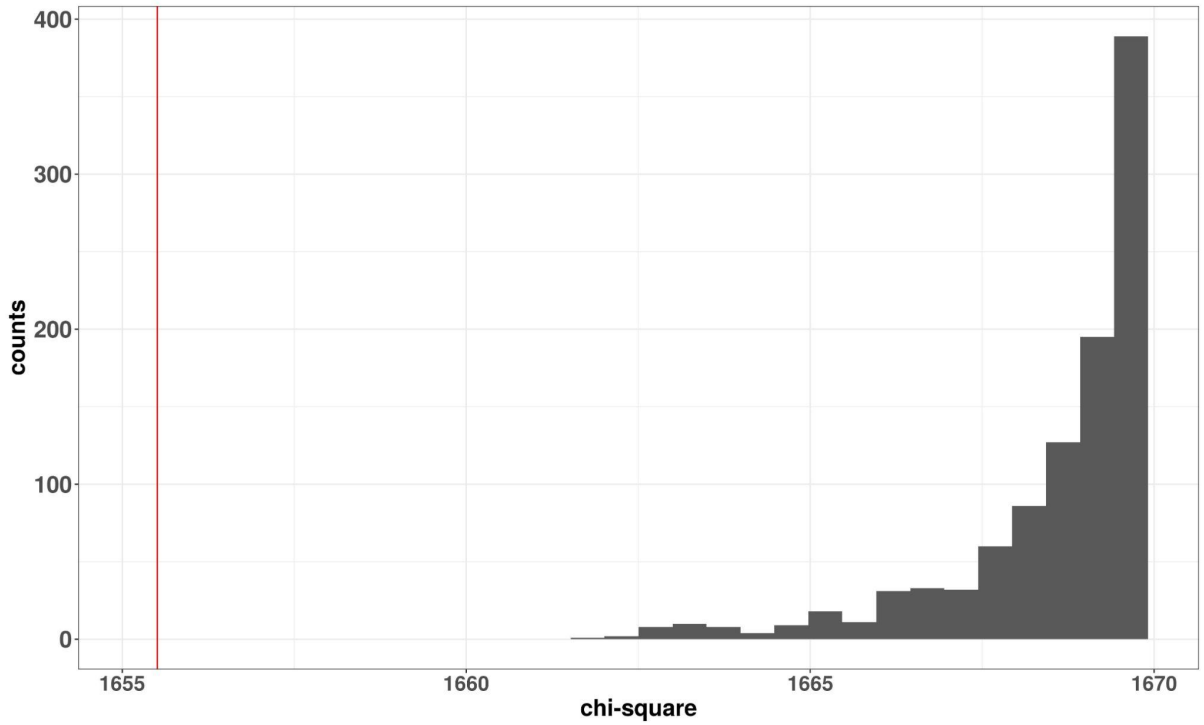
Supplementary Figure S4. The profiles of chromatin states and recombination rate in intergenic regions between genes of divergent orientation. All “divergent” intergenic regions larger than 100 base pairs are divided into 4 groups depending on their size, and each group has one quantile (25 %) of intergenic-region events. In each group, we segmented every intergenic region into 100 bins, then pooled all data of each bin, and calculated the fraction of 9 chromatin states and SVs and the recombination rate of each bin. In the top of this figure we show the fraction of states on the y-axis while the x-axis gives the relative position using 100 bins. At the bottom of this figure, the y-axis corresponds to the recombination rate, while the x-axis is as above. The bottom histograms show the experimental recombination rate in the 100 bins, the black dashed line giving the corresponding average. The procedures are the same as in Fig 2B. The continuous blue curve shows the prediction of the model with 10 chromatin states when using the experimentally measured state-specific recombination rates (no adjustable parameters). The blue dashed line is the corresponding average. The predictions of the model systematically overestimate recombination rates in the small intergenic regions.



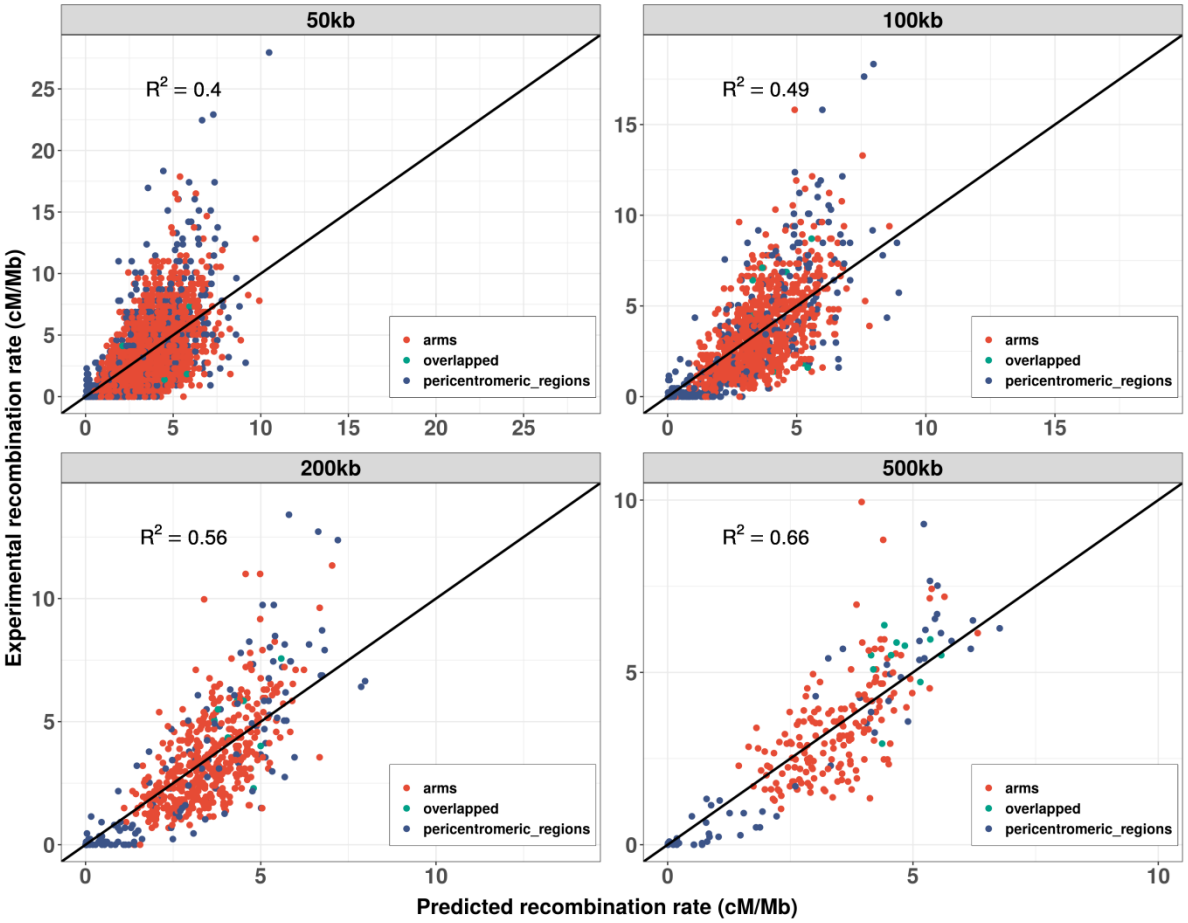
Supplementary Figure S5. The profiles of chromatin states and patterns of recombination rate in intergenic regions between genes of convergent orientation. The procedures and quantities displayed are as in Supplementary Figure S4. The predictions of the model systematically overestimate recombination rates in the small intergenic regions.



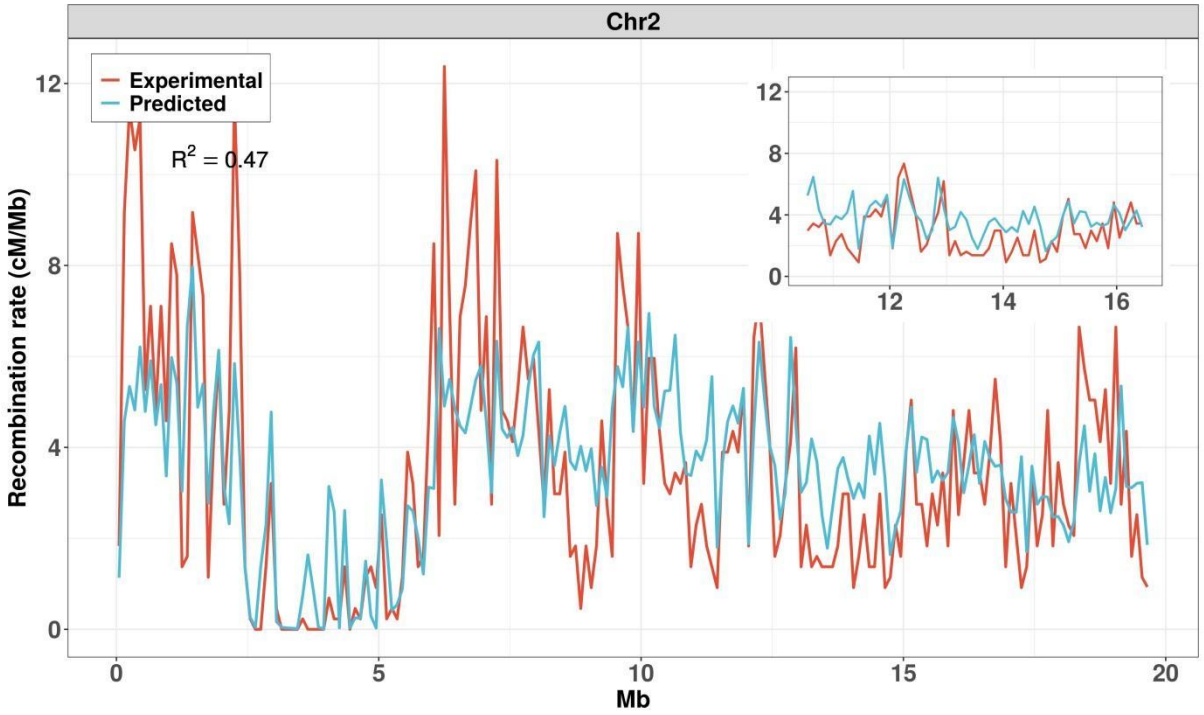
Supplementary Figure S6. The profiles of chromatin states and recombination rate in intergenic regions between genes of parallel orientation. The procedures and quantities displayed are as in Supplementary Figure S4. The predictions of the model systematically overestimate recombination rates in the small intergenic regions.

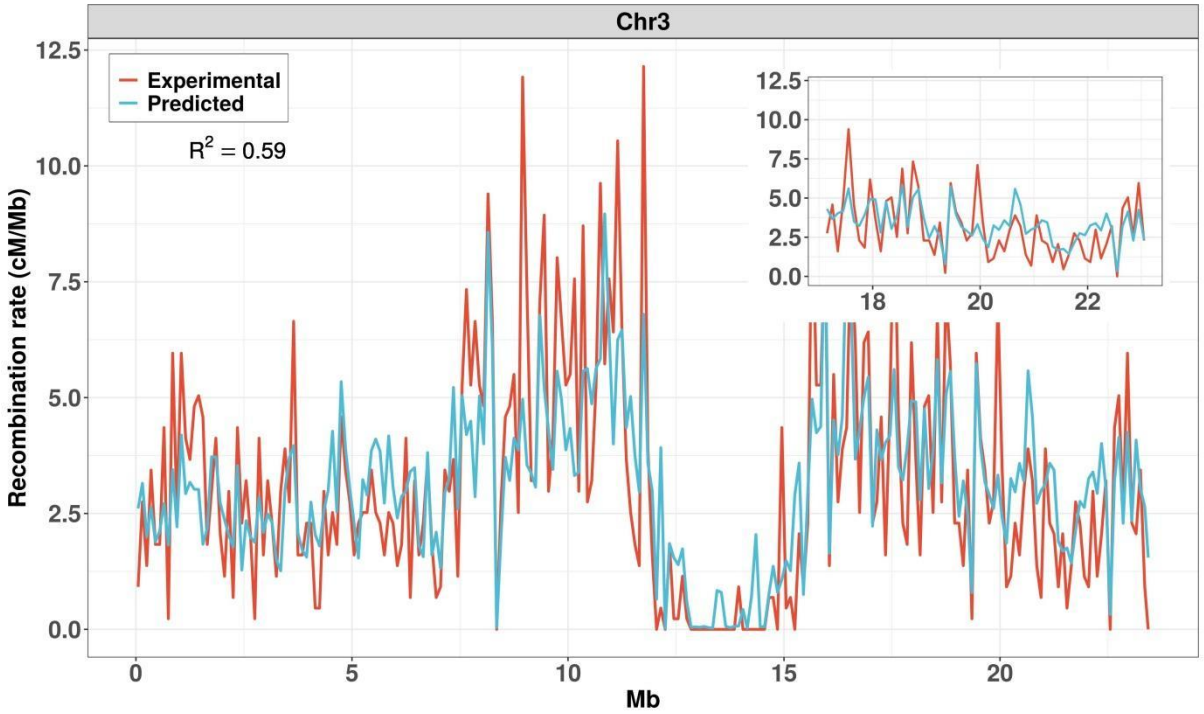


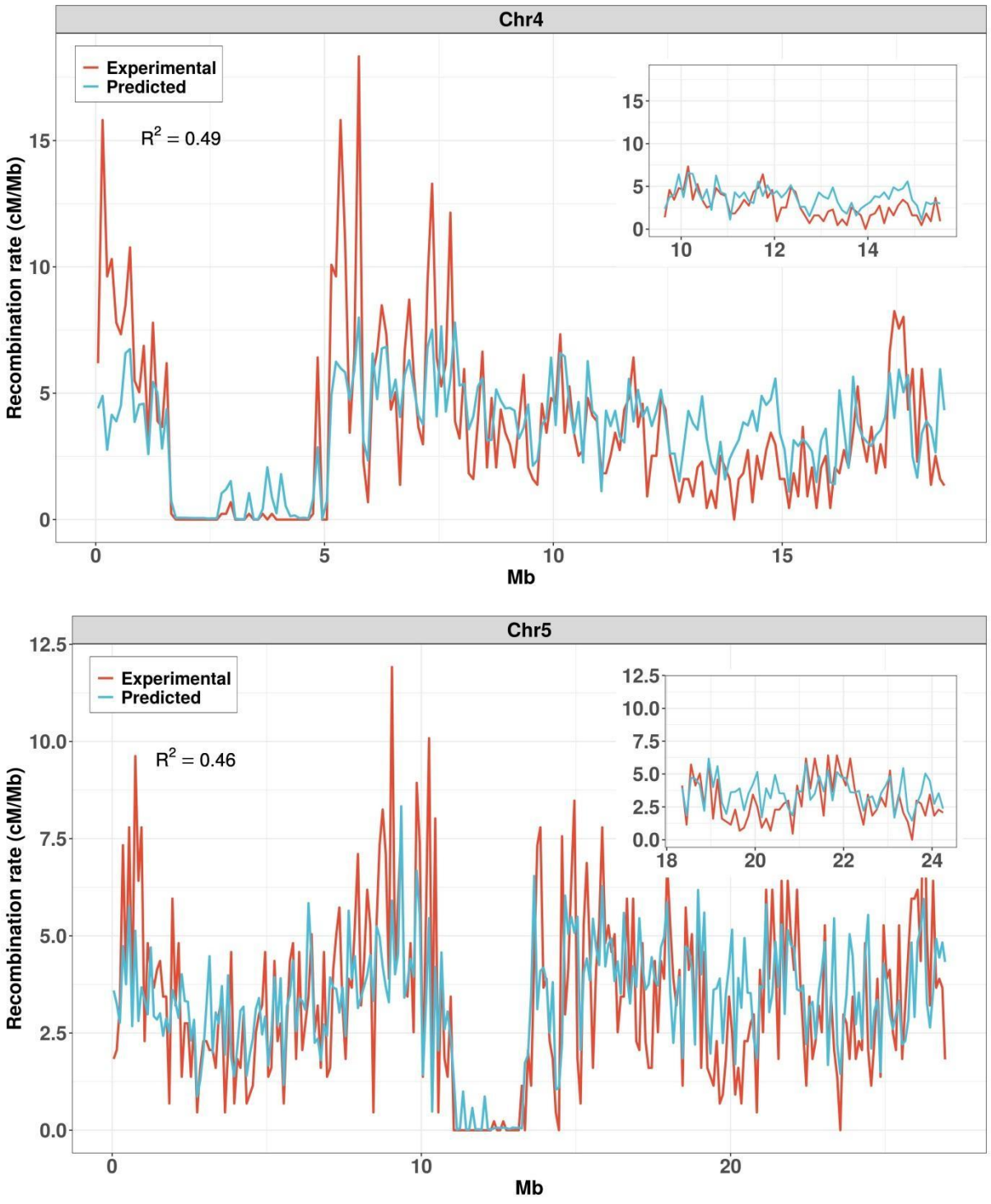
Supplementary Figure S7. Another framework to test whether recombination rate is suppressed by low SNP density. In this approach (different from the one in Main), we compare two hypotheses, H0 and H1. Under H0, we assume that there is an (unknown) “reference” recombination landscape, likely driven by genomic or epigenomic features, but common to all 5 F2 populations of Blackwell et al. (2020). (In Main, this reference landscape was implicitly assumed to be constant.) Under H1, the common landscape is further modulated by the divergence between the homologs present, thus differently in each cross and each bin. This modulation is parametrized via the function (a + b x) exp(- cx) where x is the SNP density of the bin in the considered cross. Because high SNP density is expected to lead to suppressed recombination, the test is only applied to data belonging to the first two quantiles of SNP density. We confront H0 to H1 by asking whether a good fit to the data necessitates the modulation effect. We thus compare the chi-square goodness of fit using H1 to what would be expected if there were no causal suppressive effect (the H0 hypothesis). That distribution is obtained by shuffling in each bin the values of SNP density between crosses to decorrelate recombination rate from any SNP density effect. The figure displays the histogram of the chi-square values under H0 where for each shuffling we have adjusted the parameters a, b, and c to minimize the chi-square for that shuffle. Also, the red line gives the chi-square value in the unshuffled data, corresponding to H1, showing that the recombination rate modulation, when using the SNPs between the parents of each separate cross, improves the fit far more than expected by chance (p-value $\leqslant 0.001$).



Supplementary Figure S8. Scatterplots of experimental and predicted recombination rate when the 15 parameter model calibration is done using bin sizes ranging from 50 to 500 kb. The x-axis specifies the recombination rate predicted by our quantitative model that incorporates 10 chromatin states along with contextual modulating effects, having a total of 15 adjustable parameters. The y-axis corresponds to the experimental recombination rate as produced from the Rowan *et al*. (2019) dataset. R^2 is the fraction of the variance explained by the model; it inevitably increases as bin size decreases because the CO numbers per Mb are more subject to stochastic noise.


Chr2
Experimental
Predicted
R^2 = 0.47
Recombination rate (cM/Mb)
Mb


Chr3
Experimental
Predicted
R^2 = 0.59
Recombination rate (cM/Mb)
Mb



Supplementary Figure S9. Experimental and predicted recombination landscapes of chromosomes 2 to 5. Landscapes using 100 kb bins were produced from the Rowan *et al*. dataset (red) and from our quantitative model with 15 adjustable parameters (blue). Each inset shows a corresponding zoom within the right arm. R^2 is the fraction of the recombination rate variance that is explained by the model.

3. Plant diversity applied to the improvement of plant breeding

3.1 From domestication to plant breeding – the genetic diversity of cultivated peanut

3.1.1 The origin of cultivated peanut

The cultivated peanut (*Arachis hypogaea L.*) is an allotetraploid legume crop (2n=4x=40), which is a worldwide important oil crop. The cultivated peanut belongs to the genus *Arachis* which originated from South America. The early human management of wild peanuts can be traced back to about 8500 years ago according to radiocarbon dating by accelerator mass spectrometry (AMS) using the macrofossils of peanuts. Collected in the western slopes of the Andes in northern Peru, these peanuts with morphologies corresponding to the wild species were recognized as the one managed by humans probably during the early stage of peanut domestication (Dillehay et al., 2007). To date, there are 80 peanut species including wild and cultivated ones that have been collected and described (Krapovickas & Gregory, 1994; Valls & Simpson, 2005), and all wild species in the *Arachis* genus were found in South America. The distribution of peanut species covers the eastern Andes Mountains in Bolivia, northern Argentina, central/northeastern/southeastern Brazil, eastern Paraguay and the west half of Uruguay (Figure 10) (Bertioli et al., 2011). *Arachis* species that show a large-scale distribution reflects their broad adaptability in diverse regions, including places on the Atlantic coast in Brazil and Uruguay or the Andes Mountains of Northwestern Argentina. In terms of the characteristics of morphology, cytology, mating type, and geographic location, the *Arachis* species can be classified into nine taxonomic sections, namely *Arachis*, *Triseminatae*, *Extranervosae*, *Caulorrhizae*, *Heteranthae*, *Rhizomatosae*, *Procumbentes*, *Erectoides* and *Trierectoides*. Depending on internal transcribed spacers (ITS) and coding regions of rDNA, analyses suggest that *Extranervosae*, *Heteranthae*, and *Triseminatae* are the most primitive sections, and *Arachis* is the most advanced section that is in correspondence with its broader distribution in South America than the other

eight sections (Bechara et al., 2010; Wang et al., 2010; Bertioli et a., 2011). Among nine sections, section *Arachis* has drawn a lot of attention because it consists of the cultivated peanut and its progenitors (Bertioli et al., 2011). According to the karyotypes, most species in section *Arachis* have metacentric chromosomes, and diploid species with 20 chromosomes can be into three types of genomes, which are A, B, or D, respectively. Both A and B genomes have symmetric karyotypes, but the A genome is characterized by its smaller chromosomes than B genome. Unlike A and B genomes, D genome has asymmetric karyotype, *Arachis glandulifera* for instance has a number of subtelocentric and telocentric chromosomes which are assigned to this genome group (Stalker, 1991). In addition, the cultivated and wild peanuts, *A. hypogaea* and *A. monticola*, are both tetrapolyploid species with the AABB genome (Husted, 1936; Smartt et al., 1978). Furthermore, A and B genome species within section *Arachis* can be mainly grouped into two divisions according to different molecular studies, the D genome and several diploid species with 18 chromosomes are more similar to B genome species (Bechara et al., 2010; Bravo et al., 2006; Cunha et al., 2008; Gimenes et al., 2007; Halward et al., 1992; Milla et al., 2005; Moretzsohn et al., 2004; Tallury et al., 2005; Tang et al., 2008). Using 5S and 18S-26S rDNA and heterochromatin detection by DAPI, the relationship among A or B genome wild species of section *Arachis* were separately identified by comparing with two subgenomes of the cultivated peanut. The A genome of *A. hypogaea*, the cultivated peanut, was closely related to *A. duranensis*, *A. villosa*, *A. schininii* and *A. correntina* (Robledo et al., 2009). On the other hand, the B genome of *A. hypogaea* falls in the same group, named as B *sensu stricto*, with *A. ipaensis*, *A. magna*, *A.gregoryi*, *A. valida*, and *A.williamsii*, and this group is characterized by the lack of centromeric heterochromatin (Robledo & Seijo, 2010).

As an allotetraploid, the A and B genomes *A. hypogaea* behave like a diploid organism during meiosis that has chromosomes pair as bivalents during

meiosis. The exact origin of cultivated peanuts has been of interest for researchers and breeders. Gregory and Gregory (1979) conducted 1,075 cross combinations within *Arachis* genus including the cultivated and wild peanuts belonging to section *Arachis*, and the cultivated peanut, leading to successful interspecies hybridizations only with wild species from section *Arachis*, implying that the progenitors of cultivated peanut come from wild species in section *Arachis*. Kochert et al (1996) utilized nuclear and chloroplast restriction fragment length polymorphism (RFLP) analysis to conclude that *A. duranensis* and *A. ipaensis* are the most probable progenitors of cultivated peanut and of another wild species with the AABB genome, *A . monticola*. Moreover, Moretzsohn et al (2013) carried out the sequence analysis based on intron and microsatellite markers to further strengthen this conjecture. Another study relied on the availability of hybridization specifically using *A. duranensis* and *A. ipaensis* that also provides supportive evidence (Fávero et al., 2006). In their study, they first successfully hybridized *A. duranensis* and *A. ipaensis*, and then produced the synthetic amphidiploid of this cross combination induced by colchicine treatment. Furthermore, the hybrids can be obtained by separately hybridizing the synthetic amphidiploid and six botanical varieties of *A. hypogaea*, a result that thus strongly supports the claim that *A. duranensis* and *A. ipaensis* are the two ancestors of the cultivated peanut.

For establishing fundamental knowledge concerning the evolution and domestication of cultivated peanut, Bertioli et al (2016) performed genome sequencing of *A. duranensis* V14167 and *A. ipaensis* K30076 which led successfully to the production of the synthetic amphidiploid (Fávero et al., 2006). Corresponding to the karyotype result that indicates the A genome has smaller chromosomes than the B genome, all *A.duranensis* pseudomolecules were smaller than their *A. ipaensis* counterparts partly because frequencies of local duplications and of transposable elements are lower in *A.duranensis*. The analysis of their collinear chromosomes showed that the regions of

A.duranensis are about 80 - 90% of the length in the corresponding regions of *A. ipaensis*. Sequence analysis was carried out for comparing the combined sequences of two diploid ancestors and the sequence of cultivated peanut (cv. Tifrunner); the result indicated that *A. hypogaea* is more similar to the B-genome ancestor than the A-genome one. Then, those authors utilized the number of synonymous substitutions per synonymous site (K_s) to infer the evolutionary divergence of two diploid ancestors and the corresponding cultivated peanut genome. The estimated divergence times of *A. duranensis* V14167 and *A. ipaensis* K30076 from the sub genomes in *A. hypogaea* are about 247,000 and 9,400 year ago, respectively. The result showed the high similarity between *A. ipaensis* and the B-genome of *A. hypogaea*, not only indicating the genetic bottleneck and reproductive isolation in these two species but suggest an interesting hypothesis for peanut domestication. Due to their reproduction nature, *Arachis* species develop pods under the ground and have their seed dispersal in a limited area. The population of Arachis species only moved 1 km in more than a thousand years. *A. ipaensis* is the only B-genome *Arachis* species identified in regions of *A. duranensis*, but *A. magna*, the closest relative of *A. ipaensis*, was found at a few hundred km to the north from the intersected region between *A. duranensis*, *A. ipaensis* and *A. hypogaea*, indicating that the population of *A. ipaensis* was possibly established by human transport from the north and eventually had allopolyploidization with *A. duranensis* to form the current cultivated peanut (Figure 11).

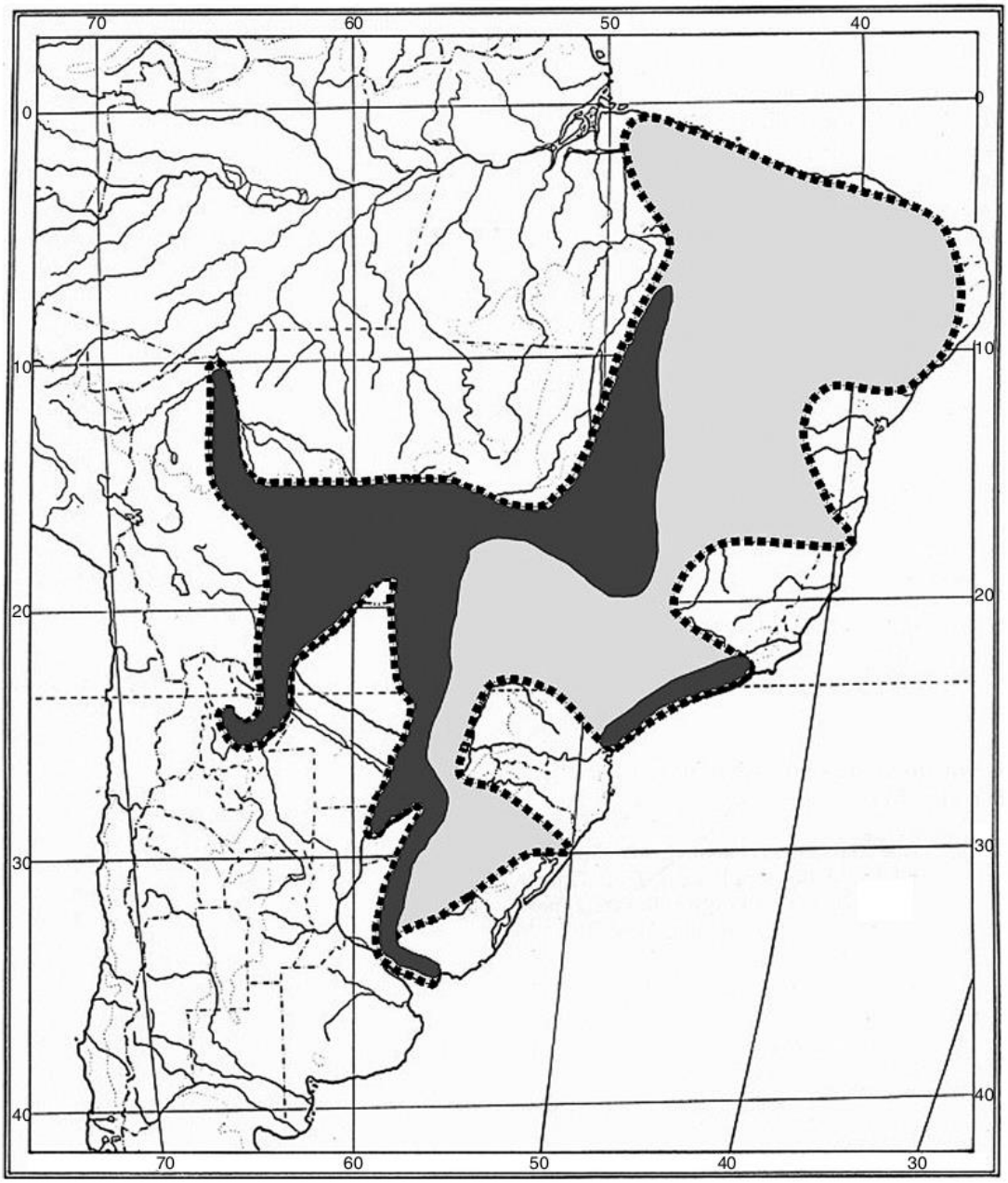


Figure 10. (adapted from Bertioli et al., 2011) The geographic distribution of 80 species in the genus *Arachis*. The dashed line covers the whole distribution of all species including the region in dark gray containing the section *Arachis* and the region in light gray containing the other 8 taxonomic sections within the genus *Arachis*. In the southeast of Brazil, along the coast, the dark gray area is the location of *Arachis stenosperma*. In general, this distribution was influenced by human transport and management due to use as food.

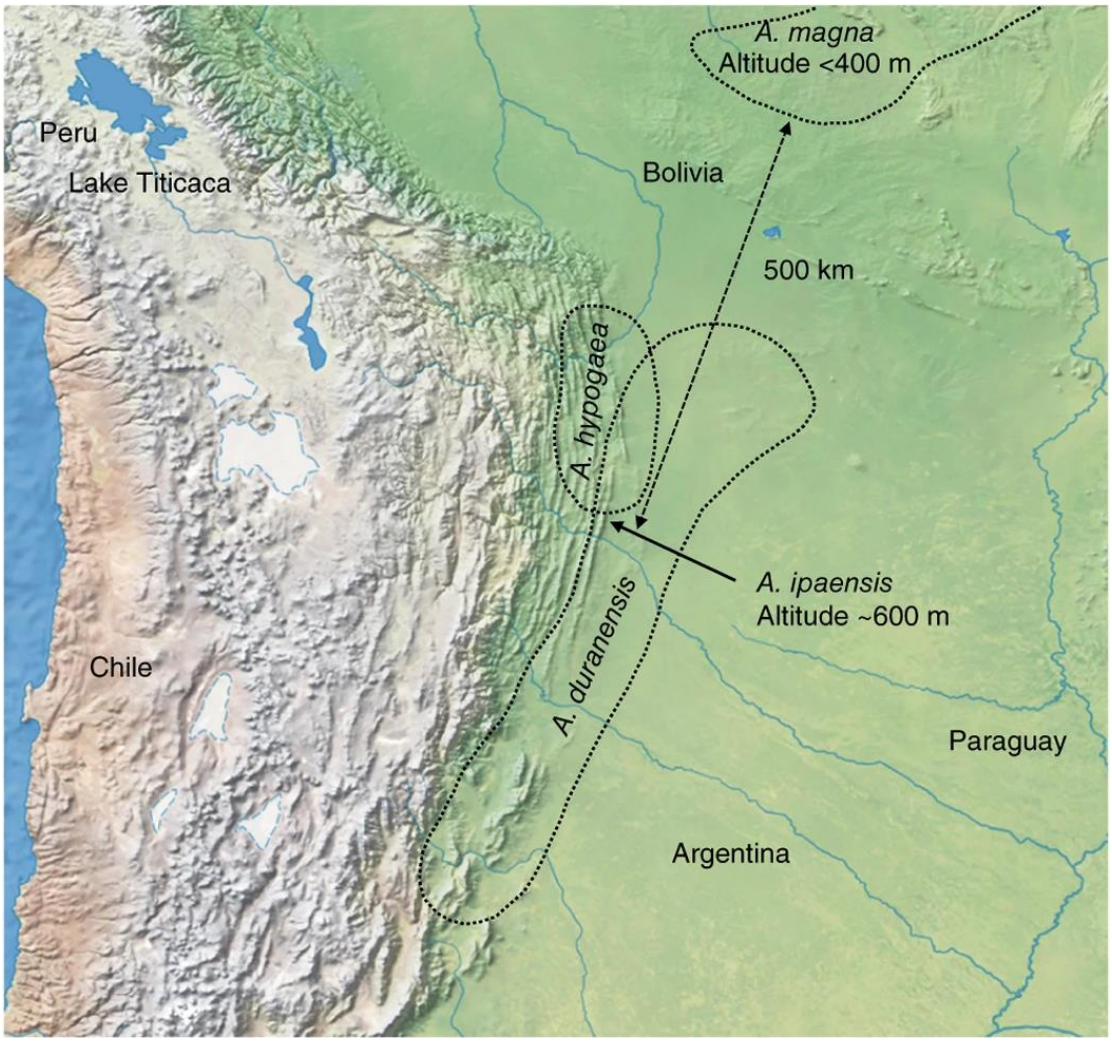


Figure 11. (adapted from Bertioli et al., 2016) The approximate known distributions of the cultivated peanut, its two ancestors, and related species. *A. ipaensis*, one of two diploid ancestors of *A. hypogaea*, is known to have the only location close to the region of *A. duranensis* (the other diploid ancestors of *A. hypogaea*) and *A. hypogaea* subsp. *hypogaea* var. *hypogaea* (recognized as the center of diversity with the earliest cultivated peanut). In addition, *A. ipaensis* is the only B-genome species close to the distribution region of *A. duranensis*, and the relative of *A. ipaensis*, A. magna, was distributed at the place about 500 km to the north from the region with *A. duranensis*. Furthermore, the divergence of *A. ipaensis* genome from the B genome of cultivated peanut is estimated to go back to about 9,400 years ago, suggesting that the current location of *A. ipaensis* populations were

probably established by people who transported seeds from the north. Then, *A. ipaensis* and *A. duranensis* formed the allotetraploidy *A. hypogaea*.

3.1.2 Genetic variation and germplasm conservation of cultivated peanut

Even though it was known that the cultivated peanut has limited genetic diversity resulting from serious genetic bottlenecks (Burow et al., 2001; Foncéka et al., 2009), the morphological variation within this species can be divided into two subspecies, *hypogaea* and *fastigiata*, which are further classified into several botanical varieties. *A. hypogaea* subsp. *hypogaea* has two botanical varieties, *hypogaea* and *hirsuta*. Compared with the subspecies *fastigiata*, *hypogaea* and *hirsuta* that belong to subsp. *hypogaea* have a longer live cycle and no flowers on the central stem. With vegetative and reproductive sides stems regularly alternated, var. *hypogaea* not only has landraces which are found along the Amazon River in Brazil and Bolivia, but also acquired the modern market types including “Virginia” and “Runner”. The other botanical variety, var. *hirsuta*, exhibiting more hirsute leaflets and even an extended life cycle, is localized on the coast of Peru. On the contrary, subsp. *fastigiata*, with four botanical varieties, has a shorter cycle, flowers on the central stem, and a disorganized distribution of reproductive and vegetative stems. Usually with two seeds inside fruits, var. *vulgaris*, also referring to the “Spanish” type, is distributed in the Uruguay river basin. Different from var. *vulgaris*, var. *fastigiata*, corresponding to the “Valencia” type, has more than two seeds inside its fruits and smooth pericarps. Its distribution includes Paraguay and the central and northeastern part of Brazil, extending to Peru. The other two botanical varieties of subsp. *fastigiata*, *aequatoriana* and *peruviana*, have reticulated pericarps, more than two seeds inside fruits and a limited distribution around the world.

The conservation of genetic variation of cultivated and wild peanut species has been maintained by a number of *ex situ* worldwide collections. There are six important *Arachis* germplasm collections around the world. Currently, two out of six collections are based in India. ICRISAT (International Centre for Research in the Semi-Arid Tropics) has 15,622 accessions of *Arachis* genus, including more than 15,000 *A. hypogaea* accessions from 92 countries and 480 wild accessions collected from six countries. Another collection in India is ICAR-NBPGR (National Bureau of Plant Genetic Resources) that conserves 13,755 accessions in total, of which 81 are wild accessions belonging to 16 wild *Arachis* species. The United States has two collections, USDA-ARS (United States Department of Agriculture) and TAMU (Texas AgriLife Research Center, Texas A&M University). At the present, the USDA-ARS collection consists of 9,753 accessions, including 9,194 cultivated and 559 wild accessions. On the other hand, TAMU contains a total of about 6,500 accessions with more landraces collected from South America and wild *Arachis* species (954 accessions from at least 76 species) than USDA-ARS. In China, OCRI-CAAS (Oilseed Crops Research Institute, CAAS) maintains more than 8,600 accessions with 8,307 *A. hypogaea* ones, 234 ones from 37 wild Arachis species and 123 wild species hybrids. The last one, EMBRAPA (Empresa Brasileira de Pesquisa Agropecuária), is set up in Brazil, famous for being the largest and broadest conservation in wild *Arachis* species. The EMBRAPA collection is composed of 1,559 accessions from 84 wild species (79 known wild species and 5 recently discovered new species) and 2,508 accessions of the cultivated peanut. Taken together, even though it should be noted that a certain amount of accessions conserved by these collections are duplicated, these six collections account for two thirds of the worldwide conserved collections and thus represent well the genetic diversity of cultivated peanut.

Since the growing accessions in germplasm collections can become too massive to be deal with, Frankel (1984) developed the concept of a “core

collection” which can represent the overall genetic diversity of the total germplasm collection. Three of six above-mentioned collections, ICRISAT, USDA-ARS and OCRI-CAAS, have further founded core collections based on their original collections. Holbrook et al (1993) first stratified the entire USDA-ARS collection into 9 sets, and performed multivariate analysis to classify each set into groups when the morphological data was available. Finally, 10% of samples were randomly selected from each of these groups to establish the USDA core collection. The ICRISAT core collection was set up using a similar strategy as the USDA one (Upadhyaya et al., 2003). The whole ICRISAT core collection (14,310 accessions) was stratified first by six botanical varieties and then by country of origin. The accessions from the same botanical variety from small and nearby countries were merged into the same group, leading to 75 groups in total. Based on the multivariate analysis using 14 morphological data, 10% of individuals from each cluster in each group were chosen to establish the core collection with 1,704 accessions. The final core collection was based on the entire OCRI-CAAS collection (Jiang et al., 2008). After the progressive stratification using botanical varieties and the origin of countries, 6,390 accessions were further clustered into 258 groups using the multivariate analysis of morphological and biochemical data. Eventually, 5-10% of accessions were selected from each cluster to form the core collection with 576 samples. Even though a core collection that contains 10% of the entire collection reduces substantially the management work, it is still hard to directly use a core collection to screen phenotypes because of cost and the time required. To make the collection still more manageable, Upadhyaya and Ortiz (2001) suggested the development of a “mini core collection” with 1% of the entire collection using the similar sampling strategy to establish a mini core collection, which can still represent a good part of the genetic diversity of the whole collection. Thus, this concept led to the production of three mini cores derived from the USDA, ICRISAT and OCRI-CAAS collection (Holbrook & Dong, 2005; Jiang et al., 2010; Upadhyaya et al., 2002).

With the advance of molecular tools, different genotyping systems have been utilized to investigate the genetic diversity of these germplasm collections. Kottapalli *et al.* (2007) performed their analysis based on 72 accessions from the USDA mini core collection using 67 SSR markers that provide reliable polymorphisms. The result gave an average gene diversity, average number of alleles per marker and polymorphism information content (PIC) values of 0.18, 7.9 and 0.15, respectively. Furthermore, cluster analysis based on genetic distance indicated that two subspecies, *fastigiata* and *hypogaea*, can be grouped into two major clusters, corresponding to their morphological classification. Another study incorporating more accessions from the USDA mini core and more SSR markers led to similar results for the average number of alleles per marker and for the population structure but resulted in greater differences for the average genetic diversity and PIC (0.59 and 0.53 respectively), suggesting that the analysis of genetic diversity can be influenced by different choices of markers and accessions (Wang et al., 2011). Jiang et al (2010) also compared the genetic diversity based on SSR markers to compare the OCRI-CAAS and ICRISAT, and concluded that the genetic distance between two mini cores is larger than the genetic distance within a core collection. In addition, Jiang et al (2013) utilized 103 SSR markers to genotype the OCRI-CAAS mini core collection and concluded that this Chinese mini core collection, with an average number of alleles per marker of 5.1, mean PIC of 0.213 and mean genetic diversity of 0.265, is less diverse than the USDA mini core collection. However, in this study Jiang *et al.* found line-specific alleles not identified in the USDA core collection. Considering these two studies (Jiang et al., 2010, 2013) comparing different mini core collections, one can conclude that these collections act on the complementary role to each other for representing the current peanut germplasm diversity.

More recently, the progress in sequencing technology has facilitated the usage of SNP markers to genotype not only mini core collections but also the larger core collections to identify finer details. Pandey et al. (2017) developed a 58 K SNP array from DNA resequencing and RNA sequencing data of 41 peanut accessions. Using this 58 K SNP array, Otyama et al (2019) genotyped the USDA mini core, and showed that these accessions can be separated into four or five groups. Among subgroups, only 43 accessions were classified into groups in agreement with the main market type groups, and the other accessions were either clustered in groups not corresponding to their market types or were classified as mixed groups. Some accessions even lacked the taxonomic classification. Since the classification for the subspecies still stayed the same as previous studies using SSR markers, this result indicated that SNP markers can provide information of finer population structure, improving methods based on morphology to define different botanical varieties, a complex and subjective task based on measurements and phenotyping in the field. Furthermore, the USDA core collection was genotyped by the Arachis_Axiom2 SNP array (Otyama et al., 2020). Several perspectives were pointed out in that study. First, the initial 791 accessions from that core collection, with substantial phenotypic difference between samples, can be replaced by one with 671 accessions by merging clusters having 99% identity, indicating that the phenotypic approach could be misleading. Second, these accessions were grouped into five clusters depending on their genetic distance, and the cluster containing accessions mainly from west-central South America (Bolivia, Peru, and Ecuador) also have the “synthetic-tetraploid” accession of *A. duranensis* and *A. ipaënsis*, in agreement with the fact that the earliest landraces were found in this area and that the tetraploid peanut originated in Southeast Bolivia. Third, genetic clusters have little correspondence with country of origin, suggesting that the seeds were widely distributed in the 18th and 19th centuries.

scientific reports

Check for updates

OPEN

Assessment of genetic diversity and SNP marker development within peanut germplasm in Taiwan by RAD-seq

Yu-Ming Hsu[1,2,3], Sheng-Shan Wang[4], Yu-Chien Tseng[5], Shin-Ruei Lee[3], Hsiang Fang[3], Wei-Chia Hung[3], Hsin-I. Kuo[5] & Hung-Yu Dai[3]✉

The cultivated peanut (*Arachis hypogaea* L.) is an important oil crop but has a narrow genetic diversity. Molecular markers can be used to probe the genetic diversity of various germplasm. In this study, the restriction site associated DNA (RAD) approach was utilized to sequence 31 accessions of Taiwanese peanut germplasm, leading to the identification of a total of 17,610 single nucleotide polymorphisms (SNPs). When we grouped these 31 accessions into two subsets according to origin, we found that the "global" subset (n = 17) was more genetically diverse than the "local" subset (n = 14). Concerning botanical varieties, the var. *fastigiata* subset had greater genetic diversity than the other two subsets of var. *vulgaris* and var. *hypogaea*, suggesting that novel genetic resources should be introduced into breeding programs to enhance genetic diversity. Principal component analysis (PCA) of genotyping data separated the 31 accessions into three clusters largely according to the botanical varieties, consistent with the PCA result for 282 accessions genotyped by 14 kompetitive allele-specific PCR (KASP) markers developed in this study. The SNP markers identified in this work not only revealed the genetic relationship and population structure of current germplasm in Taiwan, but also offer an efficient tool for breeding and further genetic applications.

Originated from South America, the cultivated peanut (*Arachis hypogaea* L.) is an allotetraploid (AABB, 2n = 4x = 40) and an important legume crop worldwide. Humans benefit from peanut seeds as food and source of oil due to their high percentage of proteins and fatty acids[1]. The annual production of peanuts has increased in the past 20 years to reach 53 million tons in 2020 according to FAOSTAT (http://www.fao.org/faostat). To fulfill the increasing peanut demand under the threat of climate change, breeding new varieties is an effective strategy to improve peanut qualitative and quantitative traits.

The conservation of *Arachis* germplasm and exploitation of their genetic diversity are crucial for the breeding of the cultivated peanut. Presently, several gene banks are renowned for their *Arachis* germplasm including the International Crops Research Institute for the Semi-Arid Tropics (ICRISAT), United States Department of Agriculture (USDA), and the Oil Crops Research Institute of the Chinese Academy of Agricultural Sciences (OCRI-CAAS). More than 15,000, 9,000 and 8,000 accessions were collected in ICRISAT, USDA, and OCRI-CAAS[2], respectively. On the other hand, understanding the genetic diversity of in-hand germplasm is the prerequisite before launching breeding programs, and the utilization of molecular markers is the predominant strategy to evaluate the genetic diversity of germplasm at present[3]. Cultivated peanut has its low genetic diversity due to the recent hybridization of its two ancestors and selection in breeding programs[4–7]. Even though the narrow genetic diversity of cultivated peanuts has hindered the development of molecular markers, it has been possible to develop and utilize simple sequence repeat (SSR) markers to assess the genetic diversity in cultivated peanut[8–11]. In particular, the population structures of 92 accessions in the US Peanut Mini Core Collection and 196 major peanut cultivars in China were revealed by SSR markers[12,13]. Although SSR markers were widely used

[1]Université Paris-Saclay, CNRS, INRAE, Univ Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91405 Orsay, France. [2]Université Paris Cité, CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), 91405 Orsay, France. [3]Crop Science Division, Taiwan Agricultural Research Institute, Taichung 413008, Taiwan, ROC. [4]Crop Improvement Division, Tainan District Agricultural Research and Extension Station, Tainan 71246, Taiwan, ROC. [5]Agronomy Department, National Chiayi University, Chiayi 60004, Taiwan, ROC. ✉email: hydai@tari.gov.tw

for identifying genetic diversity of peanut populations, these studies had limited population size due to the challenging genotyping process.

Recently, the peanut genome projects made possible by next generation sequencing (NGS) have revolutionized genetic research in cultivated peanuts. So far, the genomes of *Arachis hypogaea* L. and its two diploid ancestors, *A. duranensis* (AA) and *A. ipaensis* (BB), have been sequenced[6,7,14]. These high quality genome sequences have paved the way for developing high-throughput single nucleotide polymorphism (SNP) markers e.g. via genotyping-by-sequencing (GBS) that can then facilitate peanut molecular breeding. The 58 K SNP array '*Axiom_Arachis*', developed by resequencing 41 peanut accessions, was used to identify genetic diversity across 384 *Arachis* genotypes including USDA Mini Core Collection and wild species[15,16], while 787 accessions from the U.S. Peanut core collection were genotyped by the 14 K '*Arachis_Axiom2*' SNP array to reveal their genetic diversity[17]. Compared to SNP arrays, GBS is a more cost-effective technique based on sequencing of the reduced genome associated with restriction sites using NGS[18,19]. In peanut research, this technique was applied in SNP development, enabling the construction of genetic maps for quantitative trait locus (QTL) mapping and the analysis of population structure[20–22].

In Taiwan, peanut breeding programs can be traced back to the late 1950s. To date, most varieties developed locally have been obtained by conventional breeding based on evaluating morphological traits. In such breeding programs, the parental selection mainly relied on the pedigree information or the collection source to infer genetic relationships. Thus, exploiting available molecular tools to characterize the present peanut varieties in Taiwan should allow improved breeding programs in the future. In this study, we performed the restriction site-associated DNA (RAD) approach to sequence 31 genotypes—including current elite varieties developed in Taiwan and important accessions introduced from abroad—to reveal the underlying genetic diversity. Furthermore, 14 kompetitive allele-specific PCR (KASP) markers were designed and then used to genotype 282 other accessions. Overall, this work reveals the genetic structure of peanut germplasm in Taiwan through SNP markers identified by RAD-seq and these markers can be used for a number of applications such as variety identification and breeding programs.

Materials and methods

Plant materials and DNA extraction. 31 peanut accessions, maintained by Taiwan Agricultural Research Institute (TARI) and Tainan District Agricultural Research and Extension Station (Tainan DARES), were chosen for RAD-seq construction. These accessions consist of elite cultivars, advanced breeding lines, and "introduced" old accessions acquired in South American countries close to the geographic origin of peanut (Supplementary Table S1). Among 31 peanut accessions, there are 13 Spanish, 11 Valencia, 3 Virginia, and 4 Runner type accessions. For the genotyping via KASP markers, 282 peanut accessions were obtained from the National Plant Genetic Resources Center in TARI, including 66 Spanish, 27 Valencia, 49 Virginia and 88 Runner type accessions. The plant materials utilized in this study conform to relevant international, national and institutional guidelines.

The DNA extraction of all accessions was based on young leaves collected from seedlings within two weeks using the modified CTAB method which replaces phenol and chloroform with potassium acetate to remove protein and polysaccharides[23]. The DNA samples extracted from the modified CTAB method can directly be used for KASP genotyping, but need further purification to ensure their quality for RAD-seq library construction. Thus, after extracting DNA of 31 accessions used for RAD-seq, we utilized the QIAGEN kit (DNeasy Blood & Tissue Kit; Qiagen, https://www.qiagen.com/, Hilden, Germany) to purify these DNA samples which were then quantified and qualified by NanoDrop 2000 spectrophotometer (Thermo Fisher Scientific Inc., https://www.thermofisher.com, DE, USA). The purified DNA samples with (1) the 260/280 ratio from 1.8 to 2.0, (2) the 260/230 ratio from 2.0 to 2.4, and (3) the concentration ≥ 25 ng/μl were further checked for the DNA integrity by agarose gel (1.0%) electrophoresis.

Phenotypic evaluation. 24 out of the 31 peanut accessions used in RAD-seq and 282 additional peanut accessions of the germplasm were phenotyped in the fall of 2016 in TARI (coordinates 24° 01′ 47.5″ N 120° 41′ 47.4″ E), and 20 plants of each accession were evaluated for 8 quantitative traits, including days to flowering (between the sowing and flowering date), plant architecture, number of pods, yield (g/m^2), 100-pod weight, 100-seed weight, rust resistance and leaf spot resistance. The susceptibility to these two peanut diseases was quantified under natural conditions in the field since these diseases develop spontaneously during the fall, and the disease symptoms were scored from 1 (having no symptoms) to 9 (highly susceptible)[24]. Depending on the degree of inclination from verticality, plant architecture was scored from 0 (the most upright) to 9 (the most prostrate).

RAD-seq library construction and SNP calling. After finalizing DNA extraction, purification and quality control of 31 peanut accessions, we used 1.1 μg of each high-quality DNA sample to make two RAD-seq libraries from 16 and 15 accessions, respectively, following the published protocol, and *PstI* was chosen as the digestion enzyme[25]. Next-generation sequencing of each library was carried out in the Genome Research Center of Yang-Ming University using the Illumina Hiseq 2500 platform (Illumina Inc., https://www.illumina.com, CA, USA) with 100 bp single-end reads in two lanes. The sequencing data have been deposited at National Center for Biotechnology Information (NCBI) under BioProject PRJNA811600. For SNP calling, single-end reads were first debarcoded by Stacks using the program "process_radtags"[26]. Then, we used Burrows-Wheeler Alignment (BWA) v0.7.17-r1188 "aln" to align the reads of each accession onto the reference genome of cultivated peanut and its diploid ancestors for identifying SNPs used in the genetic diversity analysis and the development of KASP markers, respectively[6,14,27]. When the genome of cultivated peanut was published[14], all of the KASP markers used in this study had already been designed using the merged genomes of two diploid ancestors of cultivated peanut[6].

Thus, identified SNPs based on the genome of cultivated peanut were only used for the *in-silico* analyses that investigated the genetic diversity of the 31 accessions, but in all cases the SNP calling pipeline was the same. After the alignment was finished, Samtools and BCFtools were utilized for SNP calling and filtering, SNPs were kept with (1) base quality ≥ 20, (2) mapping quality score ≥ 20, and (3) depth ≥ 3. Then, a customized R script was used to create Variant Call Format (VCF) files encompassing qualified SNPs that discriminated the 31 accessions[28].

The development and validation of KASP markers. Among SNPs available for distinguishing 31 peanut accessions based on the genome of the two diploid ancestors[6], we extracted 1,230 homozygous SNPs with informative alleles in all 31 accessions, and then discarded 783 SNPs having Polymorphic Information Content (PIC) values lower than the average over all SNPs. Finally, 29 out of 477 SNPs with an average PIC value of 0.28 were selected for developing KASP markers. These 29 putative SNPs with 100 bp flanking sequences on both sides were used for designing KASP primers that were then synthesized by LGC genomics (http://www.lgcgroup.com, Teddington, England). The validation of KASP markers was performed on the 96-well StepOnePlus™ Real-Time PCR System (Thermo Fisher Scientific Inc., https://www.thermofisher.com, DE, USA), and each 10-µL reaction consisted of 12.5 ng of DNA, 0.14 µL of KASP assay mix and 5 µL of KASP Master Mix (2X). The PCR protocol was carried out as follows: (1) pre-read stage at 30 °C for 1 min, (2) hold stage at 94 °C for 15 min, (3) PCR stage 1 of 10 touchdown cycles using 94 °C for 20 s and 61 °C (decreasing 0.6 °C per cycle) for 1 min, (4) PCR stage 2 with 26 amplification cycles at 94 °C for 20 s and 55 °C for 1 min, and (5) post-read stage at 30 °C for 1 min. When the PCRs were completed, the fluorescent signals of samples were analyzed by the StepOne™ software for determining genotypes.

Statistical analysis. All statistical analyses were carried out in R 3.63. The PIC value and the expected heterozygosity (He) were determined for each SNP marker[29,30]. Principal component analysis (PCA) using phenotypic data was performed by the "PCA" function in the "FactoMineR" package[31]. In the "poppr" package, the "bitwise.dist" function was used to calculate the genetic distances between the 31 accessions, and these distances were calculated depending on the fraction of loci which differ between germplasm[32,33]. The "aboot" function was utilized to construct the dendrograms based on the unweighted pair group method with arithmetic mean (UPGMA) with 1000 bootstraps. In the "adegenet" package, PCA for SNP data from 31 accessions was carried out by the "glPCA" function, and population structure of 282 accessions was addressed by successive K-mean clustering and discriminant analysis of principal components (DAPC) using the "find.clusters" and "dapc" function, respectively[34,35]. In addition to the dendrogram plot, created by the "plot.phylo" function in the "ape" package[36], all visualization was performed using the "ggplot" function in the "tidyverse" package[37].

Results

SNP marker development from 31 peanut accessions using RAD-seq. In this study, 31 peanut accessions were chosen to conduct RAD-seq, of which 17 accessions were introduced from abroad and 14 accessions developed or collected in Taiwan. This collection has important agronomic traits including yield-related traits, resistances to biotic and abiotic stresses, and valuable characteristics at the genetic diversity level (Supplementary Table S1).

In the RAD-seq approach, the six-cutter enzyme, *PstI*, was utilized for the DNA digestion, and so sequencing of each accession focused on approximately 5% (100-bp extensions on both side of a *PstI* cutting site that occurs every 4,096 bp on average) of the total cultivated peanut genome (2.7 Gb). The estimated sequencing depth in the 31 accessions ranged from 4.26 (HL2) to 15.01 (Red), and the average depth was 9.47. In addition, more than 99.0% of sequenced reads from all samples were properly aligned to the reference genome. Compared to the reference genome, *A. hypogaea* cv. Tifrunner, there were 1475 to 14,471 SNPs identified from these 31 accessions with an average of 5249 SNPs, and more than 3 quarters of these polymorphisms were homozygous. In addition, the transition/transversion (Ts/Tv) ratio ranged from 0.48 to 1.19 (Table 1). In terms of the three botanical varieties of the cultivated peanut, accessions from subsp. *fastigiata* var. *vulgaris* (Spanish type), subsp. *fastigiata* var. *fastigiata* (Valencia type) and subsp. *hypogaea* var. *hypogaea* (Virginia/Runner types) had a total of 5006, 5119 and 5905 SNPs, i.e., the differences across botanical varieties were very small. Interestingly, the 31 accessions separated well according to the global and local collection, corresponding to 17 genotypes introduced from other countries and 14 genotypes from Taiwan, respectively. The global collection had an average of 6071 SNPs which was higher than the average of 4526 SNPs for the local collection. Moreover, 8 introduced accessions, collected in South America close to the center of origin of cultivated peanut, led to an average of 7139 SNPs, even higher than that of the global collection. This result suggested that the global collection germplasm from various countries had more polymorphisms than the local one containing mainly Taiwanese cultivars.

Then, the next stage of filtration was performed to keep only SNPs differentiating these 31 accessions. As a result, 3474 out of 17,610 SNPs were finally kept for the genetic diversity analysis using a tolerance of 6 missing values (20%) at most for each polymorphism.

Evaluation of genetic diversity and cluster analysis based on 31 peanut accessions. The genetic diversity of the 31 peanut accessions was quantified by a number of measures, including the expected heterozygosity (He), the major allele frequency (MAF), polymorphic information content (PIC), and genetic distance. The genetic distance was based on the bitwise distance, identical to Provesti's distance, growing with the fraction of genetically different loci between 31 accessions[32]. The pairwise comparison of the genetic distance between accessions is listed in Supplementary Table S2. On average, these 31 peanut accessions had a He of 0.19, PIC of 0.16, MAF of 0.87, and distance of 0.17. While considering botanical varieties, germplasm from subsp. *fastigiata* var. *fastigiata* had the largest average He, PIC and genetic distance (He = 0.18, PIC = 0.15, distance = 0.15) and

Germplasm	Properly mapped reads (%)	Estimated depth[a]	Filtered SNPs[b]	Homozygous SNPs[b]	Ts/Tv ratio[c]
PI153169	99.44	8.4	5666	5342	0.66
PI259717	99.26	8.98	5502	5118	0.67
PI565455	98.99	10.65	14,471	14,051	0.63
Tainung 7 (TNG7)	99.52	12.04	6166	5848	0.60
Tainung 10 (TNG10)	99.56	12.14	5884	5630	0.64
Tainan 14 (TN14)	99.18	9.7	1682	1348	1.12
Tainan 15 (TN15)	99.11	6.38	1475	1217	1.19
Tainan 18 (TN18)	99.15	5.02	1794	1461	0.68
Tainan Selection 9 (TNS 9)	99.04	13.13	6884	6387	0.59
Hualieng 1 (HL1)	99.14	9.77	3600	3230	0.67
India	99.09	10.67	3495	3174	0.72
Xiamen	99.02	11.35	5858	5483	0.70
Vietnam	99.10	7.97	2605	2301	0.75
PI118480	99.48	9.25	5564	5277	0.83
PI118989	99.28	11.02	11,823	11,367	0.53
PI155112	99.44	7.87	5163	4829	0.74
PI314817	99.38	9.16	11,051	10,653	0.73
PI338337	99.62	10.72	6187	5830	0.69
Tainan 16 (TN16)	99.15	7.63	1683	1326	1.19
Tainan 17 (TN17)	99.19	4.77	1606	1324	1.14
Hualieng 2 (HL2)	99.17	4.26	1896	1642	0.62
E01001	99.22	6.32	2376	1923	0.98
E01004	99.21	6.05	2759	2393	0.85
Red	99.17	15.01	6204	5797	0.51
NS011001	99.15	9.97	4147	3735	0.87
PI109839	99.57	10.19	6805	6499	0.48
Taichung 1 (TC1)	99.63	12.65	7390	7097	0.65
PI145681	99.47	9.51	2618	2342	0.56
PI599592	99.55	12.68	6408	6150	0.69
PI203396	99.59	8.92	4855	4593	0.59
Penghu 1 (PH1)	99.34	11.37	9115	8713	0.58

Table 1. Sequence and SNP information of our 31 accessions in the Taiwanese peanut germplasm. [a]The estimated depth was calculated by the total number of bases divided by 4.8% of 2.7 Gb, the size of reduced reference genome. [b]The SNP identification was based on the reference genome of *A. hypogaea* cv. Tifrunner. [c]Ts/Tv is the abbreviation of transition/transversion.

	Number	Mean He	Mean MAF[a]	Mean PIC[b]	Mean genetic distance
The whole collection	31	0.19	0.87	0.16	0.17
The origin of germplasm					
The global subset	17	0.19	0.86	0.15	0.17
The local subset	14	0.16	0.88	0.14	0.14
The botanical variety					
var. *vulgaris*	13	0.13	0.90	0.11	0.11
var. *fastigiata*	11	0.18	0.87	0.15	0.15
var. *hypogaea*	7	0.12	0.92	0.10	0.11

Table 2. Genetic diversity in the 31 accessions of Taiwanese peanut germplasm. [a]MAF, major allele frequency. [b]PIC, polymorphic information content.

smallest MAF (0.87), to be compared to that of the germplasm from subsp. *fastigiata* var. *vulgaris* (He = 0.13, PIC = 0.11, MAF = 0.90, distance = 0.11) or subsp. *hypogaea* var. *hypogaea* (He = 0.12, PIC = 0.10, MAF = 0.92, distance = 0.11) (Table 2), showing that Valencia type germplasm acquired higher genetic diversity than both Spanish type and Virginia/Runner type germplasm. In terms of the collection source, the global collection had larger average He, PIC and genetic distance (He = 0.19, PIC = 0.15, distance = 0.17) and smaller MAF (0.86) than

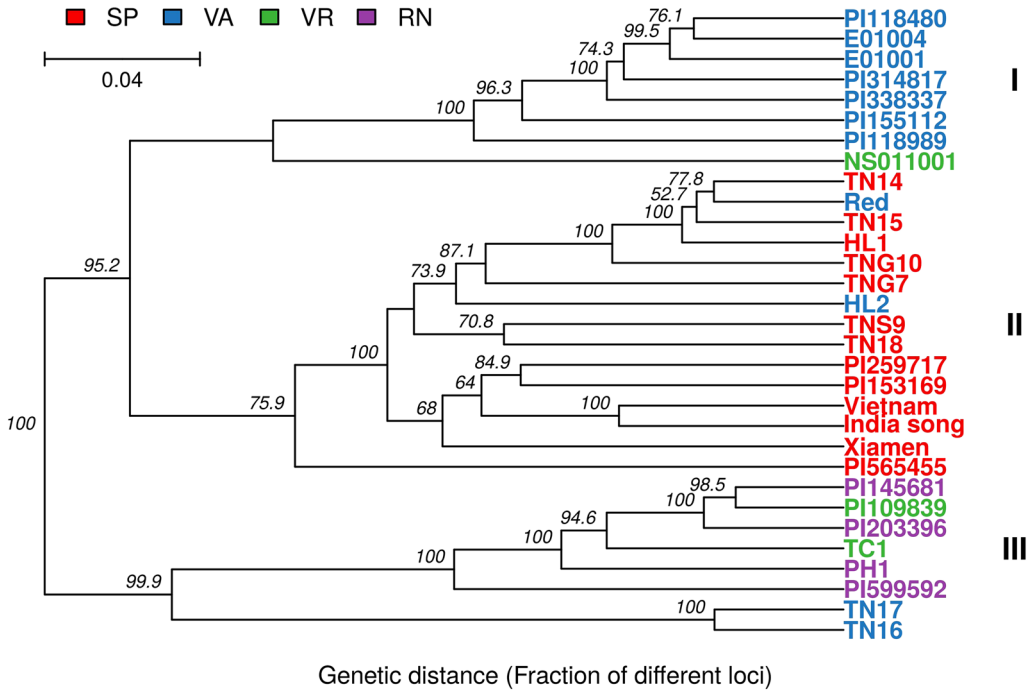


Figure 1. Dendrogram of the 31 accessions created from the unweighted pair group method with arithmetic mean (UPGMA). This dendrogram was based on the pairwise genetic distance with 1000 replicates using 3474 single nucleotide polymorphisms (SNPs). The branch length represents genetic distance, and the scale is on the top left of this figure. The numbers on the branches are bootstrap percentages. The legend shows the color of four market types, Spanish (SP), Valencia (VA), Virginia (VR), Runner (RN), and three clades clustered in this plot were named as I, II and III.

the local collection (He = 0.16, PIC = 0.14, MAF = 0.88, distance = 0.14), indicating that the global collection had greater genetic diversity than the local collection. In addition, the distance tree for cluster analysis was reconstructed based on the UPGMA method with 1,000 bootstraps (Fig. 1). The results showed that 26 out of the 31 accessions were clustered into 3 groups mainly according to three botanical varieties. Germplasm of subsp. *fastigiata* var. *fastigiata* and subsp. *fastigiata* var. *vulgaris* were clustered into Group I and II with a distance of 0.19, and germplasm of subsp. *hypogaea* var. *hypogaea* was clustered into Group III separated from Group I and II with a distance of 0.21. With the exception of 5 accessions, NS011001 (Virginia type) was clustered into group I with mainly Valencia type germplasm, while two Valencia type accessions, Red and HL2, were clustered into group II with mostly Spanish type germplasm. Interestingly, TN16 and TN17, two Valencia type cultivars, were clustered into group III, but they were separated from accessions of subsp. *hypogaea* var. *hypogaea* with a distance of 0.18.

To further investigate and compare the genetic relationship among these germplasm, PCA were performed using genetic distances between the 31 accessions calculated via 3474 SNPs. The PCA result showed that the first three principal components (PC1, PC2 and PC3) explained 24.2%, 20.8% and 8.2% of the variance, respectively, totaling 53.2% of the overall genetic distance variance (Fig. 2). However, the scatter plots of PCs suggested that PCA based on genomic data distinguished well the 31 accessions. In the three biplots of PC1/PC2, PC2/PC3 and PC1/PC3 based on PCA using 3474 SNPs, the first pair succeeded in distinguishing 31 accessions into three clear groups, mainly according to three botanical varieties, subsp. *fastigiata* var. *vulgaris* (Spanish type), subsp. *fastigiata* var. *fastigiata* (Valencia type) and subsp. *hypogaea* var. *hypogaea* (Virginia/Runner types), while the second and third pair were capable of separating TN16 and TN17 from three clusters assigned by the first pair (Fig. 2). Furthermore, the 3D scatter plot created using the three PCs displayed a relationship of 31 accessions compatible with the three PC biplots (Fig. 2).

The development and validation of KASP markers. In this study, one of our goals was to design a set of non-gel based SNP markers which could be exploited to investigate the genetic structure within the germplasm collection conserved in the National Plant Genetic Resources Center of TARI. When this project was launched, the genome of cultivated peanut was not published yet. Thus, the development of SNP markers for the KASP genotyping relied on the two diploid ancestors of cultivated peanut[6]. Note that the SNP calling pipeline used here was the same as the one that identified SNPs from the cultivated peanut genome for assessing the genetic diversity of the 31 accessions. Of the SNPs identified by the mapping to the two diploid ancestral genomes, 1230 had both alleles represented in the 31 accessions while satisfying the constraint of being homozygous and having no missing data therein. 477 of these SNPs were kept because their PIC value was higher than

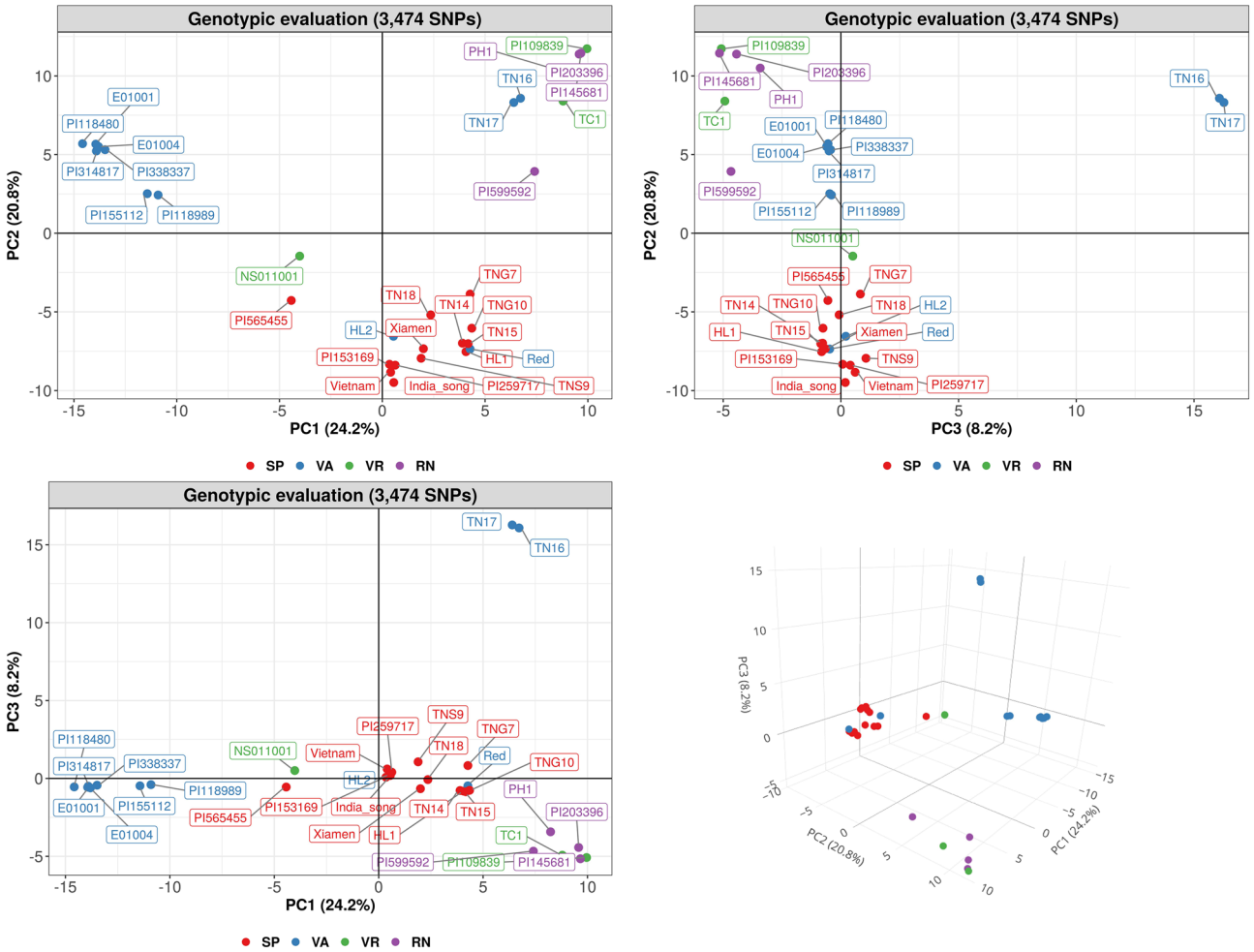


Figure 2. Principal component analysis (PCA) of the 31 accessions based on 3474 single nucleotide polymorphisms (SNPs). The 31 accessions were visualized by 4 market types, Spanish (SP), Valencia (VA), Virginia (VR), Runner (RN).

the average one of the 1230 homozygous SNPs (Supplementary Table S3). At the same time, we conducted a field experiment in the fall of 2016 to evaluate 8 agronomically quantitative traits for 24 of the 31 accessions used in RAD-seq and the other 282 peanut accessions of the TARI germplasm with 66 Spanish, 27 Valencia, 49 Virginia and 88 Runner type accessions. The summary statistics indicated that accessions from subsp. *fastigiata* had early maturity characteristics, more pods and higher yield but slightly less spot resistance compared to accessions from subsp. *hypogaea* (Supplementary Table S4). The phenotypic data from this trial enabled us to compare the capability of genotypic and phenotypic data to identify genetic relationships between peanut accessions (Supplementary Table S5).

We performed PCA separately for the genotypic data from 29 out of 477 SNPs with an average PIC value of 0.28 and for the phenotypic data (8 agronomic traits from 24 of the 31 accessions based on field experiments). For PCA based on 29 SNPs, 31 accessions were grouped into three clusters according to their botanical varieties (Fig. 3). In addition, with eigenvalues between 0.49 and 1.79, the first three PCs accounted for 67.8% of total variance (Supplementary Table S6). The top three SNP markers having the most contribution to three PCs were as follows: (1) PC1: B02_105774702, B04_1643180, B09_70140267 and B09_141920571, (2) PC2: A01_9265671, A02_65802170 and A01_90269752 and (3) PC3: B05_133797191, A09_45155599 and A01_90916564 (Supplementary Table S6). On the other hand, PC1, PC2 and PC3 in the PCA that relied on the phenotypic data of 8 agronomic traits cumulated 73.0% of the overall phenotypic variance, and these PCs had eigenvalues ranging from 1.46 to 2.41 (Supplementary Table S7). The top three traits contributing to three PCs the most were as follows: (1) PC1: yield, number of pods and days to flower, (2) PC2: plant architecture, number of pods and 100 seed weight and (3) PC3: leaf spot level, rust level and 100-pod weight (Supplementary Table S7). Unlike the PCA result using 29 SNPs, for the three biplots and 3D scatter plot from the PCA depending on phenotypic data, none provided much evidence for structure within the 24 of the 31 accessions (Supplementary Fig. S1), suggesting that 29 SNPs were able to better distinguish 31 accessions than 8 agronomic traits. These 29 SNPs were therefore designed as KASP markers.

To validate these 29 KASP markers, 282 accessions of the TARI germplasm with 66 Spanish, 27 Valencia, 49 Virginia and 88 Runner type accessions were genotyped. 14 out of 29 KASP markers showed a stable and

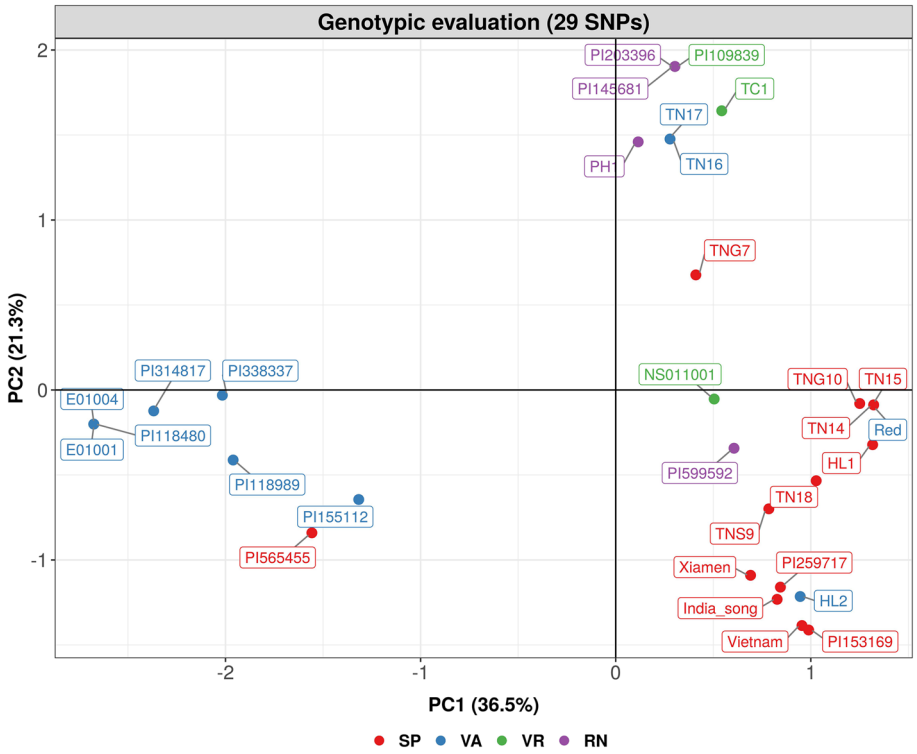


Figure 3. Principal component analysis (PCA) of the 31 accessions based on 29 single nucleotide polymorphisms (SNPs). 31 accessions were visualized by 4 market types, Spanish (SP), Valencia (VA), Virginia (VR), Runner (RN).

discernible genotyping result in the initial validation process. The population structure analysis of 282 accessions was then determined by PCA using either genetic distances between these 282 accessions calculated by the KASP-marker genotyping data (Supplementary Table S8) or the phenotypic data from the field experiment in the fall of 2016 based on 8 agronomic traits. The PCA biplots indicated that the PCA using the genotypic data performed better than the one using the phenotypic data to distinguish 282 accessions. For the scatter plots based on genomic data, PC1 and PC2 explained 36.9% and 18.3% of the variance of the genotyping data and separated these accessions into 3 groups according to three botanical varieties (subsp. *fastigiata* var. *vulgaris,* subsp. *fastigiata* var. *fastigiata* and subsp. *hypogaea* var. *hypogaea*). In addition, the KASP markers mostly contributing to the variance of PC1 and PC2 were B04_84804214, B09_6670331, A01_9265671, A02_65802170 and A05_80673567, A01_90916564 (Supplementary Table S9). On the other hand, the first two PCs from the PCA using phenotypic data accounted for 28.4% and 24.1% of phenotypic variation, and only quite roughly separated these accessions into two groups (subsp. *fastigiata* var. *vulgaris*/var. *fastigiata* and subsp. *hypogaea* var. *hypogaea*). Most Spanish and Valencia type accessions were difficult to distinguish using phenotypic data, and the grouping between Spanish/Valencia and Virginia/Runner accessions was less clear than in the PCA result based on genotyping data (Fig. 4). The major traits accounting for the variance of PC1 and PC2 were days to flowering, leaf spot level, yield and number of pods (Supplementary Table S10).

To further identify the population structure within the 282 accessions, discriminant analysis of principal components (DAPC) was performed based on increasing number of clusters (K) assigned by successive K-means. In such an approach, the Bayesian information criterion (BIC) was used to assess the model relevance, and the result showed that it was best to go to values of K of at least 3 for clustering the 282 accessions (Supplementary Fig. S2). Similarly, DAPC was conducted for 2, 3 and 4 clusters to explore the population structure of 282 accessions. At K = 2, clusters corresponded to Spanish/Valencia and Virginia/Runner type accessions. At K = 3, the 3 groups corresponded largely to Spanish, Valencia and Virginia/Runner. At K = 4, the overall grouping trend was similar to that with K = 3, but had a mixture of accessions from the three botanical varieties that were assigned into the fourth group (Supplementary Fig. S3). This result suggested that our KASP markers are effective for identifying the population structure of peanut germplasm according to the botanical varieties, and it can even illustrate the similar genetic background acquired by accessions corresponding to a mixture of botanical varieties.

Discussion

The worldwide peanut accessions accumulate more polymorphisms. Molecular markers are of importance in many aspects of plant genetics and breeding, including variety identification, positional cloning, and the exploration of genetic diversity and population structure within germplasm. Developing molecular markers in the cultivated peanut was challenging because of its narrow genetic diversity and the high sequence

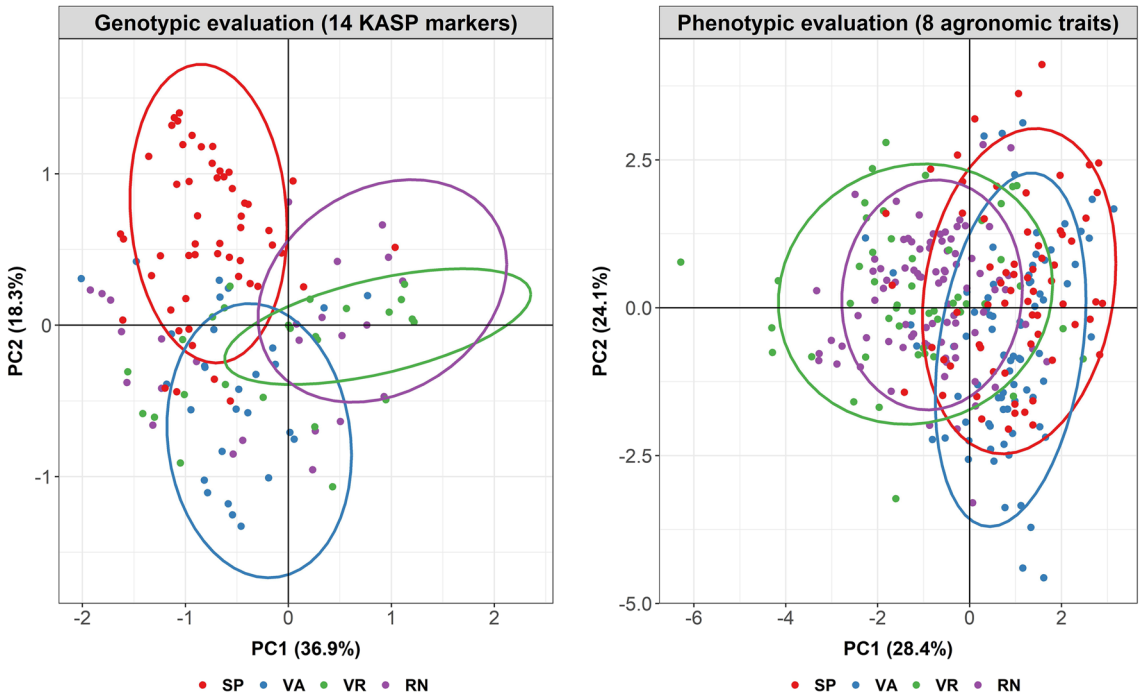


Figure 4. Principal component analysis (PCA) of 282 accessions based on genotyping data of 14 kompetitive allele-specific PCR (KASP) and phenotypic data. 282 accessions were visualized by 4 market types, Spanish (SP), Valencia (VA), Virginia (VR), Runner (RN).

similarity between its two diploid genomes[38]. The assembled genomes of cultivated peanuts and their diploid ancestors using NGS approaches has significantly boosted the genomic research in the peanut community. In particular, reduced-representation sequencing, such as GBS and RAD-seq, has been widely used in peanut research[20–22]. In this study, the RAD-seq approach was utilized to sequence 31 accessions of Taiwanese germplasm, decomposed into a "global" subset containing 17 "introduced" accessions and a "local" subset containing 14 Taiwanese accessions, 12 being current elite cultivars, 1 being a landrace, and 1 being an advanced breeding line. The global subset had a higher average number of SNPs than the local subset, suggesting that the germplasm from abroad had more polymorphisms than the local germplasm, and this can be explained by the fact that the accessions of the global subset were mainly introduced from North and South America encompassing the region of origin of domesticated cultivated peanut, supporting the idea that the origin of domesticated crops accumulates high diversity[39]. This result was compatible with previous work in soybean and sorghum based on SSR markers. Indeed, Iquira et al.[40] and Ghebru et al.[41] both found that the germplasm collection with accessions mostly from the origin of their cultivated crop had more unique alleles than the other collection with accessions from regions distant from the origin.

The introduced accessions are more genetically diverse than the local ones. The genetic diversity of these 31 accessions was then investigated using several approaches based on 3474 SNPs. As a whole, this panel had an average PIC, expected He, MAF and genetic distance of 0.16, 0.19, 0.87, 0.17 (Table 2), respectively, which was concordant with previous research using SNP genotyping[42–44]. Note that the PIC value is a marker's level of polymorphism. Markers are considered as highly informative (greater than 0.5), reasonably informative (0.25–0.5) and only slightly informative (smaller than 0.25) according to their PIC values[29]. The average PIC of 0.16 from these 31 accessions fell in this last class, while 32% of the identified SNPs corresponded to the reasonably informative class. In studies of three other germplasm collections genotyped by 48 K and 58 K SNP arrays, two collections comprising accessions of three botanical varieties like this study had a mean PIC value of 0.19[42,44], and the third germplasm collection, having only accessions from two botanical varieties, had the mean PIC value of 0.08[43], implying that the germplasm panel of 31 accessions chosen in this study preserves a high proportion of overall genetic diversity in spite of a smaller sample size compared to the ones in these three studies.

Focusing on subsets associated with the origin of our germplasm, Table 2 showed that the global subset (n = 17) had a higher average PIC value (0.15) than the local subset (0.14 with n = 14). Similarly, the average He and genetic distance of the global subset (He = 0.19, distance = 0.17) were greater than that of local subset (He = 0.16, distance = 0.14), indicating that the global subset had larger genetic diversity than the local subset. While these 31 accessions were separated into three botanical varieties, accessions from subsp. *fastigiata* var. *fastigiata* (Valencia type) had larger He (0.18), PIC (0.15) and genetic distance (0.15) on average than subsp. *fastigiata* var. *vulgaris* (Spanish type, He = 0.13, PIC = 0.11, distance = 0.11) and subsp. *hypogaea* var. *hypogaea* (Virginia/Runner types, He = 0.12, PIC = 0.10, distance = 0.11). In the 31 accessions, 11 genotypes were from

subsp. *fastigiata* var. *fastigiata* including 7 introduced accessions, 3 cultivars and 1 landrace; in particular, 5 of 7 introduced accessions were from countries in South America including Brazil, Uruguay, Peru and Venezuela. The cultivated peanut originated from South America[45], it is thus expected that genotypes of subsp. *fastigiata* var. *fastigiata* have larger genetic diversity than genotypes of subsp. *fastigiata* var. *vulgaris* and subsp. *hypogaea* var. *hypogaea* composing most cultivars in Taiwan. They also are expected to have higher diversity than this study's introduced accessions coming from regions away from the center of origin of cultivated peanut.

The peanut varieties developed in Taiwan may suffer from genetic vulnerability. Genetic relationships among the 31 accessions were investigated using the pairwise genetic distance for the construction of the dendrogram and PCA (Supplementary Table S2). In general, these 31 accessions were grouped into three clusters in line with three botanical varieties. Group I and Group II with mainly subsp. *fastigiata* var. *fastigiata* and subsp. *fastigiata* var. *vulgaris*, respectively, were separated by a genetic distance of 0.19, and Group III was separated from Group I/II by a genetic distance of 0.21 (Fig. 1). These results indicated that Group I and Group II were more closely related to one-another than to Group III, in agreement with the botanical classification and with other studies having larger sample sizes[12,16,22].

This dendrogram result also suggested that the local cultivars in Taiwan might be suffering from low genetic diversity due to the excessive exploitation of narrow genetic resources as breeding material, notably genotypes of Spanish type germplasm (Fig. 1). In the main clade within Group II, there were 9 accessions from the local subset containing 8 cultivars and 1 landrace. According to the pedigree of these 8 cultivars, many of them were genetically close to TNS9. TNS9 was developed in 1966 by pure line selection using the introduced line "Giay" from Vietnam, and this variety dominated more than 80% of peanut production in Taiwan in the 1980s because of its favorable flavor after roasting. This variety has also been widely exploited in peanut breeding programs in Taiwan, such as the development of TNG10, HL1, HL2, TN14 and TN18, all produced by the hybridization breeding method. Specifically, TNS9 was directly chosen as a parent of HL1, and indirectly contributed to the genetic background of the other four varieties by being selected as the parent of advanced breeding lines used in the breeding programs of these three varieties. Based on pedigree information, it is thus anticipated that HL2, a Valencia type variety, was grouped into the cluster with mostly Spanish type germplasm. On the other hand, TN16 and TN17, both rich in cyanidine-based anthocyanins on the seed coat, are 2 Taiwanese Valencia type cultivars derived from the same biparental breeding population using the hybridization of 2 landraces collected from central Taiwan. The dendrogram result showed that TN16 and TN17 were clustered into Group III; moreover, they were obviously separated from the other accessions in this clade. Therefore, these two varieties were not closely related to the three groups containing the other 29 accessions, suggesting that potentially locally collected genetic resources can still diversify the current Taiwanese germplasm. The same result of clustering was also found using PCA based on genetic distance (Fig. 2).

The KASP marker sets identify the population structure better than phenotypes. To understand the genetic diversity beyond the 31 accessions, 14 KASP markers developed by RAD-seq data of 31 accessions were utilized to assess the population structure of 282 peanut accessions from the germplasm conservation center in TARI. On the other hand, we also considered phenotypic data as an alternative tool for the assessment of population structure; specifically, 8 agronomic quantitative traits were evaluated in the field trial in the fall of 2016 using 306 accessions including 282 peanut accessions for the KASP marker validation and 24 out of 31 accessions used in RAD-seq.

The phenotyping results were consistent with similar field trials conducted in India and Turkey, which separated the subsp. *fastigiata* and subsp. *hypogaea* into two groups[46,47]. Similar to the results of two previous studies, we found that the subsp. *fastigiata* accessions in Taiwan had early maturity characteristics. However, our work showed that the subsp. *fastigiata* accessions have more pods than previously reported, their yield-related characteristics indicated the subsp. *fastigiata* accessions produce higher yields than subsp. *hypogaea* accessions in Taiwan (Supplementary Table S4). This result can be explained by the climate in Taiwan which influences the peanut breeding strategy. In terms of climate zones, Taiwan is separated into the north part belonging to the sub-tropical climate zone and the south part belonging to tropical climate zone allowing farmers to annually have two cropping seasons. However, the "plum rain" season between mid-May to mid-June and typhoons occurring between June and October can seriously damage the peanut yield in the end of the first cropping season or the beginning of the second cropping season, respectively. Thus, Taiwanese peanut breeders have chosen peanut accessions with early maturity characteristics, especially Spanish type peanuts, as breeding materials, reflecting the result in Supplementary Table S4 that Spanish type accessions have more pods than accessions from three other market types. For the PCA analyses based on phenotypic data, the 24 accessions used in RAD-seq could not be distinguished, and the 282 accessions used for KASP validation were grouped into two clusters mainly according to the subsp. *fastigiata* and subsp. *hypogaea* (Fig. 4, Supplementary Fig. S1). In the PCA of the 282 accessions, the traits playing the most important roles in PC1 and PC2 were days to flowering, leaf spot resistance level, yield and the number of pods, consistent with traits having significant difference between two peanut subspecies (Supplementary Tables S4, S10).

While we validated these KASP markers using 282 genotypes, the PCA results showed that these genotypes were distinctly separated into 3 groups according to three botanical varieties (Fig. 4). This conclusion was also supported by K-means clustering, in particular with the choice $K = 3$; beyond that value the BIC value didn't improve much (Supplementary Fig. S2). In addition, these KASP markers clearly distinguished subsp. *fastigiata* and subsp. *hypogaea* accessions at $K = 2$, and then separated var. *fastigiata* and var. *vulgaris* from the same subspecies *fastigiata* at $K = 3$, which was compatible with previous work[22]. However, when setting $K = 4$, the additional group had a mixture of four market types of germplasm belonging to all three botanical varieties, suggesting

that exchanges of genetic background among these accessions may have occurred. This result of a fourth cluster not corresponding to subspecies or market types was also reported in other works[12,42], and it may result from phenotyping difficulties[48,49]. In both sets, containing respectively 31 and 282 accessions, PCA was used to compare the effectiveness of molecular markers and phenotypic data to cluster samples, and it was demonstrated that PCA based on molecular markers provides more reproducible and satisfactory results than PCA based on phenotypic data (Figs. 2, 4, Supplementary Fig. S1).

Conclusion

Overall, the genetic diversity and relationship among peanut germplasm in Taiwan was revealed by SNPs identified through the RAD-approach. Our analyses suggest that one should broaden genetic diversity by introducing novel germplasm to prevent genetic vulnerability. In addition, the KASP markers successfully developed here could be useful tools for identifying the population structure of other peanut germplasm collections or for conducting further genetic studies related to breeding.

Data availability

The sequencing data of 31 accessions produced in this study have been deposited at the NCBI BioProject PRJNA811600. All the codes related to this project are available in the github site https://github.com/ymhsu/ahdivertwn.

Received: 17 March 2022; Accepted: 18 August 2022
Published online: 25 August 2022

References

1. Willett, W. *et al.* Food in the anthropocene: The EAT-Lancet Commission on healthy diets from sustainable food systems. *Lancet* **393**, 447–492 (2019).
2. Barkley, N. A., Upadhyaya, H. D., Liao, B. & Holbrook C. C. Global resources of genetic diversity in peanut (eds. Stalker, H. T. & Wilson, R. F.) 67–109 (AOCS Press, 2016).
3. Desmae, H. *et al.* Genetics, genomics and breeding of groundnut (*Arachis hypogaea* L.). *Plant Breeding* **138**, 425–444 (2019).
4. Moretzsohn, M. C. *et al.* Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome. *BMC Plant Biol.* **4**, 11. https://doi.org/10.1186/1471-2229-4-11 (2004).
5. Oteng-Frimpong, R., Sriswathi, M., Ntare, B. R. & Dakora, F. D. Assessing the genetic diversity of 48 groundnut (*Arachis hypogaea* L.) genotypes in the Guinea savanna agro-ecology of Ghana, using microsatellite-based markers. *Afr. J. Biotechnol.* **14**, 2484–2493 (2015).
6. Bertioli, D. J. *et al.* The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nat. Genet.* **48**, 438–446 (2016).
7. Zhuang, W. *et al.* The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nat. Genet.* **51**, 865–876 (2019).
8. He, G. *et al.* Microsatellites as DNA markers in cultivated peanut (*Arachis hypogaea* L.). *BMC Plant Biol.* **3**, 3. https://doi.org/10.1186/1471-2229-3-3 (2003).
9. Ferguson, M. E. *et al.* Microsatellite identification and characterization in peanut (*A. hypogaea* L.). *Theor. Appl. Genet.* **108**, 1064–1070 (2004).
10. Mace, E. S., Phong, D. T., Upadhyaya, H. D., Chandra, S. & Crouch, J. H. SSR analysis of cultivated groundnut (*Arachis hypogaea* L.) germplasm resistant to rust and late leaf spot diseases. *Euphytica* **152**, 317–330 (2006).
11. Tang, R. *et al.* Genetic diversity in cultivated groundnut based on SSR markers. *J. Genet. Genomics* **34**, 449–459 (2007).
12. Belamkar, V. *et al.* A first insight into population structure and linkage disequilibrium in the US peanut minicore collection. *Genetica* **139**, 411–429 (2011).
13. Ren, X. *et al.* Genetic diversity and population structure of the major peanut (*Arachis hypogaea* L.) cultivars grown in China by SSR Markers. *PLoS ONE* **9**, e88091. https://doi.org/10.1371/journal.pone.0088091 (2014).
14. Bertioli, D. J. *et al.* The genome sequence of segmental allotetraploid peanut *Arachis hypogaea*. *Nat. Genet.* **51**, 877–884 (2019).
15. Clevenger, J. *et al.* Genome-wide SNP genotyping resolves signatures of selection and tetrasomic recombination in peanut. *Mol. Plant* **10**, 309–322 (2017).
16. Pandey, M. K. *et al.* Development and evaluation of a high density genotyping 'Axiom_Arachis' Array with 58 K SNPs for accelerating genetics and breeding in groundnut. *Sci. Rep.* **7**, 40577. https://doi.org/10.1038/srep40577 (2017).
17. Otyama, P. I. *et al.* Genotypic characterization of the US peanut core collection. *G3 (Bethesda).* **10**, 4013–4026 (2020).
18. Baird, N. A. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS ONE* **3**, e3376. https://doi.org/10.1371/journal.pone.0003376 (2008).
19. Elshire, R. J. *et al.* A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. *PLoS ONE* **6**, e19379. https://doi.org/10.1371/journal.pone.0019379 (2011).
20. Zhao, Y. *et al.* QTL mapping for bacterial wilt resistance in peanut (*Arachis hypogaea* L.). *Mol. Breed.* **36**, 13. https://doi.org/10.1007/s11032-015-0432-0 (2016).
21. Han, S. *et al.* A SNP-based linkage map revealed QTLs for resistance to early and late leaf spot diseases in peanut (*Arachis hypogaea* L.). *Front. Plant Sci.* **9**, 1012. https://doi.org/10.3389/fpls.2018.01012 (2018).
22. Zheng, Z. *et al.* Genetic diversity, population structure, and botanical variety of 320 global peanut accessions revealed through tunable genotyping-by-sequencing. *Sci. Rep.* **8**, 14500. https://doi.org/10.1038/s41598-018-32800-9 (2018).
23. Niu, C. *et al.* A safe inexpensive method to isolate high quality plant and fungal DNA in an open laboratory environment. *Afr. J. Biotechnol.* **7**, 2818–2822 (2008).
24. Subrahmanyam, P. *et al.* Screening methods and sources of resistance to rust and late leaf spot of groundnut. *Inform. Bull.* **47**, 1–21 (1995).
25. Etter, P. D., Bassham, S., Hohenlohe, P. A., Johnson, E. A. & Cresko, W. A. SNP discovery and genotyping for evolutionary genetics using RAD sequencing. *Methods Mol. Biol.* **772**, 157–178 (2011).
26. Catchen, J., Hohenlohe, P. A., Bassham, S., Amores, A. & Cresko, W. A. Stacks: An analysis tool set for population genomics. *Mol. Ecol.* **22**, 3124–3140 (2013).
27. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. Preprint at https://arxiv.org/abs/1303.3997v2 Accessed 1 June 2021. (2013).
28. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

29. Botstein, D., White, R. L., Skolnick, M. & Davis, R. W. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *Am. J. Hum. Genet.* **32**, 314–331 (1980).
30. Weir, B. S. Diversity (ed. Weir, B. S.) 141–160 (Sinauer Associates, 1996).
31. Lê, S., Josse, J. & Husson, F. FactoMineR: An R package for multivariate analysis. *J. Stat. Softw.* **25**, 1–18 (2008).
32. Prevosti, A., Ocaña, J. & Alonso, G. Distances between populations of Drosophila subobscura, based on chromosome arrangement frequencies. *Theor. Appl. Genet.* **45**, 231–241 (1975).
33. Kamvar, Z. N., Brooks, J. C. & Grünwald, N. J. Novel R tools for analysis of genome-wide population genetic data with emphasis on clonality. *Front. Genet.* **6**, 208. https://doi.org/10.3389/fgene.2015.00208 (2015).
34. Jombart, T., Devillard, S. & Balloux, F. Discriminant analysis of principal components: A new method for the analysis of genetically structured populations. *BMC Genet.* **11**, 94. https://doi.org/10.1186/1471-2156-11-94 (2010).
35. Jombart, T. & Ahmed, I. adegenet 1.3-1: New tools for the analysis of genome-wide SNP data. *Bioinformatics* **27**, 3070–3071 (2011).
36. Paradis, E. & Schliep, K. ape 5.0: An environment for modern phylogenetics and evolutionary analyses in R. *Bioinformatics* **35**, 526–528 (2019).
37. Wickham, H. *et al.* Welcome to the tidyverse. *J. Open Source Softw.* **4**, 1686. https://doi.org/10.21105/joss.01686 (2019).
38. Pandey, M. K. *et al.* Advances in Arachis genomics for peanut improvement. *Biotechnol. Adv.* **30**, 639–651 (2012).
39. Hummer, K. E. & Hancock, J. F. Vavilovian centers of plant diversity: Implications and impacts. *Hort Sci.* **50**, 780–783 (2015).
40. Iquira, E., Gagnon, E. & Belzile, F. Comparison of genetic diversity between Canadian adapted genotypes and exotic germplasm of soybean. *Genome* **53**, 337–345 (2010).
41. Ghebru, B., Schmidt, J. & Bennetzen, L. Genetic diversity of Eritrean sorghum landraces assessed with simple sequence repeat (SSR) markers. *Theor. Appl. Genet.* **105**, 229–236 (2002).
42. Otyama, P. I. *et al.* Evaluation of linkage disequilibrium, population structure, and genetic diversity in the US peanut mini core collection. *BMC Genomics* **20**, 481. https://doi.org/10.1186/s12864-019-5824-9 (2019).
43. Abady, S. *et al.* Assessment of the genetic diversity and population structure of groundnut germplasm collections using phenotypic traits and SNP markers: Implications for drought tolerance breeding. *PLoS ONE* **16**, e0259883. https://doi.org/10.1371/journal.pone.0259883 (2021).
44. Nabi, R. B. S. *et al.* Genetic diversity analysis of Korean peanut germplasm using 48 K SNPs 'Axiom_Arachis' Array and its application for cultivar differentiation. *Sci. Rep.* **11**, 16630. https://doi.org/10.1038/s41598-021-96074-4 (2021).
45. Bertioli, D. J. *et al.* An overview of peanut and its wild relatives. *Plant Genet. Res.* **9**, 134–149 (2011).
46. Mallikarjuna, S. B. P., Upadhyaya, H. D., Kenchana Goudar, P. V., Kullaiswamy, B. Y. & Singh, S. Phenotypic variation for agronomic characteristics in a groundnut core collection for Asia. *Field Crop Res.* **84**, 359–371 (2003).
47. Yol, E., Furat, S., Upadhyaya, H. D. & Uzun, B. Characterization of groundnut (*Arachis hypogaea* L.) collection using quantitative and qualitative traits in the Mediterranean Basin. *J. Integr. Agric.* **17**, 63–75 (2018).
48. Holbrook, C. C. & Stalker, H. T. *Peanut Breeding and Genetic Resources* (ed. Janick, J.) 297–356 (Wiley, 2002)
49. Barkley, N. A. *et al.* Genetic diversity of cultivated and wild-type peanuts evaluated with M13-tailed SSR markers and sequencing. *Genet. Res.* **89**, 93–106 (2007).

Acknowledgements

We thank O.C. Martin, T. Blein and K.H. Chen for discussions. The authors also thank Dr. K.Y. Chen at National Taiwan University for providing resources and C.Y. Liu for the guidance while preparing RAD-seq libraries.

Author contributions

H.Y.D. and S.S.W. conceived the study and acquired the funding. Y.M.H. designed the bioinformatics and data analysis pipeline and wrote the manuscript. S.R.L., H.F., W.C.H. and H.I.K. performed the experiments. Y.C.T. supervised the student. All authors contributed to manuscript revisions.

Funding

This research was supported by funding from the Council of Agriculture of Taiwan (Project ID: 105AS-9.3.1-CI-C3 and 106AS-8.2.1-CI-C3). The work of Y-M. Hsu was supported by a PhD Grant provided by the Ministry of Education (Taiwan) and Université Paris-Sud/Saclay. IPS2 benefits from the support of Saclay Plant Science-SPS (ANR-17-EUR-0007).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-18737-0.

Correspondence and requests for materials should be addressed to H.-Y.D.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

3.2 The exploitation of genetic diversity for disease resistance - tomato breeding for bacterial wilt (Ralstonia sp.) resistance

3.2.1 *Ralstonia solanacearum* - the pathogen leading to bacterial wilt

Caused by *Ralstonia solanacearum*, bacterial wilt (BW) is a serious plant disease which can be found in more than 200 plant species including economically important crops such as potato, eggplant, and tomato. Having a broad host range, *R*. *solanacearum* species complex (RSSC) contains diverse strains that make it one of the most damaging plant pathogenic bacteria around the world (Denny, 2006; Genin, 2010). As a soil-borne pathogen, *R*. *solanacearum* infects plants through wounds induced by cultivation, insects or even lateral root emergence. Then, this bacterium populates from the root cortex and progressively penetrates into the xylem vessels, stem and aerial parts. Their rapid growth in the xylem eventually devastates the vascular system, affecting water transport therein, and resulting in the wilting symptoms and death of plants. *R*. *solanacearum* tends to grow and spread in high temperature (24 - 35 ℃) and moist soils (-0.5 to -1 bar), leading to the BW occurrence in tropical, subtropical and warm temperate regions worldwide. The environment and types of soils are crucial for the survival of this bacterium that influences the BW development (Denny 2006, Ramesh & Bandyopadhyay, 1993). It was reported that *R*. *solanacearum* remained alive up to 40 years without a host plant in warm soils (20 - 25 ℃) but showed different degrees of tomato infection while using various soils as the inoculum.

Traditionally, *R*. *solanacearum* are classified into five races based on the host range, including race 1 (solanaceous vegetables), race 2 (banana), race 3 (potato and tomato in temperate conditions), race 4 (ginger) and race 5 (mulberry), indicating that *solanaceae* plant species are mainly infected by

race 1 and 3 of *R*. *solanacearum*. This pathogen is grouped into six biovars on the basis of the utilization of carbon sources (Singh et al., 2015). Based on the sequence analysis of the 16S and 23S gene intergenic spacer region (ITS), the endoglucanase gene (*egl*) and the *hrpB* gene, the *R*. *solanacearum* are categorized into four genetic groups, referring to four phylotypes. Each phylotype has strains with corresponding geographical origin, indicating that (a) phylotype I strains originate mainly from Asia, (b) phylotype II strains originate from America, (c) phylotype III from Africa, and (d) phylotype IV from Indonesia and Australia (Fegan & Prior, 2005). Furthermore, Wicker et al (2012) performed multilocus sequence analysis (MLSA) to trace the evolutionary history of these four phylotypes, and identified seven chromosome housekeeping genes and two megaplasmid virulence-associated genes. Their results showed that phylotype IV is the most ancestral and distinct phylotype as well as the main donor. Ongoing diversification within phylotypes suggests that the evolutionary potential is possessed probably through the spread and adaptation of pathogens in different regions and host ranges. For example, both studies indicated that novel sequevars within phylotype I were identified in India and Taiwan using the *egl* gene sequence (Lin et al., 2014; Ramesh et al., 2014).

3.2.2 Tomato genetic resources for resistance to bacterial wilt

It has been shown that BW results in a considerable yield loss in tomato (Hartman et al., 1991; Karumannil et al., 2008), and developing resistant cultivars is the most effective approach to decrease such losses (Hanson et al., 2016). The resistance sources are found in cultivated tomato and wild species, such as *Lycopersicon esculentum* var. *cerasiforme*, *Lycopersicon pimpinellifolium* and *Lycopersicon peruvianum*, but unfortunately the BW resistance genes are linked with genes controlling small fruit sizes (Jyothi et al ., 2012). In addition, a durable resistance that adapts to all environments is

unlikely to be obtained since the resistance of host plants can be overcome by the combinatorial factors including soil type, temperature, rainfall and pathogen strains (Hayward, 1991). Wang et al (1998) carried out an experiment to evaluate the BW resistance of 35 tomato lines in 11 fields located in 11 countries. These 35 resistant tomato lines include wild accessions, breeding lines and commercial varieties identified by different breeding programs. It turned out none of these resistant lines was immune to BW. Thirty-five lines showed the average survival percentage ranging from 24.1% to 97%, and the mean survival percentage of all accessions of each 11 locations ranged from 33.7% to 86.6%. Among seven accessions with mean survival percentage over 90%, H7996 exhibits the most stable BW resistance in multiple locations.

Different studies concentrated on advanced recombinant inbred line populations. For instance, 188 F_9 lines that derived from the cross between *L. esculentum* cv. H7996 and susceptible *L. pimpinellifolium* “West Virginia 700” (Wva700) were produced (Carmeille et al., 2006, Wang et al., 2013). Using a R. solanacearum race 3-phylotype II strain, the major QTL *Bwr-6* was identified that explains 29.8% of the phenotypic variation (Carmeille et al., 2006). On the other hand, Wang et al. (2013) evaluated the same population using the race 1-phylotype I strain, and identified two major QTLs, *Bwr-6* and *Bwr-12*, which explain respectively 22.2% and 56.1% of the variation, suggesting that *Bwr-6*, identified in both studies using different strains, has a broad-spectrum resistance. *Bwr-6* and *Bwr-12* were localized respectively to regions of 15.5-cM and 2.8-cM intervals of chromosome 6 and 12. Further studies for developing markers tightly linked to these two QTLs have progressed (Abebe et al., 2020; Shin et al., 2020), but the candidate genes conferring the BW resistance still need to be confirmed by further studies. For identifying more QTLs, GWAS was applied to 191 cultivated varieties based on race 1 strain (Nguyen et al., 2021). In addition to the two QTLs mentioned

above, a major QTL, *Bwr-4*, and four environment-specific QTLs located in chromosome 1 and 8 to 10 were detected in this study.

scientific reports

Check for updates

OPEN

Whole genome resequencing and complementation tests reveal candidate loci contributing to bacterial wilt (*Ralstonia* sp.) resistance in tomato

Derek W. Barchenger[1]✉, Yu-ming Hsu[2], Jheng-yang Ou[3], Ya-ping Lin[1], Yao-cheng Lin[3], Mark Angelo O. Balendres[4], Yun-che Hsu[1], Roland Schafleitner[1] & Peter Hanson[1]

Tomato (*Solanum lycopersicum*) is one of the most economically important vegetable crops worldwide. Bacterial wilt (BW), caused by the *Ralstonia solanacearum* species complex, has been reported as the second most important plant pathogenic bacteria worldwide, and likely the most destructive. Extensive research has identified two major loci, *Bwr-6* and *Bwr-12*, that contribute to resistance to BW in tomato; however, these loci do not completely explain resistance. Segregation of resistance in two populations that were homozygous dominant or heterozygous for all *Bwr-6* and *Bwr-12* associated molecular markers suggested the action of one or two resistance loci in addition to these two major QTLs. We utilized whole genome sequence data analysis and pairwise comparison of six BW resistant and nine BW susceptible tomato lines to identify candidate genes that, in addition to *Bwr-6* and *Bwr-12*, contributed to resistance. Through this approach we found 27,046 SNPs and 5975 indels specific to the six resistant lines, affecting 385 genes. One sequence variant on chromosome 3 captured by marker Bwr3.2dCAPS located in the *Asc* (*Solyc03g114600.4.1*) gene had significant association with resistance, but it did not completely explain the resistance phenotype. The SNP associated with Bwr3.2dCAPS was located within the resistance gene *Asc* which was inside the previously identified *Bwr-3* locus. This study provides a foundation for further investigations into new loci distributed throughout the tomato genome that could contribute to BW resistance and into the role of resistance genes that may act against multiple pathogens.

Tomato (*Solanum lycopersicum* L.) is widely grown and one of the most economically important vegetable crops worldwide. Global production of tomatoes has continuously increased for the past 50 years, especially in tropical and subtropical regions. Tomato crops can be infected by disease-causing bacterial, fungal, and viral pathogens that can reduce yield, fruit quality, shelf-life, and nutritional content. Bacterial wilt (BW), caused by the *Ralstonia solanacearum* species complex (RSSC), is one of the most destructive plant pathogenic bacteria[1]. The RSSC is favored by high temperatures and humidity, and, as extreme weather events become more frequent and severe through climate change, it is anticipated that BW will become more common and destructive. Management of BW with pesticides is not a viable option because the pathogen survives in the soil for many years and has a wide host range[2]. Other management strategies include soil solarization, which is of limited effectiveness due to the existence of the pathogen deep in the soil. An integrated approach has been identified as the best way to manage the disease, including irrigation management, grafting, crop rotation, sanitation (removing weeds and plant debris and also cleaning farm equipment), and managing insect and nematode pests. Host resistance is the single most effective management strategy associated with BW[3] and planting resistant cultivars is the cheapest, simplest, and most environmentally friendly approach to limit losses[4]. Sources of resistance to BW originating

[1]World Vegetable Center, Tainan, Taiwan. [2]CNRS, INRAE, Institute of Plant Sciences Paris-Saclay (IPS2), Univ Evry, Université Paris-Saclay, 91405 Orsay, France. [3]Biotechnology Center in Southern Taiwan, Agricultural Biotechnology Research Center, Academia Sinica, Tainan, Taiwan. [4]Institute of Plant Breeding, College of Agriculture and Food Science, University of the Philippines Los Baños, Los Baños, Laguna, Philippines. ✉email: derek.barchenger@worldveg.org

from cultivated tomato and its close wild relatives, *S. pimpinellifolium* and *S. lycopersicum* var. *cerasiforme,* have been identified, but none are immune and expression of resistance is strongly influenced by pathogen strain, temperature, soil pH and the interactions among these factors[3]. Furthermore, BW resistance has been associated (linked) with small fruit weight, bitter flavor, susceptibility to root-knot nematodes, and other negative traits[5]. Variable reaction of BW resistance sources[6] coupled with quantitative inheritance of resistance complicates conventional breeding and development of resistant cultivars.

A coordinated multilocation testing of a set of resistance sources by a team of collaborators following comparable testing and evaluation protocols identified 'Hawaii 7996' (H7996) as one of the most stable resistance sources with a high survival rate across 12 field trials in 11 countries[7]. Later, INRA-CNRS, University of the Philippines Los Baños, and the World Vegetable Center (WorldVeg) developed an advanced recombinant inbred line (RIL) population (188 F_9 lines) derived from the cross of H7996 by susceptible *S. pimpinellifolium* 'West Virginia 700' (WVa700). Multi-location testing of this mapping population in nine trials, seven in Asia and two in Reunion Island, revealed the presence of two major genomic regions (*Bwr-6* and *Bwr-12*) conditioning BW resistance, as well as additional QTLs with minor or strain-specific effects[8], supporting the findings of Carmeille et al.[9] who reported major QTLs on chromosome 6 (*Bwr-6*) and minor QTLs on chromosomes 3, 4, and 8 (*Bwr-3*, *Bwr-4*, and *Bwr-8*, respectively). The molecular markers developed for the selection of *Bwr-6* and *Bwr-12* QTLs are certainly useful[4,9–11]; however, they do not completely explain the resistant phenotype and have some level of mismatch resulting in false positives and selection of susceptible individuals[12].

The QTL *Bwr-12*, located in a 2.3-cM interval of chromosome 12, accounted for much of the phenotypic variation for resistance to phylotype I isolates (recently reclassified as *R. pseudosolanacearum*)[12]. Virus-induced gene silencing assays suggested the involvement of leucine-rich repeat receptor-like kinases *Solyc12g009520* and *Solyc12g009550* located in the *Bwr-12* QTL interval with resistance to phylotype I strains[13]. Through whole genome resequencing, Kim et al.[14] identified four genes that encode putative leucine-rich repeat receptor-like proteins that were associated with resistance to BW on chromosome 12. The authors reported one SNP marker in the gene *Solyc12g009690.1* that could be tightly linked to *Bwr-12*. However, in our analysis this marker does not improve selection accuracy for BW resistance beyond previously developed molecular markers linked to the trait (unpublished data). The QTL *Bwr-6* encompasses a 15.5-cM region on chromosome 6 that may include one or more important QTLs for resistance to phylotype II isolates (classified as *R. solanacearum*) as well as more broad-spectrum resistance[12]. *Bwr-6* is a large region and molecular markers in these regions do not completely explain the broad-spectrum resistance in the offspring of 'H7996[14]. Recent efforts focused on fine-mapping the *Bwr-6* and *Bwr-12* regions to identify important resistance loci and closely linked markers have been promising[15]. The authors identified four QTLs associated with strain-specific resistance on chromosome 6 and three on chromosome 12, explaining 14–54% of the overall variability. For validation, they used a set of 80 near-isogenic lines (NILs) derived from the RILs developed by Wang et al.[8] and found significant association with the phenotype[15]. Field trials of H7996 and WorldVeg tomato lines homozygous for *Bwr-12* and *Bwr-6* under BW pressure in Benin revealed that the WorldVeg lines did not demonstrate high levels of resistance like H7996[16]. This result suggests that H7996 carries additional major BW QTL besides *Bwr-12* and *Bwr-6*. The objective of this study was to identify loci contributing to BW resistance besides *Bwr-6* and *Bwr-12* to support breeding for more durable resistance in tomato varieties.

Results and discussion

Disease resistance levels among tomato lines. None of the lines had complete resistance to both pathogen strains (Pss4 and Pss1632) in these trials, including H7996, the best-known tomato resistance source (Table 1). Wilting can occur in BW resistant tomato lines, the extent of which depends on pathogen strain, temperature, and other environmental conditions[12,17,18]. However, the proportion of wilted plants in resistant lines was usually less than in susceptible lines (Table 1). The six lines in the resistant group selected for whole genome sequencing had higher levels of resistance to both pathogen strains (average of 95 and 83% resistance to Pss4 and Pss1632, respectively) compared to the performance of the nine susceptible lines (average of 28 and 19% resistant plants for Pss4 and Pss1632, respectively) (Table 1). Both groups typically had slightly higher levels of resistance to Pss4 than Pss1632. Within the susceptible group, there were large differences in symptom expression between and within pathogen strains. TBL-2, Pant Bahar, and L390 were highly susceptible to both strains. CRA84-23–1 115 was highly resistant to Pss4 (90% resistant) but highly susceptible to Pss1632 (10% resistant) (Table 1). CRA84-57-1 140, T-245, and ST/2 had moderately low levels of resistance to both strains (Table 1). These results support the extensive body of literature highlighting the complexity of host-pathogen interactions in the tomato-BW pathosystem, as reviewed by Hayward et al.[3]. Furthermore, the higher level of virulence of Pss1632 was previously reported[12]. When challenged with Pss4, LS-89 and F7 80 Pink were the most resistant accessions (100% resistant), while Pant Bahar, L390, and LA3501 were the most susceptible (0%) (Table 1). The accession F_7 80-465-10-pink was the most resistant to Pss1632 (92.5%), while TBL-2 was the most susceptible (100% of symptomatic plants) (Table 1). The resistant and susceptible reactions of the accessions screened in this study were generally in alignment with the previous work of Kunwar et al.[12] employing a partly overlapping set of materials. Hai et al.[17] reported that LA3501 was resistant to BW strain Pss186 but susceptible to Pss4. Strain- and environment-specific reactions have been previously reported[8,12] and these will likely limit the development of widely applicable molecular markers associated with BW resistance. To account for the variability of resistance in the accessions, only the five most resistant or most susceptible individual plants per accession were selected for sequencing and downstream analysis.

Whole genome sequencing of 15 tomato varieties for genome wide variant detection. The read depth of the sequencing ranged from 24.7 × (LE415 Anagha) to 56.8 × (H7997), with an average read depth

Tomato line	Country of origin	Resistant percent screened against Pss4	Resistant percent screened against Pss1632	Average percent resistance
LS-89	Japan	100	85	92.5
Hawaii 7997	USA	95	82.5	88.8
F_7 80-465-10-pink	Philippines	85	92.5	88.8
F_7 80 pink	Philippines	100	72.5	86.3
Hawaii 7996	USA	95	75	85
LE415 Anagha	India	95	90	82.5
CRA84-23-1 115	Guadeloupe	90	15	52.5
CRA84-57-1 140	Guadeloupe	60	30	45
T-245	Sri Lanka	40	35	37.5
S/T2	Philippines	30	35	32.5
Rodade	South Africa	20	25	22.5
LA3501	USA	0	20	10
TBL-2	France	10	0	5
L390	Taiwan	0	10	5
PantBahar	India	0	5	2.5

Table 1. Average resistance percentage of the highly resistant and highly susceptible tomato lines used for sequencing two weeks after inoculation with two different strains of *Ralstonia* sp, Pss4 (race 1, biovar 3, *R. pseudosolanacearum*) and Pss1632 (race 3, biovar 2, *R. solanacearum*), during the hot season (June–July) in 2018. Five individual plants with extremes in the phenotype (highly susceptible early in the evaluation, highly resistant late in the evaluation) were selected for sequencing.

Tomato line	Estimated read depth	Genome coverage ratio (%)	Properly mapped paired reads (%)	All SNPs	All InDels	Homozygous SNPs	Homozygous InDels	Phenotypic response
F7 80-465-10-pink	46.4	99.2	99.2	529,584	207,522	327,246	166,770	R
LE415 Anagha	24.7	98.9	99.5	410,103	172,062	157,920	135,478	R
LS 89	35.3	98.7	98.7	1,643,618	303,559	1,327,260	251,884	R
Hawaii 7997	56.8	98.8	99.0	876,848	223,157	634,321	181,368	R
Hawaii 7996	34.2	98.9	98.9	1,136,702	247,511	849,093	201,316	R
F7 80 pink	44.1	99.4	99.3	534,965	213,438	327,984	168,569	R
TBL-2	41.7	99.2	99.3	627,186	196,732	352,923	155,867	S
Pant Bahar	25.3	98.6	99.4	359,227	157,239	136,709	126,529	S
L390	32.5	99.5	99.0	397,321	185,729	225,619	154,056	S
CRA84-23-1 115	26.4	98.6	99.0	991,748	221,898	602,564	170,932	S
LA3501	27.8	98.4	98.6	1,637,262	315,105	1,331,932	263,758	S
Rodade	26.5	99.1	99.2	606,730	192,083	392,549	159,238	S
CRA84-57-1 140	53.0	99.3	99.3	689,382	220,246	219,025	153,402	S
T-245	51.9	99.6	99.2	1,023,995	244,588	113,967	130,427	S
S/T2	52.7	99.0	99.0	1,040,560	237,605	775,087	191,734	S

Table 2. Summary statistics of the sequence quality, coverage and polymorphisms of the bacterial wilt (Pss4 (race 1, biovar 3, *Ralstonia pseudosolanacearum*) and Pss1632 (race 3, biovar 2, *R. solanacearum*) resistant and susceptible tomato lines.

of 38.6 × (Table 2). Genome coverage and properly mapped pair-end reads were always greater than 98% in our experiment (Table 2). When compared to the 'Heinz 1706' annotated genome (v. SL4.0), we identified an average of 883,682 SNPs and 222,565 indels. LS-89 had the greatest number of SNPs, at 1,643,618 followed by LA3501 with 1,637,262, while the greatest number of indels were identified for LA3501 (Table 2). The highly susceptible cultivar Pant Bahar had the fewest number of SNPs and indels with 359,227 and 157,239, respectively (Table 2). The number of polymorphisms identified in our study is in line with several other studies using different accessions of domesticated tomato species[19–21], which was generally fewer than 2 million SNPs, although results were based on different versions of the 'Heinz 1706' reference genome.

Three resistant and six susceptible accessions (F7_80P, F7_80465P, CRA84-57-1, L390, LE415, Pant Bahar, Rodade, T-245, and TBL-2) formed a distinct cluster based on similarities in the high-quality SNPs identity in this study (Fig. 1). However, the highly unique and BW susceptible line LA3501 had a strong interactive force on the other accessions, which could make this cluster of lines appear more similar than they actually were. LA3501 contains an introgression on chromosome 6 derived from *S. pennellii* which provides strain-specific BW resistance[17]; this DNA fragment probably contributed to the genetic uniqueness of this line compared to most

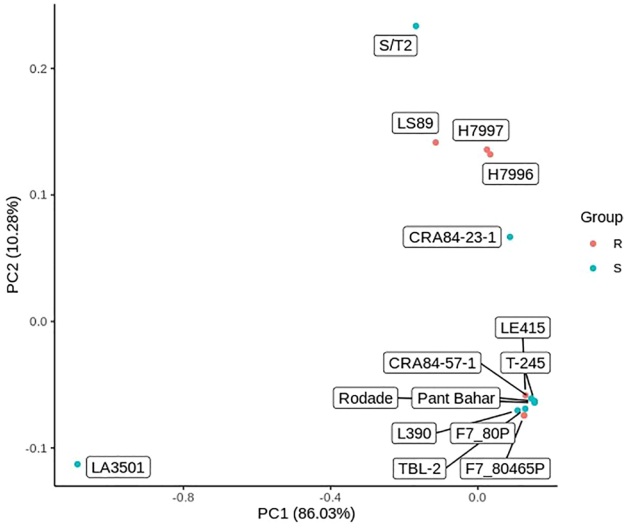


Figure 1. The Principal Coordinate Analysis based on all of the high-quality polymorphisms of the bacterial wilt (*Ralstonia* sp.) resistant (R; red) and susceptible (S; blue) tomato lines. H7796 is Hawaii 7996 and H7997 is Hawaii 7997.

other lines in our study. We found that H7996 and H7997 were genetically similar while the other accessions in our study appeared more unique (Fig. 1).

We compared the SNP distribution of all accessions, and found that the six resistant accessions had higher SNP density in the regions around *Bwr-6* and *Bwr-12* than the nine susceptible accessions (Fig. 2 and Supplemental Fig. 1). However, we also observed that resistant and susceptible lines shared many regions with similar SNP distribution (Fig. 2). Since our objective was to identify loci that contribute to BW resistance not explained by *Bwr-6* and *Bwr-12*, those regions with similar SNP distributions common in resistant and susceptible accessions were removed from further consideration as candidates for discovery of new resistant loci. To comprehensively screen the candidate polymorphisms that contributed to resistance, we compared each resistant accession with all nine susceptible accessions, and removed SNPs that were identified in any of the susceptible accessions. This comparison allowed us to extract variants that are uniquely found in each resistant line but not in any of the susceptible lines.

In the first stage of comparison, we retained only homozygous polymorphisms for further analysis. The accessions had an average of homozygous 518,279 SNPs and 174,088 indels (Table 2). Then, we compared each of the six resistant lines individually with all nine of the susceptible lines and retained variants that were uniquely identified in resistant lines. With these two filters, only about 8% of total variants of resistant accessions were retained. Among the resistant accessions, LS-89 had the greatest number of unique variants with 313,359 SNPs and 42,444 indels, while the other resistant accessions have an average of 27,046 unique SNPs and 5,975 unique indels (Fig. 3). Kim et al.[14] conducted a similar analysis using two susceptible and seven resistant accessions, including H7996, for comparison and found 5,259 SNPs to be polymorphic between resistant and susceptible groups. LS-89 is a BW-resistant rootstock cultivar developed in Japan originating as a selection from either H7996[22] or H7998[23], although both H7996 and H7998 were reported to originate from the same source (PI 127805A)[24]. However, it is possible that H7996, H7997 and several other Hawaii-prefixed lines were selections out of a genetically diverse accession 'HSBW' (Hot Set Bacterial Wilt)[25]. LS-89 should not differ greatly from H7996 but we found that LS-89 was genetically distinct from H7996, H7997 and the other resistance sources in our experiment (Fig. 4) although it was not compared with H7998 which was not included in our analysis. LS-89 might be derived from a different HSBW selection but since this original source is lost, no follow-up is possible. There is a chance that the seed source held by the World Vegetable Center is incorrect, despite it having a similar resistance reaction as the original LS-89[26].

Comparison of WGS variants with QTL mapping. Based on these polymorphisms specific to resistant lines, we compared them among the 6 resistant lines and previous studies that identified QTLs associated with the bacterial wilt resistance. The proportion of common polymorphisms among the resistant tomato lines varied across the chromosomes (Fig. 4). Only two polymorphisms on chromosome 12 were common among all six resistant lines (Fig. 4), which were near but not within the previously identified resistance QTL *Bwr-12*[8,14]. The number of unique polymorphisms were high and ranged from 196,901 on chromosome 2 to 1,429 polymorphisms on chromosome 10 (Fig. 4). There were 25 polymorphisms that were common among 5 of the 6 resistant lines and 66 polymorphisms that were common among 4 of the resistant lines (Fig. 5), all of which were within the region previously identified by Kim et al.[14] and near the large resistance QTL *Bwr-6* (22.2–39.6 Mb)[8]. Multiple QTLs within the large *Bwr-6* and *Bwr-12* loci have been previously reported[15]; therefore, the common polymorphisms on chromosomes 6 and 12 found here warrant further investigation as they could be within candidate genes contributing to resistance that are linked to the major QTLs *Bwr-6* and *Bwr-12* but have not yet been

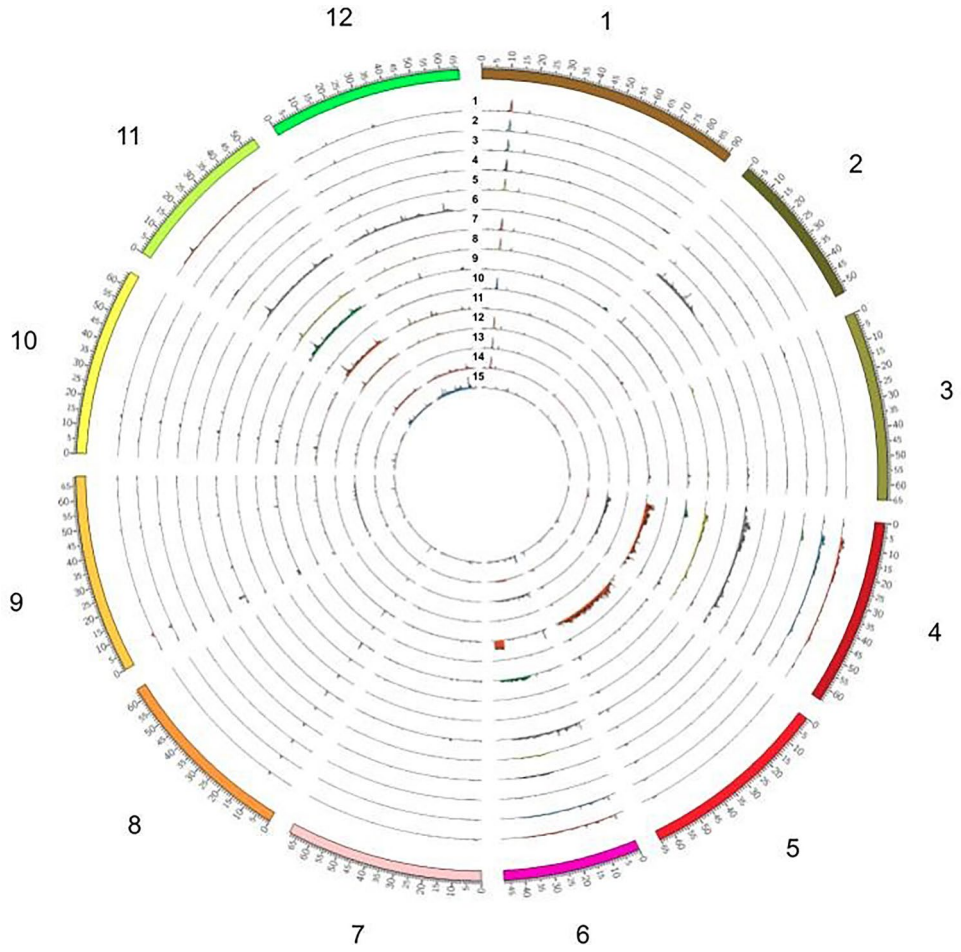


Figure 2. The distribution of SNPs across the genome for 15 bacterial wilt (*Ralstonia* sp.) resistant and susceptible tomato lines. The histograms represent the number of SNPs in 100-kb for the 15 tomato accessions. The lines are numbered (1) Hawaii 7996, (2) Hawaii 7997, (3) LE415, 4) F7_80P, (5) F7_80465P, (6) LS89, (7) Bahar, (8) CRA84_115, (9) CRA84_140, (10) L390, (11) LA3501, (12) Rodade, (13) ST2, (14) T_245, and (15) TBL_2.

fully characterized. The majority of the unique polymorphisms were from LS-89 (Fig. 3), which underlies the genetic distinctiveness of this line (Fig. 5). Interestingly, we found that our other resistance sources form two distinct clusters based on genetic similarity, with H7996 and H7997 being similar and with F7_80P and F7_80465P being extremely similar and clustering closely with LE415 (Fig. 4). This genetic structure could be a contributing factor in the overall lack of common polymorphisms in our study and a preponderance of polymorphisms that were common among only two or three sources.

We then predicted the functional effects of variants uniquely identified in 6 resistant lines targeting protein-coding genes. The vast majority of the variants were detected in intergenic or intronic regions (Fig. 3), with fewer than 1,000 SNPs being located in genic regions in most entries with the exception of LS-89, which contained 6,500 SNPs in protein-coding regions (Supplemental Table 1). For the variants in UTR, the 3′UTRs had 1.64 to 2.65 times more variants than 5′UTRs. The ratio of nonsynonymous and synonymous mutation ranged from 0.56 to 0.94. Frameshift mutations were the most frequent type of mutation we identified (Supplemental Table 1).

The details of candidate genes are provided in Supplemental Table 2. A large number of polymorphisms were unique to LS-89 and not present in the other resistant lines. In total, we found high impact mutations specific to the six resistant lines in 385 genes. The polymorphisms identified here were not uniformly distributed among the 12 chromosomes and most were located on chromosomes 2 and 4 (Fig. 4 and Supplemental Fig. 2). Using H7996, Kim et al.[14] found 265 resistant-specific SNPs located in coding regions, with most SNPs located on chromosomes 6 and 12 near *Bwr-6* and *Bwr-12* QTLs.

As expected, the three parental lines (CLN3641F1-5-11-14-4-25-20-11-7(F), CLN4018F1-6-7U14-29-21-14-5 and H7996) were resistant against BW strain Pss4 used in our experiment. Based on molecular marker results, all F_2 plants in both mapping populations had either the homozygous dominant or heterozygous alleles at *Bwr-6* and *Bwr-12*, as did the three parental lines (Supplemental Table 3). The two F_2 populations showed different segregation patterns for inheritance of resistance to Pss4 strain: CLN4397-4 did not deviate significantly from a 3:1 (resistant to susceptible) ratio while CLN4398-8 showed a 9:7 ratio (Table 3). Given that the populations were homozygous for both *Bwr-6* and *Bwr-12*, there were apparently two additional independent loci contributing

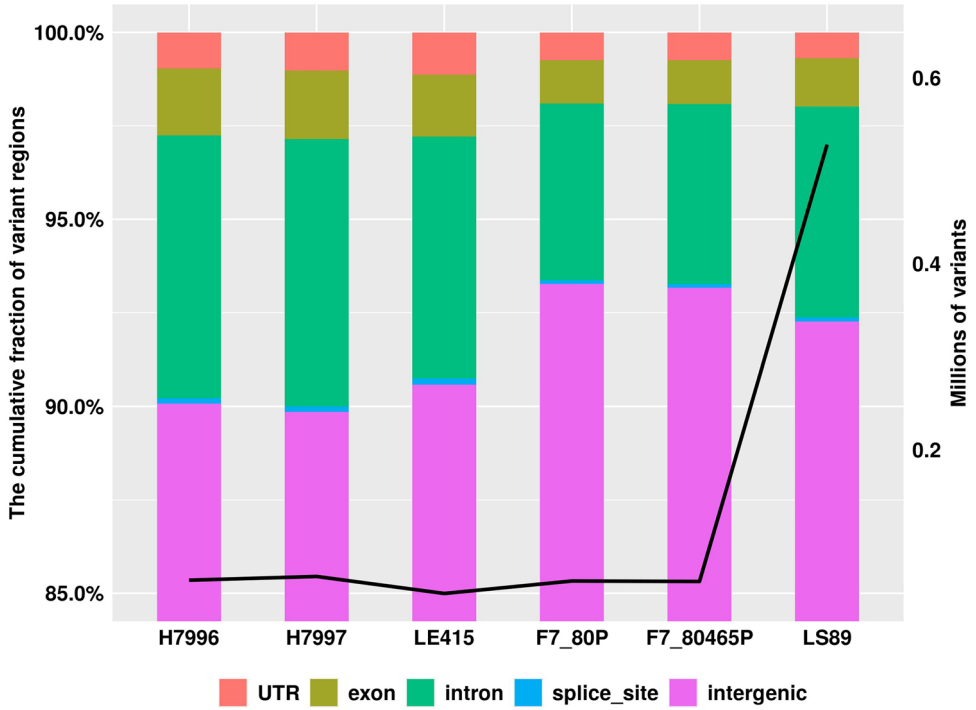


Figure 3. The proportion and number of SNPs acquired by genomic features of the six highly bacterial wilt (*Ralstonia* sp.) resistant tomato lines. The bars represent the proportion of genomic features in which SNPs of tomato lines are located, and the black line is the number of SNPs contained in each of the tomato lines. In the legend, "UTR" includes 5′UTRs and 3′UTRs, and "splice_site" includes the donors, receptors and regions of splice sites.

to resistance in CLN4398-8 and one additional independent locus in CLN4397-4. The role of multiple loci or complex inheritance patterns associated with resistance to BW in tomato has been widely reported[8,15,27–33], which supports our findings. However, one study identified a single dominant gene conferring resistance to BW in H7996[34] and H7998[35]. The difference in findings is not necessarily contradictory but could be due to different pathogen strains used for screening in inheritance studies.

Validation of CAPS markers in two F2 populations confirmed resistant genes to bacterial wilt. To validate the identified polymorphisms, molecular markers were developed and first tested in the parental lines (CLN3641F1-5-11-14-4-25-20-11-7(F), CLN4018F1-6-7U14-29-21-14-5 and H7996) of our segregating populations (Table 4). Selection of polymorphisms for molecular marker development was based on the presence of the polymorphism in the highly resistant parent H7996 as well as location of polymorphisms within genes putatively associated with tolerance to stress (Supplemental Table 1). While the molecular markers developed here were polymorphic for the parental lines (data not shown), most markers were unable to accurately predict BW resistance phenotypes in the segregating F_2 populations. Marker Bwr3.2dCAPS located on chromosome 3 was significantly associated with the phenotypic response in the CLN4398 population (Table 5). A minor QTL on chromosome 3 was previously found to contribute to resistance derived from H7996[8,9,28]. The reported size of *Bwr-3* is quite large, spanning most of the distal end of chromosome 3[9,28]and Bwr3.2dCAPS is within this region, supporting our results. Furthermore, marker Bwr3.2dCAPS was located within the *Asc* gene (Solyc03g114600.4.1) which confers resistance to *Alternaria alternata* f. sp. *lycopersici* (AAL). The Bwr3.2dCAPS marker is based on the deletion of the 102nd arginine in the *Asc* gene, resulting in a high-impact frameshift mutation that affects transcription and translation. The *Asc* locus was first identified by Gilchrist and Grogan[36] and two alleles were found with resistance to the pathogen being dominant although the heterozygous condition conferred intermediate resistant phenotypes in AAL-toxin sensitivity assays. The *Asc* locus was later mapped to chromosome 3[37–39] and was found to mediate resistance to sphinganine-analog mycotoxins (SAM)-induced apoptosis[40]. Interestingly, the homologous *LAG1-like Asc1* gene has been found to rescue tomato hair roots from SAM-induced cell death[41] and the *Asc* gene has been found to be upregulated when plants were infested with *Bactericera cockerelli* infectious with *Candidatus Liberibacter solanacearum*[42], potentially indicating *Asc* has multiple functions including response to bacterial infection and could be contributing to resistance to *Ralstonia* sp.

Conclusion

In this study, we utilized whole genome sequence data analysis, based on pairwise comparison of BW resistant and susceptible lines to identify candidate genes contributing to resistance above the levels conferred by *Bwr-6* and *Bwr-12*. Through this approach we found 27,046 SNPs and 5,975 indels specific to the resistant lines and

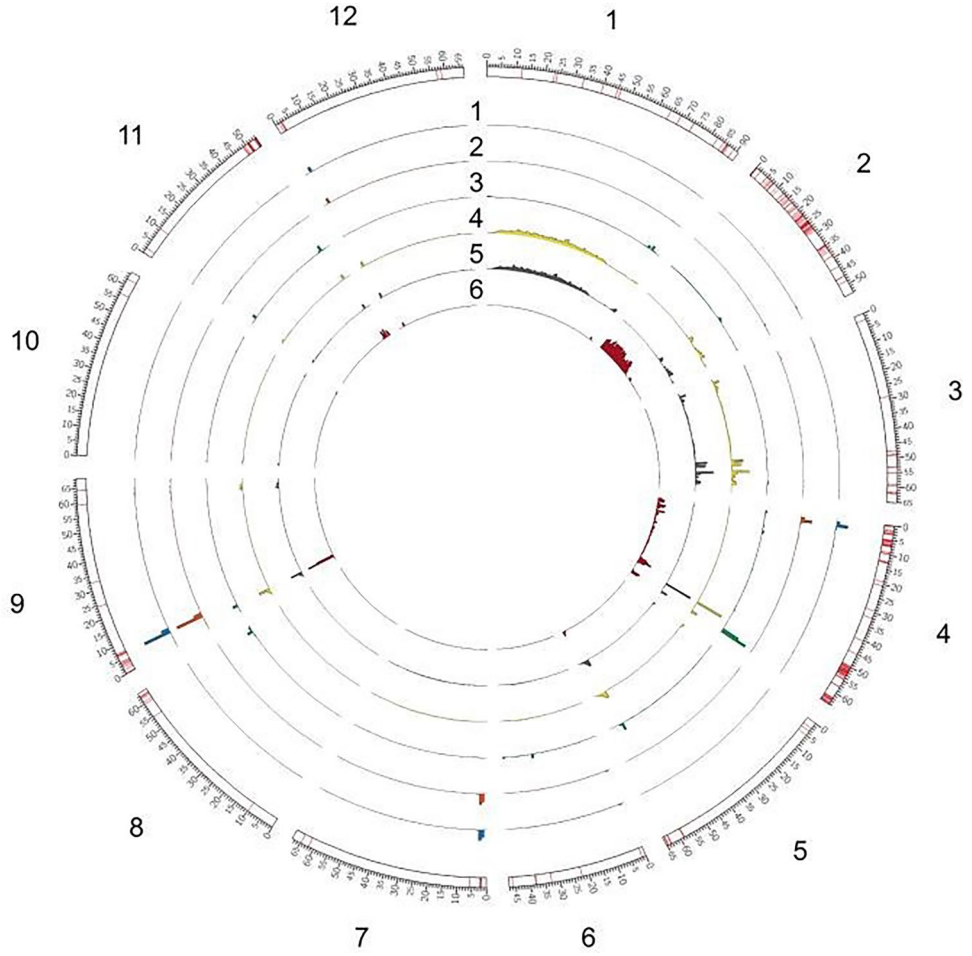


Figure 4. The genome-wide distribution of filtered variants and highly-affected genes of six bacterial wilt (*Ralstonia* sp.) resistant tomato lines. The 12 chromosomes are numbered clockwise, and the red bands on the outermost bars are genes highly affected by polymorphisms of 6 resistant accessions. The six histograms display the number of SNPs in 1-mb windows of 6 resistant tomato accessions. The lines are numbers (1) Hawaii 7996, (2) Hawaii 7997, (3) LE415, (4) F7_80P, (5) F7_80465P, and (6) LS89.

causing high impact mutations in 385 genes. Furthermore, in addition to *Bwr-6* and *Bwr-12*, we found one or two independent loci contributed BW resistance based on inheritance patterns. Association between the phenotype and a newly developed molecular marker, Bwr3.2dCAPS in the previously reported *Asc* gene, was statistically significant but it did not completely explain the resistance phenotype. This study provides a basis for further investigations into new loci distributed throughout the genome that could contribute to BW resistance in tomato.

Materials and methods

Plant materials and inoculation. To identify highly resistant and susceptible individual plants for sequencing, six resistant tomato lines were selected (LS-89, H7997, F_7-80-465-10-pink, F7-80-pink, H7996, and LE415 Anagha) and nine susceptible lines (CRA84-23-1 115, CRA84-57-1 140, T-245, S/T2, 'Rodade', LA3501, TBL-2, L390, and 'Pant Bahar'), previously reported by Kunwar et al.[12]. The lines were inoculated with two virulent strains of *Ralstonia* sp., Pss4 (race 1, biovar 3[17], *R. pseudosolanacearum*) and Pss1632 (race 3, biovar 2, *R. solanacearum*), representing the former designations of Phylotype I and Phylotype II, respectively. The bioassay was conducted during the hot season (June–July) of 2018 in a controlled environment greenhouse (19 ± 4 °C night and 39 ± 4 °C day) in Shanhua, Tainan, Taiwan (lat. 23.1°N; long. 120.3°E; elevation 12 m) and plants were fertilized weekly. The experiment followed a completely randomized design (CRD) with two replications, each with 20 plants for each of the strains used. The plants were inoculated at the 4–6 true leaf stage by drenching with a bacterial suspension (10^8 CFU/ml) on the soil surface at a ratio of 1:10 (v:v) inoculum to potting mix. The individual plants were scored using a standardized scale twice a week for two weeks. The resistance percentage was calculated based on the number of asymptomatic plants during each time point. The highly resistant lines had a higher percent resistance after two weeks, while the highly susceptible lines had a low percent resistance within the first week after inoculation.

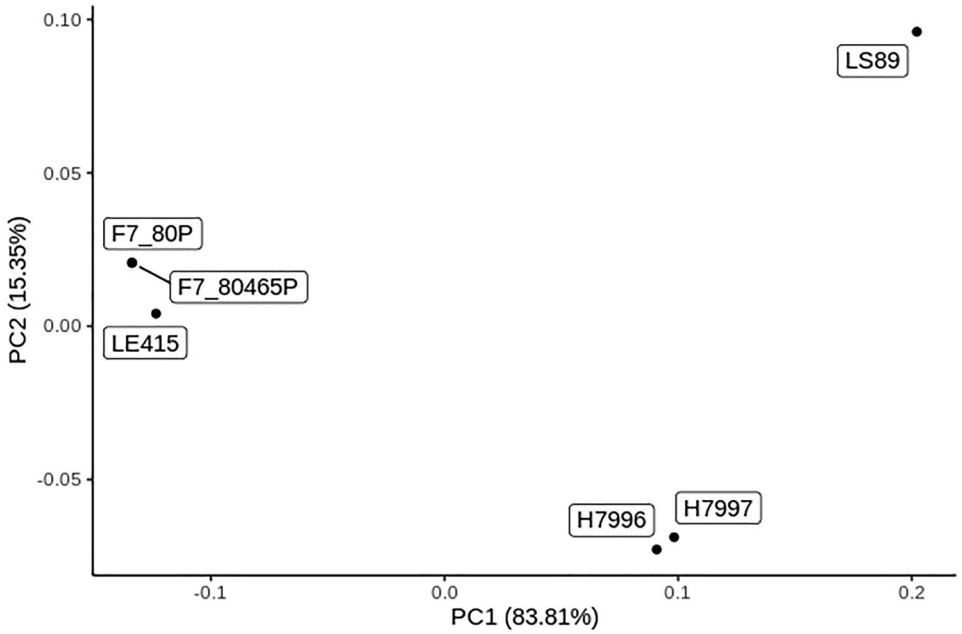


Figure 5. The Principal Coordinate Analysis based on the polymorphisms of the six bacterial wilt (*Ralstonia* sp.) resistant tomato lines used in this study. F7_80P and F7_80465P share the same PC1 and PC2. H7996 is Hawaii 7996 and H7997 is Hawaii 7997.

Population	Expected ratio	AUDPC ≤ 35 (resistant)	AUDPC > 35 (susceptible)	χ^2-value	*P* value
CLN4018F1-6-7U14-29–21-14–5	1:0	30	0	–	–
Hawaii 7996	1:0	30	0	–	–
CLN4398-8	3:1	107	93	49.3	< 0.001
	9:7			0.6	0.4331
CLN3641F1-5-11-14-4-25-20-11-7(F)	1:0	30	0	–	–
Hawaii 7996'	1:0	30	0	–	–
CLN4397-4	3:1	117	43	0.3	0.5839
	9:7			18.5	< 0.001

Table 3. Goodness of fit test for inheritance of resistance to the Pss4 isolate of bacterial wilt (race 1, biovar 3, *Ralstonia pseudosolanacearum*) for the two F_2 populations (CLN4398-8 and CLN4397-4) derived from CLN4018F1-6-7U14-29-21-14-5 by 'Hawaii 7996' and CLN3641F1-5-11-14-4-25-20-11-7(F) by 'Hawaii 7996', respectively.

DNA isolation, library preparation, and sequencing. For whole genome resequencing, five individual plants within each of the six resistant and nine susceptible lines were selected. Selection of plants was based on extremes in phenotype with susceptible individual plants selected based on early symptom occurrence, while resistant plants were selected by absence of symptoms at the final evaluation. DNA was extracted from each of the five plants using the Qiagen DNeasy kit following the manufacturer's instructions (Qiagen; Hilden, Germany), quantified using a fluorometer (Qubit 2.0, Invitrogen, Waltham, MA, USA) and pooled in equal amounts for each accession. The total DNA concentration, and DNA quality were determined using the TapeStation system (Agilent, Santa Clara, CA, USA). DNA libraries were generated using the NEBNext Ultra II DNA Library Prep Kit for Illumina (New England Biolabs, Ipswich MA, USA) according to the manufacturer's instructions. The quality of the libraries was assessed using the TapeStation system with D1000 High Sensitivity ScreenTape. Next-generation sequencing using the HiSeq Illumina platform with 150 bp paired-end reads was conducted by Welgene Biotech Co., Ltd. (Taipei, Taiwan). Total DNA was isolated from leaf tissue collected prior to inoculation and stored at −80 °C until the phenotyping experiment was completed.

Sequence analysis. For the whole genome sequencing analysis, the quality of reads was checked using FastQC (v. 0.11.7)[43]. All reads were trimmed based on an average Phred quality score of 20 for 4 consecutive bases and we discarded reads shorter than 50 bp using Trimmomatic (v.0.36)[44]. We then mapped the reads to the annotated 'Heinz 1706' reference genome (v.SL4.0)[45] using the "mem" algorithm of Burrows-Wheeler Align-

Marker	Chromosome	Position (Mp)	Restriction enzyme	Primer (5′→3′)	Product size (R/S) (bp)
Bwr1.1indel	1	8.2	–	CAGGTAAGATGGAGAACATG	81/173
				TGTTCAATGTGCTGTTCGTG	
Bwr1.2HRM	1	8.5	–	GAGATTTCCTCAAGGTTT TCCTC	127
				AGCTTGTTTATCTCTCTCTC	
Bwr3.1HRM	3	0.6	–	CCACAGACAGATTTCTCGGT	126
				GTAGTGTCCAAGTAAGGT ATAG	
Bwr3.2dCAPS	3	5.8	*BsrBI*	TTTGAATTTGTTGATCTTCTT CTCgCT	129/(105 + 24)
				ATTGATTTGGACGCGTGCTT	
Bwr4.1indel	4	2.0	–	GAGTGCGAGGAATGTATACT	(14 + 7 + 142)/(14 + 149)
				TCCAGTTTGTCTCATTTT CATCC	
Bwr4.2indel	4	2.0	–	CCAAGGTTTCGTGTATTTTAC	180/170
				TAATTGCAGCTTCCAAAT GGAC	
Bwr4.3CAPS	4	2.0	*DdeI*	CTTGAGTTTCATATTTGCTAA	(18 + 46 + 105)/(64 + 105)
				GTGTCAACATTCTTATTGTA	
Bwr4.4HRM	4	2.7	–	TGAACCCTACATTCAGTAACT TTTTCCCAACA	150
				ATGGTTGTGGATGGCGGAG	
Bwr4.5HRM	4	59.0	–	TGCAGCAATACCTTTGGA TAGGA	141
				CGCCACGCAATTTGAGACAG	
Bwr5.1HRM	5	2.2	–	TTCGCGTTTGAAGAAGAGGT	158
				TCGATTTTCGAACAAGCCTA	
Bwr7.1HRM	7	1.7	–	GAGATTTCCTCAAGGTTT TCCTA	159
				TCCCTTATCACTTAGGCCACA	
Bwr7.2HRM	7	1.89	–	TGCAACTTCCTTCCATTT TCCT	127
				TGCCCACAAATTCCATTCCA	
Bwr8.1CAPS	8	59.8	*NruI*	AGTCACACCAGATTGCAGGA	163/(132 + 31)
				GGGGATTTTCGAACGTTT AATGC	
Bwr9.1indel	9	0.3	–	CCAGCAAACCAAGTCGAT	220/161
				ATGGTCTTGTACTCAACTC	
Bwr9.2HRM	9	64.6	–	GATGTATGACAAGTCCAGTG	260
				GTGAGGCAAAGAACATAC TTCCA	

Table 4. Position (Mp), primer sequence, restriction enzyme required, and product size for the molecular markers developed and evaluated in this study for validation in the F_2 populations.

Population	AUDPC	R/H	S	*P* value
CLN4398-8	0–35	90	18	0.0178
	36–105	64	28	

Table 5. Association between the phenotypic response when inoculated with the Pss4 isolate of bacterial wilt (race 1, biovar 3, *Ralstonia pseudosolanacearum*) of the CLN4398 F_2 population and the Bwr3.2dCAPs molecular marker determined by Fisher's Exact Test in R.

ment (BWA-MEM; v0.7.17)[46] and the average number of reads was 1.15×10^8. Minimum coverage depth was set to 25×, but most of the time mean read depth was ~ 50×.

Variant calling. Variant calling was performed using Genome Analysis Toolkit (GATK; v4.1.6.0)[47] the Picard Toolkit (v2.21.9)[48] and samtools (v1.10)[49]. First, PCR duplicates were removed using MarkDuplicates for each sample and then HaplotypeCaller, GenotypeGVCFs, and VariantFiltration sequentially were used for vari-

ant calling, the filtration of variants to get the first version of homozygous SNP, and indels. For the filters in VariantFiltration, there were six filters for SNPs and three for indels. For SNPs, SNPs with FisherStrand (FS) equal to or less than 60, StrandOddsRatio (SOR) equal to or less than 3, RMSMappingQuality (MQ) equal to or greater than 40, MappingQualityRankSumTest (MQRankSum) equal to or greater than -12.5 and ReadPosRankSum (ReadPosRankSum) equal to or greater than -8.0 were retained. For indels, variants with FS equal to or less than 200, ReadPosRankSum equal to or greater than -20. We used the threshold QualByDepth (QD) as equal to or greater than 2 for both SNPs and indels were kept. The first version of homozygous variants was used to recalibrate the bam files of each sample using BaseRecalibrator and BQSR, then variant calling was again performed based on recalibrated bam files to get the final version of homozygous SNPs and indels written in the Variant Call Format (VCF) files. SNPs with read depth > 10, no missing data, and no heterozygous sites were retained, resulting in about 1.8 million SNPs. These SNPs were then used to calculate the. Principal Coordinate Analysis (PCA) of the genetic distance with TASSEL 5.0 and in R-3.6.3[50].

A customized script in R-3.6.3 was developed to compare the variants of six resistant lines with nine susceptible lines. To comprehensively screen the candidate markers that contributed to the resistance, each resistant line was compared individually with all susceptible lines and only variants polymorphic between the individual resistant lines and all susceptible lines were retained. Then, the variant annotation and effect prediction based on these variants only from six resistant lines was performed using SnpEff 4.3t[51]. The distribution of variants and highly affected genes were visualized by Circos (v 0.69–8)[52].

Molecular marker development. Based on the polymorphisms specific to resistant lines with high impact differences in predicted effects, nine loci predicted to encode proteins with putative functions associated with resistance to bacterial wilt were selected. In each selected locus, molecular markers were designed to test for associations between the sequence polymorphism in candidate genes and the resistant phenotype, which could not be explained by *Bwr-6* and *Bwr-12* QTLs. A total of 15 molecular markers were designed for validation, eight high resolution melting (HRM) markers, four insertion-deletion (indel) markers, two cleaved amplified polymorphic sequence (CAPS) markers, and one derived cleaved amplified polymorphic sequence (dCAPS) marker. All molecular markers were first used to genotype the parental lines and only those that were confirmed to be polymorphic were selected to genotype the validation populations. For the gel-based molecular markers, the PCR reactions included 2 µL DNA, 2 µL 10 × PCR buffer with 1.5 mM MgCl2 (10 × GOLD Buffer), 0.15 mM dNTPs, 0.25 U Taq polymerase (Gold Taq 250 U) and 0.5 mM for forward and reverse primers. The PCR temperature profile was as follows: 95 °C for 10 min, 35 cycles for 95 °C for 30 s., 55 °C for 45 s. and 72 °C for 45 s., followed by 72 °C for 5 min and final hold at 15 °C. The PCR product were separated on 6% polyacrylamide gels alongside a 50-bp DNA ladder in TBE buffer (90 mM Tris, 90 mM Boric acid, 2 mM EDTA, pH 8.4, VWR) at 160 V and 400 mA for 30–55 min. The polyacrylamide gels were stained by DNA fluorescent dye (FluoroStainTM DNA Fluorescent Staining Dye; Green, 5,000X, SMOBIO) for 10 min. The stained polyacrylamide gels were visualized using a blue-light imaging system (BIO-1000F). For the HRM molecular markers, the reactions were performed using a total volume of 20 µL containing 20 ng of PCR fragment on a Corbett Rotor Gene 6000. The reaction used the SensiFAST™ HRM Kit and followed the manufacturer's instructions. For PCR, 5 min pre- denaturation at 95 °C was followed by 50 cycles of 95 °C for 10 s, 60 °C for 30 s, and 72 °C for 35 s. For the HRM analysis, the amplicons spanned from 65 to 95 °C, rising by 0.1 °C each step. The Rotor-Gene Q software version v2.2 was used to analyze the melting curve results.

Validation. For marker validation, two F_2 populations coded CLN4397-4 (CLN3641F1-5-11-14-4-25-20-11–7(F) × H7996 [160 individuals]) and CLN4398-8 (CLN4018F1-6-7U14-29-21-14-5 × H7996 [200 individuals]) were developed, all of which were homozygous for both the *Bwr-6* and *Bwr-12* QTLs except for a few heterozygotes in the CLN4398 population. All lines, including one susceptible check (L390) and parental lines, were grown in the greenhouse as previously mentioned and fertilized weekly. At the 4–6 true leaf stage, the F_2 populations were screened with the Pss4 strain by drench inoculation as described above. Plants were scored using a standardized 0 to 5 rating scale twice weekly for two weeks after inoculation. The scores were used to calculate the area under the disease progress curve (AUDPC) and the deviation from expected segregation ratios of resistance in the two F_2 populations was determined using the χ^2 test in R-3.6.3[50].

Sequencing data were submitted to the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA).

Ethical statement. Experimental research and field studies on plants (either cultivated or wild), including the collection of plant material, complies with relevant institutional, national, and international guidelines and legislation.

Data availability

The Illumina sequencing data have been deposited at NCBI under BioProject PRJNA725647. (reviewer linkhttps://dataview.ncbi.nlm.nih.gov/object/PRJNA725647?reviewer=d15n1ajijjhsspov22ta9s50fa) All other data are available at the World Vegetable Center repository, HARVEST (worldveg.org/harvest3).

Received: 24 May 2021; Accepted: 7 April 2022
Published online: 19 May 2022

References

1. Mansfield, J. *et al.* Top 10 plant pathogenic bacteria in molecular plant pathology. *Mol. Plant Pathol.* **13**, 614–629. https://doi.org/10.1111/j.1364-3703.2012.00804.x (2012).
2. Imada, K., Sakai, S., Kajihara, H., Tanaka, S. & Ito, S. Magnesium oxide nanoparticles induce systemic resistance in tomato against bacterial wilt disease. *Plant Pathol.* **65**, 551–556. https://doi.org/10.1111/ppa.12443 (2016).
3. Hayward, A. C. Biology and epidemiology of bacterial wilt caused by *Pseudomonas solanacearum*. *Ann. Rev. Phytopathol.* **29**, 65–87. https://doi.org/10.1146/annurev.py.29.090191.000433 (1991).
4. Hanson, P. *et al.* Conventional and molecular marker-assisted selection for pyramiding of genes for multiple disease resistance in tomato. *Sci. Hortic.* **201**, 346–354. https://doi.org/10.1016/j.scienta.2016.02.020 (2016).
5. Scott, J. W., Wang, J. F. & Hanson, P. M. Breeding tomato for resistance to bacterial wilt, a global view. *Acta Hortic.* **695**, 161–172 (2005).
6. Hanson, P. M. *et al.* Variable reaction of tomato lines to bacterial wilt evaluated at several locations in Southeast Asia. *HortScience* **31**, 143–146. https://doi.org/10.21273/HORTSCI.31.1.143 (1996).
7. Wang J.F., Hanson, P.M. & Barnes, J.A. *Worldwide evaluation of an international set of resistance sources to bacterial wilt in tomato* in Bacterial wilt disease: Molecular and ecological aspects (eds. Prior, P., Allen, C. & Elphinstone, J.) 269–275 (Springer, Berlin, 1998).
8. Wang, J. F. *et al.* Identification of major QTLs associated with stable resistance of tomato cultivar 'Hawaii 7996' to *Ralstonia solanacearum*. *Euphytica* **190**, 241–252. https://doi.org/10.1007/s10681-012-0830-x (2013).
9. Carmeille, A. *et al.* Identification of QTLs for *Ralstonia solanacearum* race 3-phylotype II resistance in tomato. *Theor. Appl. Genetic.* **113**, 110–121. https://doi.org/10.1007/s00122-006-0277-3 (2006).
10. Geethanjali, S., Chen, K. Y., Pastrana, D. V. & Wang, J. F. Development and characterization of tomato SSR markers from genomic sequences of anchored BAC clones on chromosome 6. *Euphytica* **173**, 85–97. https://doi.org/10.1007/s10681-010-0125-z (2010).
11. Geethanjali, S., Kadirvel, R., de la Pena, R., Rao, E. S. & Wang, J. F. Development of tomato SSR markers from anchored BAC clones of chromosome 12 and their application for genetic diversity analysis and linkage mapping. *Euphytica* **178**, 283–295. https://doi.org/10.1007/s10681-010-0331-8 (2011).
12. Kunwar, S. *et al.* Characterization of tomato (*Solanum lycopersicum*) accessions for resistance to phylotype I and phylotype II strains of the *Ralstonia solanacearum* species complex under high temperatures. *Plant Breed.* **139**, 389–401. https://doi.org/10.1111/pbr.12767 (2020).
13. Cheng, C.P. Utilizing and developing molecular markers to improve breeding efficiency for tomato bacterial wilt disease resistance-Evaluating the *Bwr12* QTL on chromosome 12 for resistance to Phylotype I of bacterial wilt. Project Report NSC 101–2324-B-002–019 (2014). (in Mandarin)
14. Kim, B. *et al.* Identification of a molecular marker tightly linked to bacterial wilt resistance in tomato by genome-wide SNP analysis. *Theor. Appl. Genetic.* **131**, 1017–1030. https://doi.org/10.1007/s00122-018-3054-1 (2018).
15. Shin, I. S. *et al.* Construction of a single nucleotide polymorphism marker based QTL map and validation of resistance loci to bacterial wilt caused by *Ralstonia solanacearum* species complex in tomato. *Euphytica* **216**, 54. https://doi.org/10.1007/s10681-020-2576-1 (2020).
16. Zohoungbogbo, H., Quenum, A., Honfoga, J., Chen, J.R., Achigan-Dako, E., Kenyon, L., *et al.* Evaluation of resistance sources of tomato (*Solanum lycopersicum* L.) to phylotype I strains of *Ralstonia solanacearum* species complex in Benin. (2021). (Submitted).
17. Hai, T. T. H., Esch, E. & Wang, J. F. Resistance to Taiwanese race 1 strains of *Ralstonia solanacearum* in wild tomato germplasm. *Eur. J. Plant Pathol.* **122**, 471–479. https://doi.org/10.1007/s10658-008-9314-1 (2008).
18. Albuquerque, G. M. R. *et al.* Stability analysis of reference genes for RT-qPCR assays involving compatible and incompatible Ralstonia solanacearum-tomato 'Hawaii 7996' interactions. *Sci. Rep.* **11**, 18719 (2021).
19. Aflitos, S. *et al.* Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. *Plant J.* **80**, 136–148. https://doi.org/10.1111/tpj.12616 (2014).
20. Tranchida-Lombardo, V. *et al.* Whole-genome re-sequencing of two Italian tomato landraces reveals sequence variations in genes associated with stress tolerance, fruit quality and long shelf-life traits. *DNA Res.* **25**, 149–160. https://doi.org/10.1093/dnares/dsx045 (2018).
21. Wang, Y. *et al.* Application of whole genome resequencing in mapping of a *Tomato yellow leaf curl virus* resistance gene. *Sci. Rep.* **8**, 9592. https://doi.org/10.1038/s41598-018-27925-w (2018).
22. Nakaho, K. *et al.* Involvement of a vascular hypersensitive response in quantitative resistance to *Ralstonia solanacearum* on tomato rootstock cultivar LS-89. *Plant Pathol.* **66**, 150–158. https://doi.org/10.1111/ppa.12547 (2017).
23. Ishihara, T., Mitsuhara, I., Takahashi, H. & Nahako, K. Transcriptome analysis of quantitative resistance-specific response upon Ralstonia solanacearum infection in tomato. *PLoS* **7**, e46763. https://doi.org/10.1371/journal.pone.0046763 (2012).
24. Hanson, P., Licardo, O., Hanudin Wang, J. F. & Chen, J. T. Diallel analysis of bacterial wilt resistance in tomato derived from different sources. *Plant Dis.* **82**, 74–78 (1998).
25. Daunay, M. C., Laterrot, H., Scott, J. W., Hanson, P. & Wang, J.-F. Tomato resistance to bacterial wilt caused by *Ralstonia solanaearum* E.F. Smith: Ancestry and peculiarities. *Rep. Tomato Genet. Coop.* **60**, 6–40 (2010).
26. Yamakawa, K. Cultivar resistance to *Pseudomonas solanacearum* in tomato and eggplant. *Shokubutsu Boeki (Plant Protection)* **32**, 197–200 (1978) (**(in Japanese)**).
27. Acosta, J. C., Gilbert, J. C. & Quinon, V. L. Heritability of bacterial wilt resistance in tomato. *Proc. Amer. Soc. Hort. Sci.* **84**, 455–462 (1964).
28. Hai, T.T.H. Characterisation and mapping of bacterial wilt (*Ralstonia solanacearum*) resistance in the tomato (*Solanum lycopersicum*) cultivar Hawaii 7996 and wild tomato germplasm. MSc Thesis, Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany. (2007).
29. Mohamed, M. E. S., Umaharan, P. & Phelps, P. H. Genetic nature of bacterial wilt resistance in tomato (Lycopersicon esculentum Mill.) accession LA 1421. *Euphytica* **96**, 323–326 (1997).
30. Prior, P., Grimault, V. & Schmit, J. *Resistance to bacterial wilt (Pseudomonas solanacearum) in tomato: Present status and prospects* in Bacterial Wilt: The disease and its causative agent *Pseudomonas solanacearum* (eds. Hayward A. C. & Hartman G. L.) 209–223 (Centre for Agriculture and Bioscience International, Wallingford, UK, 1994).
31. Scott, J. W., Somodi, G.C. & Jones, J.B. *Testing tomato genotypes and breeding for resistance to bacterial wilt in Florida* in Bacterial Wilt. (eds. Hartman G. L. & Hayward A. C.) 126–131 (Australian Centre for International Agricultural Research, Canberra. 1993).
32. Thoquet, P. *et al.* Quantitative trait loci determining resistance to bacterial wilt in the tomato cultivar Hawaii 7996. *Mol. Plant-Microbe Interact.* **9**, 826–836 (1996).
33. Thoquet, P. *et al.* Polygenic resistance of tomato plants to bacterial wilt in the French West Indies. *Mol. Plant Microbe Interact.* **9**, 837–842 (1996).
34. Grimault, V., Prior, P. & Anaïs, G. A monogenic dominant resistance of tomato to bacterial wilt in Hawaii 7996 is associated with plant colonization by *Pseudomonas solanacearum*. *J. Phytopathol.* **143**, 349–352 (1995).
35. Scott, J. W., Somodi, G. C. & Jones, J. B. Bacterial spot resistance is not associated with bacterial wilt resistance in tomato. *Proc. Fla. State Hortic. Soc.* **101**, 390–392 (1988).

36. Gilchrist, D. G. & Grogan, R. G. Production and nature of a host-specific toxin from Alternaria alternata f.sp. lycopersici. *Phytopathology* **66**, 165–171 (1976).
37. Witsenboer, H. M. A., van de Griend, E. G., Tiersma, J. B., Nijkamp, H. J. J. & Hille, J. Tomato resistance to Alternaria stem canker: localization in host genotypes and functional expressions compared to non-host resistance. *Thero. Appl. Genetic.* **78**, 457–462 (1989).
38. Tanksley, S. D. *et al.* High density maps of the tomato and potato genomes. *Genetics* **132**, 1141–1160 (1992).
39. van der Biezen, E. A. *et al.* Molecular genetic characterisation of the Asc locus of tomato conferring resistance to the fungal pathogen Alternaria alternata f. sp. lycopersici. *Euphytica* **79**, 205–217 (1994).
40. Wang, H., Li, J., Bostock, R. M. & Gilchrist, D. G. Apoptosis: A functional paradigm for programmed plant cell death induced by a host-selective phototoxic and involved during development. *Plant Cell* **8**, 375–391 (1996).
41. Brandwagt, B. F. *et al.* A longevity assurance gene homolog of tomato mediates resistance to Alternaria alternata f. sp. lycopersici toxins and fumonisin B1. *Proc. Natl. Acad. Sci. USA* **97**, 4961–4966 (2000).
42. Huot, O.B. Molecular and biological mechanisms of host plant responses to an insect vector and a bacterial pathogen. PhD Dissertation, Texas A&M University, College Station, TX. (2017).
43. Andrews, S. FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (2010).
44. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: A flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
45. Hosmani, P. S. *et al.* An improved de novo assembly and annotation of the tomato reference genome using single-molecule sequencing, Hi-C proximity ligation and optical maps. *biorxiv* https://doi.org/10.1101/767764 (2019).
46. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *ArXiv*, 1303 (2013). https://arxiv.org/abs/1303.3997
47. Van der Auwera G.A. & O'Connor B.D. Genomics in the cloud: Using Docker, GATK, and WDL in Terra (1st Edition). (O'Reilly Media, 2020).
48. Picard Toolkit. Broad Institute, GitHub Repository (2019). http://broadinstitute.github.io/picard/
49. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
50. R core team. R: A language and environment for statistical computing. (2020). https://www.R-project.org/.
51. Wei, F. J. *et al.* Somaclonal variation does not preclude the use of rice transformants for genetic screening. *Plant J.* **85**, 648–659. https://doi.org/10.1111/tpj.13132 (2016).
52. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

Acknowledgements

The authors thank Dr. Po-Xing Zheng for his assistance in bioinformatics analysis, our interns Ms. Szu-ping Chen under the supervision of Dr. Kai-yi Chen at National Taiwan University and Ms. Chia-chi Yeh under the supervision of Dr. Hung-ling Yeh at National Chiayi University for their technical assistance in this project.

Author contributions

D.W.B., Y.C.L., M.A.O.B., P.H. and R.S. conceived and coordinated the study and acquired funding. Y.M.H, J.Y.O., and Y.P.L. conducted sequence analysis. Y.C.H. conducting the phenotyping and genotyping experiments. All authors read and reviewed the manuscript.

Funding

Funding for this research was provided by the Ministry of Science and Technology (MOST) of Taiwan (Project ID: 108-2923-B-125-001-MY3 and 107-2313-B-125 -001) and the Department of Science and Technology-Philippine Council for Agriculture, Aquatic and Natural Resources Research and Development (DOST-PCAARRD) (Project ID:N9-A96-21) as part of the Manila Economic and Cultural Office (MECO)-Taipei Economic and Cultural Office (TECO) Joint Research Initiative as well as long-term strategic donors to the World Vegetable Center, Taiwan; UK aid from the UK government; U.S. Agency for International Development (USAID); Australian Centre for International Agricultural Research (ACIAR), Germany, Thailand, Philippines, Korea, and Japan.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-022-12326-x.

Correspondence and requests for materials should be addressed to D.W.B.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2022

4. Conclusion and perspectives

The research of this dissertation was built around three projects.

4.1 Machine-learning based chromatin states with the assistance of genomic features can predict fine-scale meiotic recombination variations

Meiotic recombination is a complicated biological phenomenon produced from meiosis, and it is influenced by different factors. While comparing recombination rate with one feature, we discovered that the relationship between them is mostly nonmonotonic. Furthermore, when two features are highly correlated (e.g. gene density and the intensity of H3K4me3), they show similar correlation patterns with recombination rate. This indicates that it is not biologically interpretable to naively use genomic and epigenomic features for establishing a quantitative model that can reproduce crossover landscapes. Based on the linear models using these features, not only do we get a model with weak predictive power but we also lack the ability to investigate the underlying relationship between each feature and crossover rate.

Based on 9 chromatin states identified from machine learning techniques and dependent on 16 genomic and epigenomic features, we added a 10th state associated with structural variation between Col and Ler, parents of the F_2 population for identifying CO intervals. These 10 states, obtained from a discrete classifier algorithm, allowed us to predict the recombination rate landscape along the chromosome, and in particular averaged fine-scale recombination rates in genes and intergenic regions. In addition, we found that sequence divergence, SNP density, and intergenic-region size also influence CO rate. These two genomic features indeed improved our model's prediction accuracy of crossover landscapes. Even though this model can reproduce much of the variation of experimental recombination, there are still variations that can't be predicted by this model, suggesting that this model has some limitations. First, the 9 chromatin states were built using somatic cells instead of meiotic cells. It may therefore cause a biased result, even

though we showed that the correlation between CO rate and one feature is similar when using somatic and meiotic tissues. In addition, all of these 9 chromatin states depended on Col, and we assumed that all syntenic regions between Col and Ler have the same chromatin status. Our model could benefit from having the complete data of genomic and epigenomic features from both parents for identifying the different chromatin state profiles between parents. Furthermore, the information of loop structures in meiotic chromosomes is missing, and this variation, in terms of sizes and locations, is possibly useful to improve the prediction of fine-scale CO rate. Finally, since the CO dataset we studied comes from a F_2 population, it only reflects the average of male and female meiosis rates. At present, our model has no capability to predict the difference between male and female recombination rates.

All together, our model provides useful information for predicting crossover landscapes, and it can be extended to other crosses and species. With the development of new technologies and data in the near future, this model can be improved for better explaining variations in meiotic recombination rates.

4.2 Sequencing reveals the genetic diversity of 31 peanut accessions in Taiwan and identifies a candidate gene of bacterial wilt resistance for future plant breeding

In Taiwan, the molecular breeding of peanuts has lagged behind that of other countries. In this second research project, RAD-seq was used to sequence 31 Taiwanese accessions including elite cultivars, landraces and lines introduced from different countries. The result showed that SNPs can cluster these 31 accessions into groups according to their botanical varieties. In addition, the diversity analysis indicates that accessions introduced from the geographical origin of the cultivated peanut are more diverse than other accessions. For example, the global subset and var. *fastigiata*, acquiring more accessions from South America, have higher genetic diversity than the local subset and two other botanical varieties, respectively. Furthermore, the analysis of the genetic relationship between 31 accessions suggested that one should introduce more diversity into current peanut breeding programs in Taiwan since Taiwanese elite cultivars are highly genetically related. To identify the population structure of this peanut germplasm collection, a set of KASP markers were developed, and reliably distinguished 282 peanut accessions into their corresponding botanical varieties. Besides, these markers can even identify accessions with similar genetic background but from different botanical varieties, offering an alternative and efficient tool for understanding the genetic relationships between accessions without being confounded by phenotypic data. One can rely on the sequencing data of 31 accessions to develop more non-gel based markers to accelerate and improve the breeding process, such as when performing background and foreground selection.

In our third research project we considered tomato bacterial wilt (BW) which is a destructive disease. To date, only two major QTLs have been identified. In order to identify minor QTLs, we sequenced 6 resistant and 9 susceptible lines by WGS, and performed pairwise comparison between each of resistant

lines and all susceptible lines for keeping variants uniquely found in resistant lines for the further gene function prediction. Finally, we identified 385 candidate genes highly influenced by 27,046 SNPs and 5,975 indels specifically identified in the resistant lines. Only Bwr3.2dCAPS, in the previously published *Asc* gene, was statistically significantly associated with phenotypes of a F_2 population. In this study, we thus demonstrated that pairwise comparison is useful for identifying minor QTLs. Lastly, the two F_2 populations used for validating candidate genes were developed from H7996 and an advanced breeding line. In the future, one should develop more populations depending on the other five resistant lines to validate more candidate genes which could contribute to BW resistance.

4.3 Tools developed for different aspects of genetic diversity that could facilitate plant breeding in the future

In summary, in the first project we built a quantitative model which can predict crossover landscapes in *Arabidopsis thaliana*. This work could be an effective tool for predicting crossover landscapes in crops if more techniques and data related to epigenomics in crops are generated. In the second project, by revealing the genetic diversity of the cultivated peanut, we provided useful information for future peanut breeding in Taiwan, and the sequence data in this project can be a basis for peanut molecular breeding. Finally, the sequence and BW candidate gene data of tomatoes can enable tomato breeders and geneticists to better design tomato varieties with much more durable resistance against BW. Taken together, the three projects in my thesis work can help improve future plant breeding by new ways to exploit genetic diversity.

5. Reference

Abebe, A. M., Choi, J., Kim, Y., Oh, C.-S., Yeam, I., Nou, I.-S., & Lee, J. M. (2020). Development of diagnostic molecular markers for marker-assisted breeding against bacterial wilt in tomato. *Breeding Science*, *70*(4), 462–473. https://doi.org/10.1270/jsbbs.20027

Albini, S. M., & Jones, G. H. (1987). Synaptonemal complex spreading in Allium cepa and A. fistulosum. *Chromosoma*, *95*(5), 324–338. https://doi.org/10.1007/BF00293179

Alix, K., Gérard, P. R., Schwarzacher, T., & Heslop-Harrison, J. S. (Pat). (2017). Polyploidy and interspecific hybridization: Partners for adaptation, speciation and evolution in plants. *Annals of Botany*, *120*(2), 183–194. https://doi.org/10.1093/aob/mcx079

An, X. J., Deng, Z. Y., & Wang, T. (2011). OsSpo11-4, a Rice Homologue of the Archaeal TopVIA Protein, Mediates Double-Strand DNA Cleavage and Interacts with OsTopVIB. *PLOS ONE*, *6*(5), e20327. https://doi.org/10.1371/journal.pone.0020327

Barkley, N. A., Upadhyaya, H. D., Liao, B. & Holbrook C. C. Global resources of genetic diversity in peanut (eds. Stalker, H. T. & Wilson, R. F.) 67–109 (AOCS Press, 2016).

Batley, J., Barker, G., O'Sullivan, H., Edwards, K. J., & Edwards, D. (2003). Mining for Single Nucleotide Polymorphisms and Insertions/Deletions in Maize Expressed Sequence Tag Data. *Plant Physiology*, *132*(1), 84–91. https://doi.org/10.1104/pp.102.019422

Bechara, M. D., Moretzsohn, M. C., Palmieri, D. A., Monteiro, J. P., Bacci, M., Martins, J., Valls, J. F., Lopes, C. R., & Gimenes, M. A. (2010). Phylogenetic relationships in genus Arachis based on ITS and 5.8S rDNA sequences. *BMC Plant Biology*, *10*(1), 255. https://doi.org/10.1186/1471-2229-10-255

Bennett, M. D. (1971). The Duration of Meiosis. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, *178*(1052), 277–299.

Bennetzen, J. L., Qin, M.-M., Ingels, S., & Ellingboe, A. H. (1988). Allele-specific and Mutator-associated instability at the Rpl disease-resistance locus of maize. *Nature*, *332*(6162), 369–370. https://doi.org/10.1038/332369a0

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 Genetics*, *3*(8), e132. https://doi.org/10.1371/journal.pgen.0030132

Bergerat, A., de Massy, B., Gadelle, D., Varoutas, P.-C., Nicolas, A., & Forterre, P. (1997). An atypical topoisomerase II from archaea with implications for meiotic recombination. *Nature*, *386*(6623), 414–417. https://doi.org/10.1038/386414a0

Bertioli, D. J., Seijo, G., Freitas, F. O., Valls, J. F. M., Leal-Bertioli, S. C. M., & Moretzsohn, M. C. (2011). An overview of peanut and its wild relatives. *Plant Genetic Resources*, *9*(1), 134–149. https://doi.org/10.1017/S1479262110000444

Bertioli, D. J., Cannon, S. B., Froenicke, L., Huang, G., Farmer, A. D., Cannon, E. K., Liu, X., Gao, D., Clevenger, J., & Dash, S. (2016). The genome sequences of *Arachis duranensis* and *Arachis ipaensis*, the diploid ancestors of cultivated peanut. *Nature Genetics*, 48(4), 438–446.

Bertioli, D. J., Jenkins, J., Clevenger, J., Dudchenko, O., Gao, D., Seijo, G., Leal-Bertioli, S., Ren, L., Farmer, A. D., & Pandey, M. K. (2019). The genome sequence of segmental allotetraploid peanut *Arachis hypogae*a. *Nature Genetics*, 51(5), 877–884.

Bishop, D. K. (1994). RecA homologs Dmc1 and Rad51 interact to form multiple nuclear complexes prior to meiotic chromosome synapsis. *Cell*, *79*(6), 1081–1092. https://doi.org/10.1016/0092-8674(94)90038-8

Bishop, D. K., Park, D., Xu, L., & Kleckner, N. (1992). DMC1: A meiosis-specific yeast homolog of E. coli recA required for recombination, synaptonemal complex formation, and cell cycle progression. *Cell*, *69*(3), 439–456. https://doi.org/10.1016/0092-8674(92)90446-J

Blackwell, A. R., Dluzewska, J., Szymanska-Lejman, M., Desjardins, S., Tock, A. J., Kbiri, N., Lambing, C., Lawrence, E. J., Bieluszewski, T., Rowan, B., Higgins, J. D., Ziolkowski, P. A., & Henderson, I. R. (2020). MSH2 shapes the meiotic crossover landscape in relation to interhomolog polymorphism in Arabidopsis. *The EMBO Journal*, *39*(21), e104858. https://doi.org/10.15252/embj.2020104858

Blakeslee, A. F. (1922). Variations in Datura Due to Changes in Chromosome Number. *The American Naturalist*, *56*(642), 16–31. https://doi.org/10.1086/279845

Blakeslee, A. F., & Avery, A. G. (1937). Methods of inducing chromosome doubling in plants. *Journal of Heredity*, *28*(12), 393–411. https://doi.org/10.1093/oxfordjournals.jhered.a104294

Borde, V., & de Massy, B. (2013). Programmed induction of DNA double strand breaks during meiosis: Setting up communication between DNA and the

chromosome structure. *Current Opinion in Genetics & Development*, *23*(2), 147–155. https://doi.org/10.1016/j.gde.2012.12.002

Borde, V., Robine, N., Lin, W., Bonfils, S., Géli, V., & Nicolas, A. (2009). Histone H3 lysine 4 trimethylation marks meiotic recombination initiation sites. *The EMBO Journal*, *28*(2), 99–111. https://doi.org/10.1038/emboj.2008.257

Börner, G. V., Kleckner, N., & Hunter, N. (2004). Crossover/noncrossover differentiation, synaptonemal complex formation, and regulatory surveillance at the leptotene/zygotene transition of meiosis. *Cell*, *117*(1), 29–45. https://doi.org/10.1016/s0092-8674(04)00292-2

Bravo, J. P., Hoshino, A. A., Angelici, C. M. L. C. D., Lopes, C. R., & Gimenes, M. A. (2006). Transferability and use of microsatellite markers for the genetic analysis of the germplasm of some Arachis section species of the genus Arachis. *Genetics and Molecular Biology*, *29*, 516–524. https://doi.org/10.1590/S1415-47572006000300021

Brick, K., Smagulova, F., Khil, P., Camerini-Otero, R. D., & Petukhova, G. V. (2012). Genetic recombination is directed away from functional genomic elements in mice. *Nature*, *485*(7400), 642–645. https://doi.org/10.1038/nature11089

Bu, Y., & Cao, D. (2012). The origin of cancer stem cells. *Frontiers in bioscience* (Scholar edition), 4(3), 819–830. https://doi.org/10.2741/s302

Burow, M. D., Simpson, C. E., Starr, J. L., & Paterson, A. H. (2001). Transmission genetics of chromatin from a synthetic amphidiploid to cultivated peanut (*Arachis hypogaea* L.). Broadening the gene pool of a monophyletic polyploid species. *Genetics*, *159*(2), 823–837. https://doi.org/10.1093/genetics/159.2.823

Cai, X., & Xu, S. S. (2007). Meiosis-Driven Genome Variation in Plants. *Current Genomics*, *8*(3), 151–161. https://doi.org/10.2174/138920207780833847

Cao, X., Aufsatz, W., Zilberman, D., Mette, M. F., Huang, M. S., Matzke, M., & Jacobsen, S. E. (2003). Role of the DRM and CMT3 methyltransferases in RNA-directed DNA methylation. *Current Biology: CB*, *13*(24), 2212–2217. https://doi.org/10.1016/j.cub.2003.11.052

Capilla-Pérez, L., Durand, S., Hurel, A., Lian, Q., Chambon, A., Taochy, C., Solier, V., Grelon, M., & Mercier, R. (2021). The synaptonemal complex imposes crossover interference and heterochiasmy in Arabidopsis. *Proceedings of the National Academy of Sciences*, *118*(12), e2023613118. https://doi.org/10.1073/pnas.2023613118

Carmeille, A., Caranta, C., Dintinger, J., Prior, P., Luisetti, J., & Besse, P. (2006). Identification of QTLs for Ralstonia solanacearum race 3-phylotype II

resistance in tomato. *Theoretical and Applied Genetics*, *113*(1), 110–121. https://doi.org/10.1007/s00122-006-0277-3

Carpenter, A. T. C. (1975). Electron microscopy of meiosis in Drosophila melanogaster females: II: The recombination nodule—a recombination-associated structure at pachytene? *Proceedings of the National Academy of Sciences*, *72*(8), 3186–3189. https://doi.org/10.1073/pnas.72.8.3186

Cervantes, M. D., Farah, J. A., & Smith, G. R. (2000). Meiotic DNA Breaks Associated with Recombination in S. pombe. *Molecular Cell*, *5*(5), 883–888. https://doi.org/10.1016/S1097-2765(00)80328-7

Chelysheva, L., Gendrot, G., Vezon, D., Doutriaux, M.-P., Mercier, R., & Grelon, M. (2007). Zip4/Spo22 Is Required for Class I CO Formation but Not for Synapsis Completion in Arabidopsis thaliana. *PLoS Genetics*, *3*(5), e83. https://doi.org/10.1371/journal.pgen.0030083

Chelysheva, L., Vezon, D., Belcram, K., Gendrot, G., & Grelon, M. (2008). The Arabidopsis BLAP75/Rmi1 Homologue Plays Crucial Roles in Meiotic Double-Strand Break Repair. *PLOS Genetics*, *4*(12), e1000309. https://doi.org/10.1371/journal.pgen.1000309

Chelysheva, L., Vezon, D., Chambon, A., Gendrot, G., Pereira, L., Lemhemdi, A., Vrielynck, N., Le Guin, S., Novatchkova, M., & Grelon, M. (2012). The Arabidopsis HEI10 Is a New ZMM Protein Related to Zip3. *PLoS Genetics*, *8*(7), e1002799. https://doi.org/10.1371/journal.pgen.1002799

Chen, X., Suhandynata, R. T., Sandhu, R., Rockmill, B., Mohibullah, N., Niu, H., Liang, J., Lo, H.-C., Miller, D. E., Zhou, H., Börner, G. V., & Hollingsworth, N. M. (2015). Phosphorylation of the Synaptonemal Complex Protein Zip1 Regulates the Crossover/Noncrossover Decision during Yeast Meiosis. *PLOS Biology*, *13*(12), e1002329. https://doi.org/10.1371/journal.pbio.1002329

Cheng, C.-H., Lo, Y.-H., Liang, S.-S., Ti, S.-C., Lin, F.-M., Yeh, C.-H., Huang, H.-Y., & Wang, T.-F. (2006). SUMO modifications control assembly of synaptonemal complex and polycomplex in meiosis of Saccharomyces cerevisiae. *Genes & Development*, *20*(15), 2067–2081. https://doi.org/10.1101/gad.1430406

Chetelat, R. T., Meglic, V., & Cisneros, P. (2000). A Genetic Map of Tomato Based on BC1 Lycopersicon esculentum × Solanum lycopersicoides Reveals Overall Synteny but Suppressed Recombination Between These Homeologous Genomes. *Genetics*, *154*(2), 857–867. https://doi.org/10.1093/genetics/154.2.857

Choi, K., Reinhard, C., Serra, H., Ziolkowski, P. A., Underwood, C. J., Zhao, X., Hardcastle, T. J., Yelina, N. E., Griffin, C., Jackson, M., Mézard, C., McVean,

G., Copenhaver, G. P., & Henderson, I. R. (2016). Recombination Rate Heterogeneity within Arabidopsis Disease Resistance Genes. *PLOS Genetics*, *12*(7), e1006179. https://doi.org/10.1371/journal.pgen.1006179

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. H., McVean, G., & Henderson, I. R. (2013). Arabidopsis meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters. *Nature Genetics*, *45*(11), 1327–1336. https://doi.org/10.1038/ng.2766

Choi, K., Zhao, X., Tock, A. J., Lambing, C., Underwood, C. J., Hardcastle, T. J., Serra, H., Kim, J., Cho, H. S., Kim, J., Ziolkowski, P. A., Yelina, N. E., Hwang, I., Martienssen, R. A., & Henderson, I. R. (2018). Nucleosomes and DNA methylation shape meiotic DSB frequency in Arabidopsis thaliana transposons and gene regulatory regions. *Genome Research*. https://doi.org/10.1101/gr.225599.117

Choulet, F., Alberti, A., Theil, S., Glover, N., Barbe, V., Daron, J., Pingault, L., Sourdille, P., Couloux, A., Paux, E., Leroy, P., Mangenot, S., Guilhot, N., Le Gouis, J., Balfourier, F., Alaux, M., Jamilloux, V., Poulain, J., Durand, C., … Feuillet, C. (2014). Structural and functional partitioning of bread wheat chromosome 3B. *Science*, *345*(6194), 1249721. https://doi.org/10.1126/science.1249721

Corbett-Detig, R. B., Hartl, D. L., & Sackton, T. B. (2015). Natural Selection Constrains Neutral Diversity across A Wide Range of Species. *PLOS Biology*, *13*(4), e1002112. https://doi.org/10.1371/journal.pbio.1002112

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*(6088), 1588–1590. https://doi.org/10.1126/science.1220381

Cunha, F., Nobile, P., Hoshino, A., Moretzsohn, M., Lopes, C., & Gimenes, M. (2007). Genetic relationships among Arachis hypogaea L. (AABB) and diploid Arachis species with AA and BB genomes. *Genetic Resources and Crop Evolution*, *55*, 15–20. https://doi.org/10.1007/s10722-007-9209-6

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*(3), 1373–1388. https://doi.org/10.1534/genetics.116.196014

de Boer, E., Stam, P., Dietrich, A. J. J., Pastink, A., & Heyting, C. (2006). Two levels of interference in mouse meiotic recombination. *Proceedings of the*

National Academy of Sciences, *103*(25), 9607–9612. https://doi.org/10.1073/pnas.0600418103

de Massy, B. (2013). Initiation of meiotic recombination: How and where? Conservation and specificities among eukaryotes. *Annual Review of Genetics*, *47*, 563–599. https://doi.org/10.1146/annurev-genet-110711-155423

De Muyt, A., Pyatnitskaya, A., Andréani, J., Ranjha, L., Ramus, C., Laureau, R., Fernandez-Vega, A., Holoch, D., Girard, E., Govin, J., Margueron, R., Couté, Y., Cejka, P., Guérois, R., & Borde, V. (2018). A meiotic XPF–ERCC1-like complex recognizes joint molecule recombination intermediates to promote crossover formation. *Genes & Development*, *32*(3–4), 283–296. https://doi.org/10.1101/gad.308510.117

De Muyt, A., Jessop, L., Kolar, E., Sourirajan, A., Chen, J., Dayani, Y., & Lichten, M. (2012). BLM Helicase Ortholog Sgs1 Is a Central Regulator of Meiotic Recombination Intermediate Metabolism. *Molecular Cell*, *46*(1), 43–53. https://doi.org/10.1016/j.molcel.2012.02.020

Desjardins, S. D., Ogle, D. E., Ayoub, M. A., Heckmann, S., Henderson, I. R., Edwards, K. J., & Higgins, J. D. (2020). MutS homologue 4 and MutS homologue 5 Maintain the Obligate Crossover in Wheat Despite Stepwise Gene Loss following Polyploidization. *Plant Physiology*, *183*(4), 1545–1558. https://doi.org/10.1104/pp.20.00534

Desmae, H., Janila, P., Okori, P., Pandey, M. K., Motagi, B. N., Monyo, E., Mponda, O., Okello, D., Sako, D., & Echeckwu, C. (2019). Genetics, genomics and breeding of groundnut (*Arachis hypogaea* L.). *Plant Breeding*, 138(4), 425–444.

Dewey, D. R. (1980). Some Applications and Misapplications of Induced Polyploidy to Plant Breeding. In W. H. Lewis (Ed.), *Polyploidy: Biological Relevance* (pp. 445–470). Springer US. https://doi.org/10.1007/978-1-4613-3069-1_23

Dillehay, T. D., Rossen, J., Andres, T. C., & Williams, D. E. (2007). Preceramic Adoption of Peanut, Squash, and Cotton in Northern Peru. *Science*, *316*(5833), 1890–1893. https://doi.org/10.1126/science.1141395

Dluzewska, J., Szymanska, M., & Ziolkowski, P. A. (2018). Where to Cross Over? Defining Crossover Sites in Plants. *Frontiers in Genetics*, *9*. https://www.frontiersin.org/article/10.3389/fgene.2018.00609

Dubin, M. J., Mittelsten Scheid, O., & Becker, C. (2018). Transposons: A blessing curse. *Current Opinion in Plant Biology*, *42*, 23–29. https://doi.org/10.1016/j.pbi.2018.01.003

Duroc, Y., Kumar, R., Ranjha, L., Adam, C., Guérois, R., Md Muntaz, K., Marsolier-Kergoat, M.-C., Dingli, F., Laureau, R., Loew, D., Llorente, B., Charbonnier, J.-B., Cejka, P., & Borde, V. (2017). Concerted action of the MutLβ heterodimer and Mer3 helicase regulates the global extent of meiotic gene conversion. *ELife*, *6*, e21900. https://doi.org/10.7554/eLife.21900

Ellegren, H., & Galtier, N. (2016). Determinants of genetic diversity. *Nature Reviews Genetics*, *17*(7), 422–433. https://doi.org/10.1038/nrg.2016.58

Espagne, E., Vasnier, C., Storlazzi, A., Kleckner, N. E., Silar, P., Zickler, D., & Malagnac, F. (2011). Sme4 coiled-coil protein mediates synaptonemal complex assembly, recombinosome relocalization, and spindle pole body morphogenesis. *Proceedings of the National Academy of Sciences*, *108*(26), 10614–10619. https://doi.org/10.1073/pnas.1107272108

Falque, M., Anderson, L. K., Stack, S. M., Gauthier, F., & Martin, O. C. (2009). Two Types of Meiotic Crossovers Coexist in Maize. *The Plant Cell*, 21(12), 3915–3925. https://doi.org/10.1105/tpc.109.071514

Fan, C., Xing, Y., Mao, H., Lu, T., Han, B., Xu, C., Li, X., & Zhang, Q. (2006). GS3, a major QTL for grain length and weight and minor QTL for grain width and thickness in rice, encodes a putative transmembrane protein. *Theoretical and Applied Genetics*, *112*(6), 1164–1171. https://doi.org/10.1007/s00122-006-0218-1

Favero, A. P., Simpson, C. E., Valls, J. F. M., & Vello, N. A. (2006). Study of the Evolution of Cultivated Peanut through Crossability Studies among Arachis ipaënsis, *A. duranensis*, and *A. hypogaea*. *Crop Science*, 46(4), 1546-1552. http://dx.doi.org/10.2135/cropsci2005.09-0331

Fegan, M., & Prior, P. (2005). How complex is the Ralstonia solanacearum species complex. In *Bacterial wilt disease and the Ralstonia solanacearum species complex*. APS Press. https://hal.inrae.fr/hal-02833579

Feng, S., Cokus, S. J., Schubert, V., Zhai, J., Pellegrini, M., & Jacobsen, S. E. (2014). Genome-wide Hi-C Analyses in Wild-Type and Mutants Reveal High-Resolution Chromatin Interactions in Arabidopsis. *Molecular Cell*, *55*(5), 694–707. https://doi.org/10.1016/j.molcel.2014.07.008

Fernandes, J. B., Séguéla-Arnaud, M., Larchevêque, C., Lloyd, A. H., & Mercier, R. (2018). Unleashing meiotic crossovers in hybrid plants. *Proceedings of the National Academy of Sciences*, *115*(10), 2431–2436. https://doi.org/10.1073/pnas.1713078114

Foncéka, D., Hodo-Abalo, T., Rivallan, R., Faye, I., Sall, M. N., Ndoye, O., Fávero, A. P., Bertioli, D. J., Glaszmann, J.-C., Courtois, B., & Rami, J.-F. (2009). Genetic mapping of wild introgressions into cultivated peanut: A way toward

enlarging the genetic basis of a recent allotetraploid. *BMC Plant Biology*, *9*(1), 103. https://doi.org/10.1186/1471-2229-9-103

Fozard, J. A., Morgan, C., & Howard, M. (2022). The synaptonemal complex controls cis- versus trans-interference in coarsening-based meiotic crossover patterning. *BioRxiv*, 2022.04.11.487855. https://doi.org/10.1101/2022.04.11.487855

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. *The Plant Cell*, *11*(5), 809–824. https://doi.org/10.1105/tpc.11.5.809

Fransz, P., de Jong, J. H., Lysak, M., Castiglione, M. R., & Schubert, I. (2002). Interphase chromosomes in Arabidopsis are organized as well defined chromocenters from which euchromatin loops emanate. *Proceedings of the National Academy of Sciences*, *99*(22), 14584–14589. https://doi.org/10.1073/pnas.212325299

Fung, J. C., Rockmill, B., Odell, M., & Roeder, G. S. (2004). Imposition of crossover interference through the nonrandom distribution of synapsis initiation complexes. *Cell*, *116*(6), 795–802. https://doi.org/10.1016/s0092-8674(04)00249-1

Gamba, D., & Muchhala, N. (2020). Global patterns of population genetic differentiation in seed plants. *Molecular Ecology*, *29*(18), 3413–3428. https://doi.org/10.1111/mec.15575

Garcia, V., Phelps, S. E. L., Gray, S., & Neale, M. J. (2011). Bidirectional resection of DNA double-strand breaks by Mre11 and Exo1. *Nature*, *479*(7372), 241–244. https://doi.org/10.1038/nature10515

Gaut, B. S., Seymour, D. K., Liu, Q., & Zhou, Y. (2018). Demography and its effects on genomic variation in crop domestication. *Nature Plants*, *4*(8), 512–520. https://doi.org/10.1038/s41477-018-0210-1

Gilles, L. M., Khaled, A., Laffaire, J.-B., Chaignon, S., Gendrot, G., Laplaige, J., Bergès, H., Beydon, G., Bayle, V., Barret, P., Comadran, J., Martinant, J.-P., Rogowsky, P. M., & Widiez, T. (2017). Loss of pollen-specific phospholipase NOT LIKE DAD triggers gynogenesis in maize. *The EMBO Journal*, *36*(6), 707–717. https://doi.org/10.15252/embj.201796603

Gimenes, M. A., Hoshino, A. A., Barbosa, A. V., Palmieri, D. A., & Lopes, C. R. (2007). Characterization and transferability of microsatellite markers of the cultivated peanut (Arachis hypogaea). *BMC Plant Biology*, *7*, 9. https://doi.org/10.1186/1471-2229-7-9

Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O. C., & Mézard, C. (2011). Genome-Wide Crossover Distribution in Arabidopsis thaliana Meiosis Reveals Sex-Specific Patterns along Chromosomes. *PLOS Genetics*, *7*(11), e1002354. https://doi.org/10.1371/journal.pgen.1002354

Gray, S., & Cohen, P. E. (2016). Control of Meiotic Crossovers: From Double-Strand Break Formation to Designation. *Annual Review of Genetics*, *50*, 175–210. https://doi.org/10.1146/annurev-genet-120215-035111

Greenblatt, I. M., & Brink, R. A. (1962). Twin Mutations in Medium Variegated Pericarp Maize. *Genetics*, *47*(4), 489–501. https://doi.org/10.1093/genetics/47.4.489

Gregory, M. P., & Gregory, W. C. (1979). Exotic germ plasm of Arachis L. interspecific hybrids. *Journal of Heredity*, *70*(3), 185–193. https://doi.org/10.1093/oxfordjournals.jhered.a109231

Grelon, M., Vezon, D., Gendrot, G., & Pelletier, G. (2001). AtSPO11-1 is necessary for efficient meiotic recombination in plants. *EMBO Journal*, *20*(3), 589–600.

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

Hai, T.T.H. (2007). Characterisation and mapping of bacterial wilt (Ralstonia solanacearum) resistance in the tomato (Solanum lycopersicum) cultivar Hawaii 7996 and wild tomato germplasm. MSc Thesis, Gottfried Wilhelm Leibniz Universität Hannover, Hannover, Germany.

Halward, T., Stalker, T., LaRue, E., & Kochert, G. (1992). Use of single-primer DNA amplifications in genetic studies of peanut (Arachis hypogaea L.). *Plant Molecular Biology*, *18*(2), 315–325. https://doi.org/10.1007/BF00034958

Hanson, P., Lu, S.-F., Wang, J.-F., Chen, W., Kenyon, L., Tan, C.-W., Tee, K. L., Wang, Y.-Y., Hsu, Y.-C., Schafleitner, R., Ledesma, D., & Yang, R.-Y. (2016). Conventional and molecular marker-assisted selection and pyramiding of genes for multiple disease resistance in tomato. *Scientia Horticulturae*, *201*, 346–354. https://doi.org/10.1016/j.scienta.2016.02.020

Hartman, G. L., Hong, W. F., & Wang, T. C. (1991). Survey of bacterial wilt on fresh market hybrid tomatoes in Taiwan. *PLANT PROTECTION BULLETIN [TW]*, *v.33(2):197-203*. http://worldveg.tind.io/record/16800/files/aps397.pdf

Hartung, F., Suer, S., Knoll, A., Wurz-Wildersinn, R., & Puchta, H. (2008). Topoisomerase 3α and RMI1 Suppress Somatic Crossovers and Are Essential for Resolution of Meiotic Recombination Intermediates in Arabidopsis thaliana. *PLOS Genetics*, *4*(12), e1000285. https://doi.org/10.1371/journal.pgen.1000285

Hartung, F., Wurz-Wildersinn, R., Fuchs, J., Schubert, I., Suer, S., & Puchta, H. (2007). The Catalytically Active Tyrosine Residues of Both SPO11-1 and SPO11-2 Are Required for Meiotic Double-Strand Break Induction in Arabidopsis. *The Plant Cell*, *19*(10), 3090–3099. https://doi.org/10.1105/tpc.107.054817

Hayward, A. C. (1991). Biology and Epidemiology of Bacterial Wilt Caused by Pseudomonas Solanacearum. *Annual Review of Phytopathology*, *29*(1), 65–87. https://doi.org/10.1146/annurev.py.29.090191.000433

He Yan, Wang Minghui, Dukowic-Schulze Stefanie, Zhou Adele, Tiang Choon-Lin, Shilo Shay, Sidhu Gaganpreet K., Eichten Steven, Bradbury Peter, Springer Nathan M., Buckler Edward S., Levy Avraham A., Sun Qi, Pillardy Jaroslaw, Kianian Penny M. A., Kianian Shahryar F., Chen Changbin, & Pawlowski Wojciech P. (2017). Genomic features shaping the landscape of meiotic double-strand-break hotspots in maize. *Proceedings of the National Academy of Sciences*, *114*(46), 12231–12236. https://doi.org/10.1073/pnas.1713225114

Higgins, J. D., Armstrong, S. J., Franklin, F. C. H., & Jones, G. H. (2004). The Arabidopsis MutS homolog AtMSH4 functions at an early step in recombination: Evidence for two classes of recombination in Arabidopsis. *Genes & Development*, *18*(20), 2557–2570. https://doi.org/10.1101/gad.317504

Higgins, J. D., Vignard, J., Mercier, R., Pugh, A. G., Franklin, F. C. H., & Jones, G. H. (2008). AtMSH5 partners AtMSH4 in the class I meiotic crossover pathway in Arabidopsis thaliana, but is not required for synapsis. *The Plant Journal*, *55*(1), 28–39. https://doi.org/10.1111/j.1365-313X.2008.03470.x

Holbrook, C. C., Anderson, W. F., & Pittman, R. N. (1993). Selection of a Core Collection from the U.S. Germplasm Collection of Peanut. *Crop Science*, *33*(4), cropsci1993.0011183X003300040044x. https://doi.org/10.2135/cropsci1993.0011183X003300040044x

Holbrook, C. C., & Dong, W. (2005). Development and Evaluation of a Mini Core Collection for the U.S. Peanut Germplasm Collection. *Crop Science*, *45*(4), 1540–1544. https://doi.org/10.2135/cropsci2004.0368

Hunter, N. (2015). Meiotic Recombination: The Essence of Heredity. *Cold Spring Harbor Perspectives in Biology*, *7*(12), a016618. https://doi.org/10.1101/cshperspect.a016618

Husted, L. (1936). Cytological Studies an the Peanut, *Arachis*. II. *Cytologia*, *7*(3), 396–423. https://doi.org/10.1508/cytologia.7.396

Innan, H., Terauchi, R., & Miyashita, N. T. (1997). Microsatellite polymorphism in natural populations of the wild plant Arabidopsis thaliana. *Genetics*, *146*(4), 1441–1452. https://doi.org/10.1093/genetics/146.4.1441

Jackson, C. A., Castro, D. M., Saldi, G.-A., Bonneau, R., & Gresham, D. (2020). Gene regulatory network reconstruction using single-cell RNA sequencing of barcoded genotypes in diverse environments. *ELife*, *9*, e51254. https://doi.org/10.7554/eLife.51254

Jackson, N., Sanchez-Moran, E., Buckling, E., Armstrong, S. J., Jones, G. H., & Franklin, F. C. H. (2006). Reduced meiotic crossovers and delayed prophase I progression in AtMLH3-deficient Arabidopsis. *The EMBO Journal*, *25*(6), 1315–1323. https://doi.org/10.1038/sj.emboj.7600992

Jessop, L., Rockmill, B., Roeder, G. S., & Lichten, M. (2006). Meiotic Chromosome Synapsis-Promoting Proteins Antagonize the Anti-Crossover Activity of Sgs1. *PLOS Genetics*, *2*(9), e155. https://doi.org/10.1371/journal.pgen.0020155

Jiang, H. F., Ren, X. P., Liao, B. S., Huang, J. Q., Lei, Y., Chen, B. Y., Guo, B. Z., Holbrook, C. C., & Upadhyaya, H. D. (2008). Peanut Core Collection Established in China and Compared with ICRISAT Mini Core Collection. *Acta Agronomica Sinica*, *34*(01), 25–30.

Jiang, H.-F., Ren, X.-P., Zhang, X.-J., Huang, J.-Q., Lei, Y., Yan, L.-Y., Liao, B.-S., Upadhyaya, H. D., & Holbrook, C. C. (2010a). Comparison of Genetic Diversity Based on SSR Markers Between Peanut Mini Core Collections from China and ICRISAT. *Acta Agronomica Sinica*, *36*(7), 1084–1091. https://doi.org/10.1016/S1875-2780(09)60059-6

Jiang, H.-F., Ren, X.-P., Zhang, X.-J., Huang, J.-Q., Lei, Y., Yan, L.-Y., Liao, B.-S., Upadhyaya, H. D., & Holbrook, C. C. (2010b). Comparison of Genetic Diversity Based on SSR Markers Between Peanut Mini Core Collections from China and ICRISAT. *Acta Agronomica Sinica*, *36*(7), 1084–1091. https://doi.org/10.1016/S1875-2780(09)60059-6

J.M., P., & D.A. Sleper. (1995). *Breeding Field Crops* (XV). Iowa State University Press.

Jones, G. H., & Franklin, F. C. H. (2006). Meiotic crossing-over: Obligation and interference. *Cell*, 126(2), 246–248. https://doi.org/10.1016/j.cell.2006.07.010

Jyothi, H. K., Santhosha, H. M., & Basamma. (2012). Recent advances in breeding for bacterial wilt (*Ralstonia solanacearum*) resistance in tomato—Review. *Current Biotica*, *6*(3), 370–398.

Kanfany, G., Serba, D. D., Rhodes, D., St. Amand, P., Bernardo, A., Gangashetty, P. I., Kane, N. A., & Bai, G. (2020). Genomic diversity in pearl millet inbred

lines derived from landraces and improved varieties. *BMC Genomics*, *21*(1), 469. https://doi.org/10.1186/s12864-020-06796-4

Karumannil, S., Sadhankumar, P. G., Nazeem, P., Girija, D. K., & Kesavachandran, R. (2008). DNA fingerprinting of bacterial wilt resistant tomato (Solanum lycopersicum L.) cultivars. *Veg Sci*, *35*, 105–108.

Kauppi, L., Barchi, M., Baudat, F., Romanienko, P. J., Keeney, S., & Jasin, M. (2011). Distinct properties of the XY pseudoautosomal region crucial for male meiosis. *Science*, *331*(6019), 916–920. https://doi.org/10.1126/science.1195774

Kaur, H., De Muyt, A., & Lichten, M. (2015). Top3-Rmi1 DNA Single-Strand Decatenase Is Integral to the Formation and Resolution of Meiotic Recombination Intermediates. *Molecular Cell*, *57*(4), 583–594. https://doi.org/10.1016/j.molcel.2015.01.020

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*(3), 375–384. https://doi.org/10.1016/s0092-8674(00)81876-0

Khush, G. S. (2001). Green revolution: The way forward. *Nature Reviews Genetics*, *2*(10), 815–822. https://doi.org/10.1038/35093585

Kianian, P. M. A., Wang, M., Simons, K., Ghavami, F., He, Y., Dukowic-Schulze, S., Sundararajan, A., Sun, Q., Pillardy, J., Mudge, J., Chen, C., Kianian, S. F., & Pawlowski, W. P. (2018). High-resolution crossover mapping reveals similarities and differences of male and female recombination in maize. *Nature Communications*, *9*(1), 2370. https://doi.org/10.1038/s41467-018-04562-5

Kim, K. P., Weiner, B. M., Zhang, L., Jordan, A., Dekker, J., & Kleckner, N. (2010). Sister cohesion and structural axis components mediate homolog bias of meiotic recombination. *Cell*, *143*(6), 924–937. https://doi.org/10.1016/j.cell.2010.11.015

King, J. S., & Mortimer, R. K. (1990). A polymerization model of chiasma interference and corresponding computer simulation. *Genetics*, 126(4), 1127–1138. https://doi.org/10.1093/genetics/126.4.1127

Kochert, G., Stalker, H. T., Gimenes, M., Galgaro, L., Lopes, C. R., & Moore, K. (1996). RFLP and cytogenetic evidence on the origin and evolution of allotetraploid domesticated peanut, *Arachis hypogaea* (Leguminosae). *American Journal of Botany (USA)*.

Kosambi, D. D. The estimation of map distances from recombination values. *Annals of Eugenics* 12, 172–175 (1943).

Kujur, A., Upadhyaya, H. D., Shree, T., Bajaj, D., Das, S., Saxena, M. S., Badoni, S., Kumar, V., Tripathi, S., Gowda, C. L. L., Sharma, S., Singh, S., Tyagi, A. K., & Parida, S. K. (2015). Ultra-high density intra-specific genetic linkage maps accelerate identification of functionally relevant molecular tags governing important agronomic traits in chickpea. *Scientific Reports*, *5*, 9468. https://doi.org/10.1038/srep09468

Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews Genetics*, *11*(3), 204–220. https://doi.org/10.1038/nrg2719

Li, P., Tao, Z., & Dean, C. (2015). Phenotypic evolution through variation in splicing of the noncoding RNA COOLAIR. *Genes & Development*, *29*(7), 696–701. https://doi.org/10.1101/gad.258814.115

Li, X., Xu, H., Feng, J., Zhou, X., & Chen, J. (2015). Mining of genic SNPs and diversity evaluation of landraces in loquat. *Scientia Horticulturae*, *195*, 82–88. https://doi.org/10.1016/j.scienta.2015.08.040

Lin, C.-H., Tsai, K.-C., Prior, P., & Wang, J.-F. (2014). Phylogenetic relationships and population structure of *Ralstonia solanacearum* isolated from diverse origins in Taiwan. *Plant Pathology*, *63*(6), 1395–1403. https://doi.org/10.1111/ppa.12209

Ma, J., Devos, K. M., & Bennetzen, J. L. (2004). Analyses of LTR-Retrotransposon Structures Reveal Recent and Rapid Genomic DNA Loss in Rice. *Genome Research*, *14*(5), 860–869. https://doi.org/10.1101/gr.1466204

Macaisne, N., Novatchkova, M., Peirera, L., Vezon, D., Jolivet, S., Froger, N., Chelysheva, L., Grelon, M., & Mercier, R. (2008). SHOC1, an XPF endonuclease-related protein, is essential for the formation of class I meiotic crossovers. *Current Biology: CB*, *18*(18), 1432–1437. https://doi.org/10.1016/j.cub.2008.08.041

Maia, A. de H., Luiz, A. J., & Campanhola, C. (2000). Statistical inference on associated fertility life table parameters using jackknife technique: Computational aspects. *Journal of Economic Entomology*, *93*(2), 511–518. https://doi.org/10.1603/0022-0493-93.2.511

Malik, S.-B., Ramesh, M. A., Hulstrand, A. M., & Logsdon, J. M., Jr. (2007). Protist Homologs of the Meiotic Spo11 Gene and Topoisomerase VI reveal an Evolutionary History of Gene Duplication and Lineage-Specific Loss. *Molecular Biology and Evolution*, *24*(12), 2827–2841. https://doi.org/10.1093/molbev/msm217

Mao, B., Zheng, W., Huang, Z., Peng, Y., Shao, Y., Liu, C., Tang, L., Hu, Y., Li, Y., Hu, L., Zhang, D., Yuan, Z., Luo, W., Yuan, L., Liu, Y., & Zhao, B. (2021).

Rice MutLγ, the MLH1–MLH3 heterodimer, participates in the formation of type I crossovers and regulation of embryo sac fertility. *Plant Biotechnology Journal*, *19*(7), 1443–1455. https://doi.org/10.1111/pbi.13563

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. *The Plant Cell*, *31*(3), 645–662. https://doi.org/10.1105/tpc.18.00750

Masson, J. Y., & West, S. C. (2001). The Rad51 and Dmc1 recombinases: A non-identical twin relationship. *Trends in Biochemical Sciences*, *26*(2), 131–136. https://doi.org/10.1016/s0968-0004(00)01742-4

McPeek, M. S., & Speed, T. P. (1995). Modeling interference in genetic recombination. *Genetics*, 139(2), 1031–1044. https://doi.org/10.1093/genetics/139.2.1031

Melamed-Bessudo, C., & Levy, A. A. (2012). Deficiency in DNA methylation increases meiotic crossover rates in euchromatic but not in heterochromatic regions in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, *109*(16), E981–E988. https://doi.org/10.1073/pnas.1120742109

Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N., & Grelon, M. (2015). The molecular biology of meiosis in plants. *Annual Review of Plant Biology*, *66*, 297–327. https://doi.org/10.1146/annurev-arplant-050213-035923

Milla, S. R., Isleib, T. G., & Stalker, H. T. (2005). Taxonomic relationships among *Arachis* sect. *Arachis* species as revealed by AFLP markers. *Genome*, *48*(1), 1–11. https://doi.org/10.1139/g04-089

Mirouze Marie, Lieberman-Lazarovich Michal, Aversano Riccardo, Bucher Etienne, Nicolet Joël, Reinders Jon, & Paszkowski Jerzy. (2012). Loss of DNA methylation affects the recombination landscape in *Arabidopsis*. *Proceedings of the National Academy of Sciences*, *109*(15), 5880–5885. https://doi.org/10.1073/pnas.1120841109

Miyao, A., Tanaka, K., Murata, K., Sawaki, H., Takeda, S., Abe, K., Shinozuka, Y., Onosato, K., & Hirochika, H. (2003). Target Site Specificity of the *Tos17* Retrotransposon Shows a Preference for Insertion within Genes and against Insertion in Retrotransposon-Rich Regions of the Genome. *The Plant Cell*, *15*(8), 1771–1780. https://doi.org/10.1105/tpc.012559

Moens, P. B., Kolas, N. K., Tarsounas, M., Marcon, E., Cohen, P. E., & Spyropoulos, B. (2002). The time course and chromosomal localization of recombination-related proteins at meiosis in the mouse are compatible with models that can resolve the early DNA-DNA interactions without reciprocal

recombination. *Journal of Cell Science*, *115*(Pt 8), 1611–1622. https://doi.org/10.1242/jcs.115.8.1611

Morgan, C., Fozard, J.A., Hartley, M. et al. Diffusion-mediated HEI10 coarsening can explain meiotic crossover positioning in *Arabidopsis*. *Nature Communications* 12, 4674 (2021). https://doi.org/10.1038/s41467-021-24827-w

Monroe, J. G., Srikant, T., Carbonell-Bejerano, P., Becker, C., Lensink, M., Exposito-Alonso, M., Klein, M., Hildebrandt, J., Neumann, M., Kliebenstein, D., Weng, M.-L., Imbert, E., Ågren, J., Rutter, M. T., Fenster, C. B., & Weigel, D. (2022). Mutation bias reflects natural selection in *Arabidopsis thaliana*. *Nature*, *602*(7895), 101–105. https://doi.org/10.1038/s41586-021-04269-6

Moretzsohn, M. C., Gouvea, E. G., Inglis, P. W., Leal-Bertioli, S. C. M., Valls, J. F. M., & Bertioli, D. J. (2013). A study of the relationships of cultivated peanut (*Arachis hypogaea*) and its most closely related wild species using intron sequences and microsatellite markers. *Annals of Botany*, *111*(1), 113–126. https://doi.org/10.1093/aob/mcs237

Moretzsohn, M. C., Hopkins, M. S., Mitchell, S. E., Kresovich, S., Valls, J. F. M., & Ferreira, M. E. (2004). Genetic diversity of peanut (*Arachis hypogaea* L.) and its wild relatives based on the analysis of hypervariable regions of the genome. *BMC Plant Biology*, *4*(1), 11. https://doi.org/10.1186/1471-2229-4-11

Moyers, B. T., Morrell, P. L., & McKay, J. K. (2018). Genetic Costs of Domestication and Improvement. *Journal of Heredity*, *109*(2), 103–116. https://doi.org/10.1093/jhered/esx069

Multani, D. S., Briggs, S. P., Chamberlin, M. A., Blakeslee, J. J., Murphy, A. S., & Johal, G. S. (2003). Loss of an MDR Transporter in Compact Stalks of Maize *br2* and Sorghum *dw3* Mutants. *Science*, *302*(5642), 81–84. https://doi.org/10.1126/science.1086072

Muyt, A. D., Zhang, L., Piolot, T., Kleckner, N., Espagne, E., & Zickler, D. (2014). E3 ligase Hei10: A multifaceted structure-based signaling molecule with roles within and beyond meiosis. *Genes & Development*, *28*(10), 1111–1123. https://doi.org/10.1101/gad.240408.114

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*(7267), 1130–1134. https://doi.org/10.1038/nature08479

Nguyen, T. T., Le, N. T., & Sim, S.-C. (2021). Genome-wide association study and marker development for bacterial wilt resistance in tomato (*Solanum*

lycopersicum L.). *Scientia Horticulturae*, *289*, 110418. https://doi.org/10.1016/j.scienta.2021.110418

Oliveira, E. J., Pádua, J. G., Zucchi, M. I., Vencovsky, R., & Vieira, M. L. C. (2006). Origin, evolution and genome distribution of microsatellites. *Genetics and Molecular Biology*, *29*, 294–307. https://doi.org/10.1590/S1415-47572006000200018

Ong-Abdullah, M., Ordway, J. M., Jiang, N., Ooi, S.-E., Kok, S.-Y., Sarpan, N., Azimi, N., Hashim, A. T., Ishak, Z., Rosli, S. K., Malike, F. A., Bakar, N. A. A., Marjuni, M., Abdullah, N., Yaakub, Z., Amiruddin, M. D., Nookiah, R., Singh, R., Low, E.-T. L., … Martienssen, R. A. (2015). Loss of Karma transposon methylation underlies the mantled somaclonal variant of oil palm. *Nature*, *525*(7570), 533–537. https://doi.org/10.1038/nature15365

Otyama, P. I., Kulkarni, R., Chamberlin, K., Ozias-Akins, P., Chu, Y., Lincoln, L. M., MacDonald, G. E., Anglin, N. L., Dash, S., Bertioli, D. J., Fernández-Baca, D., Graham, M. A., Cannon, S. B., & Cannon, E. K. S. (2020). Genotypic Characterization of the U.S. Peanut Core Collection. *G3 Genes|Genomes|Genetics*, *10*(11), 4013–4026. https://doi.org/10.1534/g3.120.401306

Otyama, P. I., Wilkey, A., Kulkarni, R., Assefa, T., Chu, Y., Clevenger, J., O'Connor, D. J., Wright, G. C., Dezern, S. W., MacDonald, G. E., Anglin, N. L., Cannon, E. K. S., Ozias-Akins, P., & Cannon, S. B. (2019). Evaluation of linkage disequilibrium, population structure, and genetic diversity in the U.S. peanut mini core collection. *BMC Genomics*, *20*(1), 481. https://doi.org/10.1186/s12864-019-5824-9

Pandey, M. K., Agarwal, G., Kale, S. M., Clevenger, J., Nayak, S. N., Sriswathi, M., Chitikineni, A., Chavarro, C., Chen, X., Upadhyaya, H. D., Vishwakarma, M. K., Leal-Bertioli, S., Liang, X., Bertioli, D. J., Guo, B., Jackson, S. A., Ozias-Akins, P., & Varshney, R. K. (2017). Development and Evaluation of a High Density Genotyping 'Axiom_*Arachis*' Array with 58 K SNPs for Accelerating Genetics and Breeding in Groundnut. *Scientific Reports*, *7*(1), 40577. https://doi.org/10.1038/srep40577

Panizza, S., Mendoza, M. A., Berlinger, M., Huang, L., Nicolas, A., Shirahige, K., & Klein, F. (2011). Spo11-Accessory Proteins Link Double-Strand Break Sites to the Chromosome Axis in Early Meiotic Recombination. *Cell*, *146*(3), 372–383. https://doi.org/10.1016/j.cell.2011.07.003

Pratto Florencia, Brick Kevin, Khil Pavel, Smagulova Fatima, Petukhova Galina V., & Camerini-Otero R. Daniel. (2014). Recombination initiation maps of

individual human genomes. *Science*, *346*(6211), 1256442. https://doi.org/10.1126/science.1256442

Psakhye, I., & Jentsch, S. (2012). Protein Group Modification and Synergy in the SUMO Pathway as Exemplified in DNA Repair. *Cell*, *151*(4), 807–820. https://doi.org/10.1016/j.cell.2012.10.021

Purugganan, M. D., & Fuller, D. Q. (2009). The nature of selection during plant domestication. *Nature*, *457*(7231), 843–848. https://doi.org/10.1038/nature07895

Pyatnitskaya, A., Borde, V., & De Muyt, A. (2019). Crossing and zipping: Molecular duties of the ZMM proteins in meiosis. *Chromosoma*, *128*(3), 181–198. https://doi.org/10.1007/s00412-019-00714-8

Qiao, H., Chen, J. K., Reynolds, A., Höög, C., Paddy, M., & Hunter, N. (2012). Interplay between synaptonemal complex, homologous recombination, and centromeres during mammalian meiosis. *PLoS Genetics*, *8*(6), e1002790. https://doi.org/10.1371/journal.pgen.1002790

Ramakrishna, W., Emberton, J., Ogden, M., SanMiguel, P., & Bennetzen, J. L. (2002). Structural Analysis of the Maize *Rp1* Complex Reveals Numerous Sites and Unexpected Mechanisms of Local Rearrangement. *The Plant Cell*, *14*(12), 3213–3223. https://doi.org/10.1105/tpc.006338

Ramanna, M. S., & Jacobsen, E. (2003). Relevance of sexual polyploidization for crop improvement – A review. *Euphytica*, *133*(1), 3–8. https://doi.org/10.1023/A:1025600824483

Ramesh, R., Achari, G. A., & Gaitonde, S. (2014). Genetic diversity of *Ralstonia solanacearum* infecting solanaceous vegetables from India reveals the existence of unknown or newer sequevars of Phylotype I strains. *European Journal of Plant Pathology*, *140*(3), 543–562. https://doi.org/10.1007/s10658-014-0487-5

Reynolds, A., Qiao, H., Yang, Y., Chen, J. K., Jackson, N., Biswas, K., Holloway, J. K., Baudat, F., de Massy, B., Wang, J., Höög, C., Cohen, P. E., & Hunter, N. (2013). RNF212 is a dosage-sensitive regulator of crossing-over during mammalian meiosis. *Nature Genetics*, *45*(3), 269–278. https://doi.org/10.1038/ng.2541

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. *Proceedings of the National Academy of Sciences*, *112*(12), 3823–3828. https://doi.org/10.1073/pnas.1413864112

Rog, O., & Dernburg, A. F. (2013). Chromosome pairing and synapsis during C. elegans meiosis. *Current Opinion in Cell Biology*, *25*(3), 349–356. https://doi.org/10.1016/j.ceb.2013.03.003

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 Research*, *49*(18), 10431–10447. https://doi.org/10.1093/nar/gkab828

Ross, K. J., Fransz, P., Armstrong, S. J., Vizir, I., Mulligan, B., Franklin, F. C. H., & Jones, G. H. (1997). Cytological characterization of four meiotic mutants of Arabidopsis isolated from T-DNA-transformed lines. *Chromosome Research*, *5*(8), 551–559. https://doi.org/10.1023/A:1018497804129

Rosyara, U., Kishii, M., Payne, T., Sansaloni, C. P., Singh, R. P., Braun, H.-J., & Dreisigacker, S. (2019). Genetic Contribution of Synthetic Hexaploid Wheat to CIMMYT's Spring Bread Wheat Breeding Germplasm. *Scientific Reports*, *9*(1), 12355. https://doi.org/10.1038/s41598-019-47936-5

Sattler, M. C., Carvalho, C. R., & Clarindo, W. R. (2016). The polyploidy and its key role in plant breeding. *Planta*, *243*(2), 281–296. https://doi.org/10.1007/s00425-015-2450-x

Saze, H., Scheid, O. M., & Paszkowski, J. (2003). Maintenance of CpG methylation is essential for epigenetic inheritance during plant gametogenesis. *Nature Genetics*, *34*(1), 65–69. https://doi.org/10.1038/ng1138

Scherthan, H. (2001). A bouquet makes ends meet. *Nature Reviews. Molecular Cell Biology*, *2*(8), 621–627. https://doi.org/10.1038/35085086

Schmidt, C., Fransz, P., Rönspies, M., Dreissig, S., Fuchs, J., Heckmann, S., Houben, A., & Puchta, H. (2020). Changing local recombination patterns in *Arabidopsis* by CRISPR/Cas mediated chromosome engineering. *Nature Communications*, *11*(1), 4418. https://doi.org/10.1038/s41467-020-18277-z

Schwartz, E. K., & Heyer, W.-D. (2011). Processing of joint molecule intermediates by structure-selective endonucleases during homologous recombination in eukaryotes. *Chromosoma*, *120*(2), 109–127. https://doi.org/10.1007/s00412-010-0304-7

Séguéla-Arnaud, M., Crismani, W., Larchevêque, C., Mazel, J., Froger, N., Choinard, S., Lemhemdi, A., Macaisne, N., Van Leene, J., Gevaert, K., De Jaeger, G., Chelysheva, L., & Mercier, R. (2015). Multiple mechanisms limit meiotic crossovers: TOP3α and two BLM homologs antagonize crossovers in parallel to FANCM. *Proceedings of the National Academy of Sciences*, *112*(15), 4713–4718. https://doi.org/10.1073/pnas.1423107112

Sequeira-Mendes, J., Aragüez, I., Peiró, R., Mendez-Giraldez, R., Zhang, X., Jacobsen, S. E., Bastolla, U., & Gutierrez, C. (2014). The Functional Topography of the *Arabidopsis* Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States. *The Plant Cell*, 26(6), 2351–2366. https://doi.org/10.1105/tpc.114.124578

Sheridan, S. D., Yu, X., Roth, R., Heuser, J. E., Sehorn, M. G., Sung, P., Egelman, E. H., & Bishop, D. K. (2008). A comparative analysis of Dmc1 and Rad51 nucleoprotein filaments. *Nucleic Acids Research*, *36*(12), 4057–4066. https://doi.org/10.1093/nar/gkn352

Shin, I. S., Hsu, J.-C., Huang, S.-M., Chen, J.-R., Wang, J.-F., Hanson, P., & Schafleitner, R. (2020). Construction of a single nucleotide polymorphism marker based QTL map and validation of resistance loci to bacterial wilt caused by *Ralstonia solanacearum* species complex in tomato. *Euphytica*, *216*(3), 54. https://doi.org/10.1007/s10681-020-2576-1

Shull, G. H. (1908). The Composition of a Field of Maize. *Journal of Heredity*, *os-4*(1), 296–301. https://doi.org/10.1093/jhered/os-4.1.296

Sidhu, G. K., Fang, C., Olson, M. A., Falque, M., Martin, O. C., & Pawlowski, W. P. (2015). Recombination patterns in maize reveal limits to crossover homeostasis. *Proceedings of the National Academy of Sciences*, *112*(52), 15982–15987. https://doi.org/10.1073/pnas.1514265112

Singh, S., Gautam, R. K., Singh, D. R., Sharma, T. V. R. S., Sakthivel, K., & Roy, S. D. (2015). Genetic approaches for mitigating losses caused by bacterial wilt of tomato in tropical islands. *European Journal of Plant Pathology*, *143*(2), 205–221. https://doi.org/10.1007/s10658-015-0690-z

Smartt, J., Gregory, W. C., & Gregory, M. P. (1978). The genomes of Arachis hypogaea. 1. Cytogenetic studies of putative genome donors. *Euphytica*, *27*(3), 665–675. https://doi.org/10.1007/BF00023701

Song, M. J., Potter, B. I., Doyle, J. J., & Coate, J. E. (2020). Gene Balance Predicts Transcriptional Responses Immediately Following Ploidy Change in *Arabidopsis thaliana*. *The Plant Cell*, *32*(5), 1434–1448. https://doi.org/10.1105/tpc.19.00832

Soppe, W. J. J., Jasencakova, Z., Houben, A., Kakutani, T., Meister, A., Huang, M. S., Jacobsen, S. E., Schubert, I., & Fransz, P. F. (2002). DNA methylation controls histone H3 lysine 9 methylation and heterochromatin assembly in *Arabidopsis*. *The EMBO Journal*, *21*(23), 6549–6559. https://doi.org/10.1093/emboj/cdf657

Stacey, N. J., Kuromori, T., Azumi, Y., Roberts, G., Breuer, C., Wada, T., Maxwell, A., Roberts, K., & Sugimoto-Shirasu, K. (2006). *Arabidopsis* SPO11-2

functions with SPO11-1 in meiotic recombination. *The Plant Journal: For Cell and Molecular Biology*, *48*(2), 206–216. https://doi.org/10.1111/j.1365-313X.2006.02867.x

Stadler, L. J. (1928). Mutations in Barley Induced by X-Rays and Radium. *Science*, *68*(1756), 186–187. https://doi.org/10.1126/science.68.1756.186

Stalker, H. T. (1991). A New Species in Section Arachis of Peanuts with a D Genome. *American Journal of Botany*, *78*(5), 630–637. https://doi.org/10.2307/2445084

Storlazzi, A., Gargano, S., Ruprich-Robert, G., Falque, M., David, M., Kleckner, N., & Zickler, D. (2010). Recombination proteins mediate meiotic spatial chromosome organization and pairing. *Cell*, *141*(1), 94–106. https://doi.org/10.1016/j.cell.2010.02.041

Storlazzi, A., Xu, L., Schwacha, A., & Kleckner, N. (1996). Synaptonemal complex (SC) component Zip1 plays a role in meiotic recombination independent of SC polymerization along the chromosomes. *Proceedings of the National Academy of Sciences*, *93*(17), 9043–9048.

Stroud, H., Do, T., Du, J., Zhong, X., Feng, S., Johnson, L., Patel, D. J., & Jacobsen, S. E. (2014). Non-CG methylation patterns shape the epigenetic landscape in *Arabidopsis*. *Nature Structural & Molecular Biology*, *21*(1), 64–72. https://doi.org/10.1038/nsmb.2735

Stroud, H., Greenberg, M. V. C., Feng, S., Bernatavichute, Y. V., & Jacobsen, S. E. (2013). Comprehensive Analysis of Silencing Mutants Reveals Complex Regulation of the *Arabidopsis* Methylome. *Cell*, *152*(1), 352–364. https://doi.org/10.1016/j.cell.2012.10.054

Sturtevant, A. H. (1913) The linear arrangement of six sex-linked factors in Drosophila, as shown by their mode of association. *Journal of Experimental Zoology*, 14, 43–59.

Sudupak, M. A., Bennetzen, J. L., & Hulbert, S. H. (1993). Unequal exchange and meiotic instability of disease-resistance genes in the Rp1 region of maize. *Genetics*, *133*(1), 119–125. https://doi.org/10.1093/genetics/133.1.119

Takeo, S., Lake, C. M., Morais-de-Sá, E., Sunkel, C. E., & Hawley, R. S. (2011). Synaptonemal Complex-Dependent Centromeric Clustering and the Initiation of Synapsis in *Drosophila* Oocytes. *Current Biology*, *21*(21), 1845–1851. https://doi.org/10.1016/j.cub.2011.09.044

Tallury, S. P., Hilu, K. W., Milla, S. R., Friend, S. A., Alsaghir, M., Stalker, H. T., & Quandt, D. (2005). Genomic affinities in *Arachis* section *Arachis* (Fabaceae): Molecular and cytogenetic evidence. *Theoretical and Applied Genetics*, *111*(7), 1229–1237. https://doi.org/10.1007/s00122-005-0017-0

Tang, R., Zhuang, W., Gao, G., He, L., Han, Z., Shan, S., Jiang, J., & Li, Y. (2008). Phylogenetic Relationships in Genus *Arachis* Based on SSR and AFLP Markers. *Agricultural Sciences in China*, *7*(4), 405–414. https://doi.org/10.1016/S1671-2927(08)60083-8

Tanneti, N. S., Landy, K., Joyce, E. F., & McKim, K. S. (2011). A pathway for synapsis initiation during zygotene in *Drosophila* oocytes. *Current Biology: CB*, *21*(21), 1852–1857. https://doi.org/10.1016/j.cub.2011.10.005

Thomas, C. M., Jones, D. A., Parniske, M., Harrison, K., Balint-Kurti, P. J., Hatzixanthis, K., & Jones, J. D. (1997). Characterization of the tomato *Cf-4* gene for resistance to *Cladosporium fulvum* identifies sequences that determine recognitional specificity in Cf-4 and Cf-9. *The Plant Cell*, *9*(12), 2209–2224. https://doi.org/10.1105/tpc.9.12.2209

Tischfield, S. E., & Keeney, S. (2012). Scale matters. *Cell Cycle*, *11*(8), 1496–1503. https://doi.org/10.4161/cc.19733

Türkösi, E., Cseh, A., Darkó, É., & Molnár-Láng, M. (2016). Addition of Manas barley chromosome arms to the hexaploid wheat genome. *BMC Genetics*, *17*, 87. https://doi.org/10.1186/s12863-016-0393-2

Türkösi, E., Darko, E., Rakszegi, M., Molnár, I., Molnár-Láng, M., & Cseh, A. (2018). Development of a new 7BS.7HL winter wheat-winter barley Robertsonian translocation line conferring increased salt tolerance and (1,3;1,4)-β-D-glucan content. *PLOS ONE*, *13*(11), e0206248. https://doi.org/10.1371/journal.pone.0206248

Underwood, C. J., Choi, K., Lambing, C., Zhao, X., Serra, H., Borges, F., Simorowski, J., Ernst, E., Jacob, Y., Henderson, I. R., & Martienssen, R. A. (2018). Epigenetic activation of meiotic recombination near *Arabidopsis thaliana* centromeres via loss of H3K9me2 and non-CG DNA methylation. *Genome Research*, *28*(4), 519–531. PubMed. https://doi.org/10.1101/gr.227116.117

Upadhyaya, H. D., Bramel, P. J., Ortiz, R., & Singh, S. (2002). Developing a Mini Core of Peanut for Utilization of Genetic Resources. *Crop Science*, *42*(6), 2150–2156. https://doi.org/10.2135/cropsci2002.2150

Upadhyaya, H. D., Ortiz, R., Bramel, P. J., & Singh, S. (2003). Development of a groundnut core collection using taxonomical, geographical and morphological descriptors. *Genetic Resources and Crop Evolution*, *50*(2), 139–148. https://doi.org/10.1023/A:1022945715628

Van de Peer, Y., Ashman, T.-L., Soltis, P. S., & Soltis, D. E. (2021). Polyploidy: An evolutionary and ecological force in stressful times. *The Plant Cell*, *33*(1), 11–26. https://doi.org/10.1093/plcell/koaa015

Van de Peer, Y., Maere, S., & Meyer, A. (2009). The evolutionary significance of ancient genome duplications. *Nature Reviews Genetics*, *10*(10), 725–732. https://doi.org/10.1038/nrg2600

Vongs, A., Kakutani, T., Martienssen, R. A., & Richards, E. J. (1993). *Arabidopsis thaliana* DNA Methylation Mutants. *Science*, *260*(5116), 1926–1928. https://doi.org/10.1126/science.8316832

Wang, C. T., Wang, X. Z., Tang, Y. Y., Chen, D. X., Cui, F. G., Zhang, J. C., & Yu, S. L. (2011). Phylogeny of *Arachis* based on internal transcribed spacer sequences. *Genetic Resources and Crop Evolution*, *2*(58), 311–319. https://doi.org/10.1007/s10722-010-9576-2

Wang, J.-F., Hanson, P., & Barnes, J. A. (1998). Worldwide Evaluation of an International Set of Resistance Sources to Bacterial Wilt in Tomato. In P. Prior, C. Allen, & J. Elphinstone (Eds.), *Bacterial Wilt Disease: Molecular and Ecological Aspects* (pp. 269–275). Springer. https://doi.org/10.1007/978-3-662-03592-4_39

Wang, J.-F., Ho, F.-I., Truong, H. T. H., Huang, S.-M., Balatero, C. H., Dittapongpitch, V., & Hidayati, N. (2013). Identification of major QTLs associated with stable resistance of tomato cultivar 'Hawaii 7996' to *Ralstonia solanacearum*. *Euphytica*, *190*(2), 241–252. https://doi.org/10.1007/s10681-012-0830-x

Wang, K., Wang, M., Tang, D., Shen, Y., Miao, C., Hu, Q., Lu, T., & Cheng, Z. (2012). The Role of Rice HEI10 in the Formation of Meiotic Crossovers. *PLOS Genetics*, *8*(7), e1002809. https://doi.org/10.1371/journal.pgen.1002809

Wang, M., Shilo, S., Zhou, A., Zelkowski, M., Olson, M. A., Azuri, I., Shoshani-Hechel, N., Melamed-Bessudo, C., Marand, A. P., Jiang, J., Schnable, J. C., Underwood, C. J., Henderson, I. R., Sun, Q., Pillardy, J., Kianian, P. M. A., Kianian, S. F., Chen, C., Levy, A. A., & Pawlowski, W. P. (2022). Machine learning reveals conserved chromatin patterns determining meiotic recombination sites in plants. *BioRxiv*, 2022.07.11.499557. https://doi.org/10.1101/2022.07.11.499557

Wang, M. L., Sukumaran, S., Barkley, N. A., Chen, Z., Chen, C. Y., Guo, B., Pittman, R. N., Stalker, H. T., Holbrook, C. C., Pederson, G. A., & Yu, J. (2011). Population structure and marker–trait association analysis of the US peanut (*Arachis hypogaea* L.) mini-core collection. *Theoretical and Applied Genetics*, *123*(8), 1307–1317. https://doi.org/10.1007/s00122-011-1668-7

Wang, M., Wang, P., Lin, M., Ye, Z., Li, G., Tu, L., Shen, C., Li, J., Yang, Q., & Zhang, X. (2018). Evolutionary dynamics of 3D genome architecture following

polyploidization in cotton. *Nature Plants*, *4*(2), 90–97. https://doi.org/10.1038/s41477-017-0096-3

Wang, Y., & Copenhaver, G. P. (2018). Meiotic Recombination: Mixing It Up in Plants. *Annual Review of Plant Biology*, *69*(1), 577–609. https://doi.org/10.1146/annurev-arplant-042817-040431

Wicker, E., Lefeuvre, P., de Cambiaire, J.-C., Lemaire, C., Poussier, S., & Prior, P. (2012). Contrasting recombination patterns and demographic histories of the plant pathogen *Ralstonia solanacearum* inferred from MLSA. *The ISME Journal*, *6*(5), 961–974. https://doi.org/10.1038/ismej.2011.160

Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P., & Schulman, A. H. (2007). A unified classification system for eukaryotic transposable elements. *Nature Reviews Genetics*, *8*(12), 973–982. https://doi.org/10.1038/nrg2165

Wijnker, E., & de Jong, H. (2008). Managing meiotic recombination in plant breeding. *Trends in Plant Science*, *13*(12), 640–646. https://doi.org/10.1016/j.tplants.2008.09.004

Wright, S. I., Ness, R. W., Foxe, J. P., & Barrett, S. C. H. (2008). Genomic Consequences of Outcrossing and Selfing in Plants. *International Journal of Plant Sciences*, *169*(1), 105–118. https://doi.org/10.1086/523366

Xu, Y. (2010). *Molecular Plant Breeding*. CABI Publishing.

Xue, M., Wang, J., Jiang, L., Wang, M., Wolfe, S., Pawlowski, W. P., Wang, Y., & He, Y. (2018). The Number of Meiotic Double-Strand Breaks Influences Crossover Distribution in *Arabidopsis*. *The Plant Cell*, *30*(10), 2628–2638. https://doi.org/10.1105/tpc.18.00531

Yancey-Wrona, J. E., & Camerini-Otero, R. D. (1995). The search for DNA homology does not limit stable homologous pairing promoted by RecA protein. *Current Biology*, *5*(10), 1149–1158. https://doi.org/10.1016/S0960-9822(95)00231-4

Yelina, N. E., Choi, K., Chelysheva, L., Macaulay, M., de Snoo, B., Wijnker, E., Miller, N., Drouaud, J., Grelon, M., Copenhaver, G. P., Mezard, C., Kelly, K. A., & Henderson, I. R. (2012). Epigenetic remodeling of meiotic crossover frequency in *Arabidopsis thaliana* DNA methyltransferase mutants. *PLoS Genetics*, *8*(8), e1002844. https://doi.org/10.1371/journal.pgen.1002844

Yokoo, R., Zawadzki, K. A., Nabeshima, K., Drake, M., Arur, S., & Villeneuve, A. M. (2012). COSA-1 Reveals Robust Homeostasis and Separable Licensing and Reinforcement Steps Governing Meiotic Crossovers. *Cell*, *149*(1), 75–87. https://doi.org/10.1016/j.cell.2012.01.052

Yu, H., Wang, M., Tang, D., Wang, K., Chen, F., Gong, Z., Gu, M., & Cheng, Z. (2010). OsSPO11-1 is essential for both homologous chromosome pairing and crossover formation in rice. *Chromosoma*, *119*(6), 625–636. https://doi.org/10.1007/s00412-010-0284-7

Zakharyevich, K., Tang, S., Ma, Y., & Hunter, N. (2012). Delineation of joint molecule resolution pathways in meiosis identifies a crossover-specific resolvase. *Cell*, *149*(2), 334–347. https://doi.org/10.1016/j.cell.2012.03.023

Zapata Luis, Ding Jia, Willing Eva-Maria, Hartwig Benjamin, Bezdan Daniela, Jiao Wen-Biao, Patel Vipul, Velikkakam James Geo, Koornneef Maarten, Ossowski Stephan, & Schneeberger Korbinian. (2016). Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. *Proceedings of the National Academy of Sciences*, *113*(28), E4052–E4060. https://doi.org/10.1073/pnas.1607532113

Zemach, A., Kim, M. Y., Hsieh, P.-H., Coleman-Derr, D., Eshed-Williams, L., Thao, K., Harmer, S. L., & Zilberman, D. (2013). The *Arabidopsis* Nucleosome Remodeler DDM1 Allows DNA Methyltransferases to Access H1-Containing Heterochromatin. *Cell*, *153*(1), 193–205. https://doi.org/10.1016/j.cell.2013.02.033

Zhang, L., Espagne, E., de Muyt, A., Zickler, D., & Kleckner, N. E. (2014). Interference-mediated synaptonemal complex formation with embedded crossover designation. *Proceedings of the National Academy of Sciences*, *111*(47), E5059–E5068. https://doi.org/10.1073/pnas.1416411111

Zhang, L., Wang, S., Yin, S., Hong, S., Kim, K. P., & Kleckner, N. (2014). Topoisomerase II mediates meiotic crossover interference. *Nature*, *511*(7511), 551–556. https://doi.org/10.1038/nature13442

Zhang, X., Yazaki, J., Sundaresan, A., Cokus, S., Chan, S. W.-L., Chen, H., Henderson, I. R., Shinn, P., Pellegrini, M., Jacobsen, S. E., & Ecker, J. R. (2006). Genome-wide high-resolution mapping and functional analysis of DNA methylation in *Arabidopsis*. *Cell*, *126*(6), 1189–1201. https://doi.org/10.1016/j.cell.2006.08.003

Zhuang, W., Chen, H., Yang, M., Wang, J., Pandey, M. K., Zhang, C., Chang, W.-C., Zhang, L., Zhang, X., & Tang, R. (2019). The genome of cultivated peanut provides insight into legume karyotypes, polyploid evolution and crop domestication. *Nature Genetics*, 51(5), 865–876.

Zickler, D. (2006). From early homologue recognition to synaptonemal complex formation. *Chromosoma*, *115*(3), 158–174. https://doi.org/10.1007/s00412-006-0048-6

Zickler, D., & Kleckner, N. (2015). Recombination, Pairing, and Synapsis of Homologs during Meiosis. *Cold Spring Harbor Perspectives in Biology*, *7*(6), a016626. https://doi.org/10.1101/cshperspect.a016626

Ziolkowski, P. A., Berchowitz, L. E., Lambing, C., Yelina, N. E., Zhao, X., Kelly, K. A., Choi, K., Ziolkowska, L., June, V., Sanchez-Moran, E., Franklin, C., Copenhaver, G. P., & Henderson, I. R. (2015). Juxtaposition of heterozygous and homozygous regions causes reciprocal crossover remodelling via interference during *Arabidopsis* meiosis. *ELife*, *4*, e03708. https://doi.org/10.7554/eLife.03708

Zohoungbogbo, H., Quenum, A., Honfoga, J., Chen, J.-R., Achigan-Dako, E., Kenyon, L., & Hanson, P. (2021). Evaluation of Resistance Sources of Tomato (*Solanum lycopersicum* L.) to Phylotype I Strains of Ralstonia solanacearum Species Complex in Benin. *Agronomy*, 11(8), 1513.

Synthèse en français

Introduction

La diversité génétique est définie comme étant le degré de variation des séquences d'ADN entre les individus (chromosomes) d'une espèce (population). En évolution, la diversité génétique est façonnée par les mutations spontanées, la dérive génétique et la sélection, changements moléculaires héritables qui peuvent rendre les populations ou les espèces mieux adaptées aux changements de leur environnement. En agriculture, notamment dans le domaine de la sélection végétale, l'exploitation de la diversité génétique est cruciale pour créer de nouvelles variétés ayant un rendement plus élevé et des caractéristiques améliorées. Selon la définition de Poehlman et Sleper (1995), la sélection végétale est "l'art et la science de l'amélioration de l'hérédité des plantes au profit de l'humanité". En d'autres termes, les sélectionneurs doivent exploiter leur boîte à outils pour créer, évaluer et manipuler la diversité génétique afin de mieux concevoir et développer de nouvelles variétés pour répondre à la demande.

Tout d'abord, la mesure de la diversité génétique dans une collection (“germplasm” en anglais, “germoplasme” en français) est l'un des processus importants permettant aux sélectionneurs de comprendre leur matériel de sélection en vu de leurs objectifs d’amélioration. L'évaluation de la diversité génétique peut s'appuyer sur des marqueurs morphologiques, biochimiques, cytologiques et moléculaires. Grâce aux progrès des outils moléculaires, le marquage moléculaire est devenu un outil courant et efficace pour évaluer la diversité génétique. En outre, l'identification de gènes candidats pour des phénotypes cibles de caractères souhaitables peut faciliter les programmes de sélection. Les sélectionneurs ont souvent recours à l'hybridation pour accumuler des phénotypes cibles et sélectionner des descendances présentant des caractéristiques souhaitables. Par exemple, lorsque les

sélectionneurs mettent en œuvre des programmes de sélection de la résistance, ils utilisent souvent le pyramidage des gènes pour combiner plusieurs gènes de résistance en un seul génotype. Ainsi, les gènes candidats identifiés par différentes approches peuvent effectivement aider les sélectionneurs à mieux concevoir les programmes de sélection. Enfin, dans tout programme de sélection, les sélectionneurs choisissent les descendances en fonction du remaniement du génome provoqué par la recombinaison méiotique. Ce phénomène spécifique qui se produit au cours de la méiose génère différentes combinaisons d'allèles qui peuvent être utilisées pour la sélection. Si l'on parvient à élucider le mécanisme sous-jacent qui contrôle le nombre et la distribution des crossovers, les sélectionneurs devraient bénéficier de ces connaissances pour mieux manipuler la variation génétique de leur matériel de sélection et sélectionner des idéotypes avec plus d'efficacité.

Dans mon travail de thèse, j'ai commencé par modéliser la recombinaison méiotique chez *Arabidopsis thaliana* afin d'établir un modèle quantitatif capable de prédire la structuration à petite échelle des paysages de recombinaison. Ensuite, j'ai participé à un projet d'évaluation de la diversité génétique de l'arachide à Taïwan. Sur la base des données RAD-seq de 31 accessions, j'ai non seulement fait l'analyse de la diversité mais j'ai aussi sélectionné des SNPs pour obtenir un ensemble de marqueurs KASP qui peuvent être utilisés pour l'amélioration génétique. Enfin, j'ai utilisé les données de séquençage du génome entier de plusieurs lignées de tomates résistantes et sensibles pour identifier un gène de résistance candidat contre la maladie du flétrissement bactérien, et ce résultat et cette approche pourront être appliqués dans les futurs travaux de sélection. Ces trois projets commencent par des aspects différents mais partagent l'objectif de développer des outils pour mieux évaluer et manipuler la diversité génétique dans les futurs programmes de sélection végétale.

Projet 1. Modélisation quantitative du paysage de recombinaison chez *Arabidopsis thaliana*

La méiose est un type de division cellulaire qui fait passer l'organisme de l'état diploïde à l'état haploïde. La formation de crossing-overs (CO) au cours de ce processus entraîne le remaniement du génome qui crée différentes combinaisons d'allèles. Les CO jouent donc un rôle important dans la génération de la variation phénotypique et génomique pour les programmes de sélection, animale comme végétale. Les régions péricentromériques, qui présentent une forte densité d'éléments transposables, n'ont qu'un nombre limité de CO (Choulet et al., 2014). Cependant, dans les plantes cultivées, ces régions contiennent un nombre considérable de gènes, de sorte que le recrutement de CO dans ces régions pourraient faciliter les études génétiques comme le clonage positionnel et la sélection de nouvelles combinaisons alléliques pour la sélection.

Suite aux cassures double brin, la formation de CO est une des issues possibles pour la réparation de l'ADN (Mercier et al., 2015). La formation de CO présente deux caractéristiques. Premièrement, le nombre de CO est strictement régulé, même si la taille des chromosomes varie largement entre les différents organismes (Fernandes et al., 2018). Habituellement, un bivalent a obligatoirement au moins un CO (Jones & Franklin, 2006; Zickler & Kleckner, 2016). En outre, la distribution des CO est très hétérogène le long des chromosomes. Par exemple, les événements CO ont été détectés dans seulement 13% du chromosome 3B du blé (Choulet et al., 2014). Ainsi, une telle hétérogénéité de l'occurrence de CO le long des chromosomes suggère que des déterminants de la formation du CO existent pour créer des régions chaudes et froides de CO à des échelles plus fines. En général, l'occurrence des CO est associée aux régions des promoteurs et de chromatine ouverte, et les CO sont fortement supprimés dans les régions hétérochromatiques telles que les centromères. Par exemple, les CO d'Arabidopsis se localisent préférentiellement dans les régions subtélomériques et péricentromériques

(appauvries en nucléosomes) mais pas dans les centromères. De plus, les CO d'Arabidopsis provenant du croisement Col/Ler sont associés aux répétitions (A/T), CTT/GAA, CT et CCN selon une analyse à échelle fine, et les points chauds des CO du riz sont enrichis en répétitions de séquences simples et en classes de transposons d'ADN, notamment PIF, Harbinger et Stowaway, mais sont dépourvus de rétrotransposons. Ces résultats suggèrent que la formation de CO peut-être associée à divers types de motifs d'ADN et d'éléments transposables (Marand et al., 2019 ; Rowan et al., 2019).

Pour comprendre comment les caractéristiques génomiques et épigénomiques influencent les paysages de recombinaison à des échelles fines, nous avons choisi Arabidopsis thaliana qui a le plus grand jeu de données de CO (17 077 CO de 2 182 plantes, publiés par Rowan et al. (2019)) ainsi que d'autres données concernant des caractéristiques génomiques et épigénomiques pour établir un modèle quantitatif. Ici, nous avons d'abord comparé une caractéristique génomique ou épigénomique avec le taux de recombinaison, mais trouvons que la dépendance à une caractéristique est généralement non monotone. En outre, certaines caractéristiques fortement corrélées avec d'autres ont montré un modèle de corrélation similaire entre elles et le taux de recombinaison. Ainsi, un modèle basé naïvement sur la combinaison de ces caractéristiques génomiques et épigénomiques entraînera une complexité combinatoire ingérable. Pour surmonter cette difficulté, nous avons combiné un ensemble de données avec 9 états chromatiniens définis par 16 caractéristiques génomiques et épigénomiques, avec les informations de variation structurelle entre Col et Ler, conduisant à 10 classes discrètes (états) comme points de départ de notre modélisation qui segmente l'ensemble du génome d'Arabidopsis par ces classes (Sequeira-Mendes et al., 2014). En utilisant la proportion de ces 10 états le long du génome, nous avons pu leur associer un taux de recombinaison puis prédire de manière assez fiable les profils de recombinaison autour des gènes et dans les régions intergéniques. En outre,

ce modèle simple a particulièrement bien prédit les régions flanquantes des gènes avec des taux de recombinaison accrus, ce qui est cohérent avec l'expérience (Choi et al., 2013; Marand et al., 2017; Kianian et al., 2018).

Sur la base d'une analyse plus approfondie, nous avons découvert de nouvelles tendances. Premièrement, le taux de recombinaison est supprimé dans les régions intergéniques de taille inférieure à environ 1,5 kb. En outre, les régions présentant un faible niveau de divergence de séquence, représenté par la densité de polymorphismes de nucléotides simples (SNP), ont un plus faible taux de recombinaison. Dans l'ensemble, nous avons intégré 10 états, la taille des régions intergéniques et la densité de SNP pour construire un modèle quantitatif. Avec un bon pouvoir prédictif et une très faible complexité, ce modèle permet de reproduire une grande partie de la variation du taux de recombinaison chez A. thaliana. Notre résultat montre l'importance de différents effets contextuels modulant le taux de CO aux petites échelles génomiques.

Projet 2. Évaluation de la diversité génétique dans le germoplasme de l'arachide taïwanaise

L'arachide cultivée (Arachis hypogaea L.) est une légumineuse allotétraploïde (2n=4x=40) qui est une culture oléagineuse d'importance mondiale. L'arachide cultivée appartient au genre Arachis, originaire d'Amérique du Sud. Il est connu que l'arachide cultivée a une diversité génétique limitée résultant d'importants goulots d'étranglement génétiques dus à la polyploïdisation (Burow et al., 2001; Foncéka et al., 2009). Ainsi, la conservation du matériel génétique d'Arachis et la compréhension de sa diversité génétique sont toutes deux essentielles pour le travail de sélection de l'arachide. La conservation de la variation génétique des espèces cultivées et sauvages d'arachide a été assurée par un certain nombre de collections ex situ dans le monde. Par exemple, les collections les plus représentatives sont celles de l'Institut international de recherche sur les cultures des zones tropicales semi-arides

(ICRISAT), du ministère de l'Agriculture des États-Unis (USDA), et de l'Institut de recherche sur les cultures oléagineuses de l'Académie chinoise des sciences agricoles (OCRI-CAAS), qui comptent respectivement plus de 15 000, 9 000 et 8 000 entrées (Barkley et al., 2016). D'autre part, l'étude de la diversité génétique du germoplasme par des marqueurs moléculaires est le moyen le plus efficace et le plus courant à ce jour (Desmae et al., 2019). Récemment, les progrès technologiques ont facilité les projets de séquençage du génome d'Arachis hypogaea L. et de ses deux ancêtres diploïdes, A. duranensis (AA) et A. ipaensis (BB), puis ont accéléré le développement de marqueurs SNP à haut débit, tels que le génotypage par séquençage (GBS), pour la sélection moléculaire de l'arachide (Bertioli et al., 2016 ; Bertioli et al., 2019 ; Zhuang et al., 2019).

Les programmes nationaux de sélection de l'arachide à Taïwan ont commencé depuis la fin des années 1950. Cependant, la plupart des programmes n’ont bénéficié d'aucune utilisation de marqueurs moléculaires, ce qui signifie que la sélection parentale dépendait principalement des informations du pedigree pour connaître leur relation génétique et que l'évaluation des populations de sélection était uniquement basée sur les traits morphologiques. Ainsi, le développement d'outils moléculaires pour étudier le matériel génétique actuel de l'arachide améliorerait l'efficacité des futurs programmes de sélection de l'arachide à Taïwan. Dans ce projet, l'approche associé aux sites de restriction (RAD) de l’ADN a été utilisée pour séquencer 31 accessions taïwanaises comprenant des cultivars élites locaux et étrangers afin d'identifier leur diversité génétique sous-jacente.

En exploitant les données de l'approche RAD, j'ai identifié 3474 SNPs qui ont été utilisés pour l'analyse de la diversité génétique. Mes mesures de cette diversité sont basées sur la valeur du contenu d'information polymorphe (PIC), l'hétérozygotie attendue (He), et la distance génétique. Tout d'abord, je me suis concentré sur les sous-ensembles associés à l'origine du

germoplasme, conduisant à un sous-ensemble “global” contenant 17 accessions "introduites" et un sous-ensemble “local” contenant 14 accessions taïwanaises. Mes résultats montrent que trois mesures de diversité sont plus grandes dans le sous-ensemble global que dans le sous-ensemble local, ce qui indique que le sous-ensemble global est plus diversifié génétiquement que le sous-ensemble local. En termes de variétés botaniques, les 31 accessions dans ce germoplasme peuvent être regroupées en trois variétés botaniques, subsp. fastigiata var. fastigiata (type Valencia, n = 11), subsp. fastigiata var. vulgaris (type Spanish, n = 14) et subsp. hypogaea var. hypogaea (types Virginia/Runner, n = 7). Le résultat a révélé que les accessions du type Valencia avaient une moyenne de He (0,18), PIC (0,15) et une distance génétique (0,15) plus grandes que celles du type Spanish (He = 0,13, PIC = 0,11, distance = 0,11) et du type Virginia/Runner (He = 0,12, PIC = 0,10, distance = 0,11). De manière intéressante, nous avons constaté que le groupe de type Valencia a plus d'accessions introduites d'Amérique du Sud, lieu proche de l'origine de l'arachide cultivée, que les deux autres groupes (Bertioli et al., 2011). Il n'est donc pas surprenant que les accessions de type Valencia aient une plus grande diversité génétique que les accessions de type espagnol et de type Virginia/Runner qui composent la plupart des cultivars à Taïwan.

Sur la base de la distance génétique par paire entre les 31 accessions, j'ai étudié plus en détail les relations génétiques à l'aide d’un dendrogramme et de l'ACP. Dans l'ensemble, ces 31 accessions peuvent être regroupées en trois groupes selon trois variétés botaniques. Cependant, le résultat du dendrogramme a également montré que les cultivars du germoplasme de type espagnol à Taïwan sont fortement corrélés entre eux, ce qui suggère que ces cultivars locaux souffrent probablement d'une vulnérabilité génétique car la diversité génétique du matériel de sélection utilisé dans les programmes de sélection est peut-être trop étroite.

Afin d'identifier plus efficacement la structure de la population du germoplasme actuel, mes collaborateurs et moi-même avons développé 14 marqueurs KASP en utilisant les données RAD-seq de 31 accessions, et avons validé ces marqueurs à l'aide de 282 accessions d'arachide du centre de conservation du germoplasme de l'Institut de recherche agricole de Taïwan (TARI). En outre, nous avons également considéré les données phénotypiques de 8 traits quantitatifs agronomiques comme un outil alternatif pour évaluer la structure de la population. Les résultats par ACP ont montré que ces marqueurs KASP séparent clairement les 282 génotypes en 3 groupes correspondant aux trois variétés botaniques. En revanche, l'ACP basée sur les données phénotypiques n'a que grossièrement regroupé les mêmes 282 accessions en deux sous-espèces d'arachide, ce qui suggère que les marqueurs moléculaires sont plus stables et plus efficaces que les données phénotypiques pour identifier les structures de population des collections de germoplasme.

Projet 3. Identification d'un gène contribuant à la résistance au flétrissement bactérien chez la tomate

La tomate (Solanum lycopersicum L.) est l'une des cultures légumières les plus importantes du point de vue économique dans le monde, et sa production mondiale n'a cessé d'augmenter depuis les années 1970. Le rendement et la qualité de la tomate peuvent être endommagés par diverses maladies causées par des pathogènes bactériens, fongiques ou même viraux, et le flétrissement bactérien (BW) est l'une des maladies végétales graves qui entraîne une perte de rendement considérable chez la tomate (Hartman et al., 1991 ; Karumannil et al., 2008). Le flétrissement bactérien est causé par Ralstonia solanacearum qui possède une large gamme d'hôtes et une grande adaptabilité aux températures élevées et aux sols humides. Le complexe d'espèces R. solanacearum (RSSC) contient diverses souches qui en font l'une des bactéries phytopathogènes les plus nuisibles au monde, en particulier dans les régions tropicales, subtropicales et tempérées chaudes

(Denny, 2006 ; Genin, 2010). Pour diminuer la perte de rendement causée par la BW, le développement de cultivars résistants est l'approche la plus efficace (Hanson et al., 2016).

Les sources de résistance se trouvent dans les tomates cultivées et les espèces sauvages, telles que Lycopersicon esculentum var. cerasiforme, Lycopersicon pimpinellifolium et Lycopersicon peruvianum. Wang et al (1998) ont mené une expérience pour évaluer la résistance au BW de 35 lignées de tomates dans 11 champs situés dans 11 pays, et H7996 présente la résistance au BW la plus stable dans plusieurs endroits. De plus, Wang et al. (2013) ont identifié deux QTLs majeurs de H7996, Bwr-6 et Bwr-12, mais les lignées de sélection avancées avec ces deux QTLs à l'état homozygote, développées par Worldveg, n'atteignent pas le même niveau de résistance que H7996 contre le BW, ce qui suggère que H7996 a d'autres gènes de résistance qui restent à être identifiés (Zohoungbogbo et al., 2021). Ainsi, l'objectif de ce projet est d'identifier d'autres QTLs de résistance en plus de Bwr-6 et Bwr-12 qui confèrent la résistance au BW.

Après le test d'inoculation utilisant deux souches virulentes, six lignées de tomates résistantes et neuf lignées sensibles ont été sélectionnées pour du séquençage en génome entier. Après l'alignement des séquences et l'appel de variants, 883 682 SNP et 222 565 InDels ont été identifiés dans les 15 accessions. Ensuite, j'ai conçu un pipeline d'analyse interne qui compare chacune des six accessions résistantes avec les neuf accessions sensibles, et je n'ai conservé une variation dans une accession résistante que si elle était absente de toutes les accessions sensibles. Sur la base de cette analyse, j'ai exclu 92% des variants des 6 accessions résistantes. En ce qui concerne les polymorphismes spécifiques aux lignées résistantes, ils n'ont pas seulement été identifiés dans des régions comprenant Bwr-6 et Bwr-12 qui peuvent justifier une caractérisation moléculaire plus poussée, mais également sur d'autres chromosomes. Ensuite, j'ai effectué une prédiction

pour étudier comment la fonction des gènes codant pour les protéines est influencée par les polymorphismes spécifiquement trouvés dans les 6 lignées résistantes. Au final, il y avait 385 gènes identifiés à partir des six lignées résistantes qui devraient avoir des polymorphismes à fort impact, et la plupart de ces polymorphismes étaient situés sur les chromosomes 2 et 4.

Pour valider ces variants candidats, des marqueurs moléculaires ont été développés et testés dans deux populations F2 (CLN4397 et CLN4398), dérivées de H7996 et de chacune des deux lignées avancées, avec toutes les plantes contenant des locus Bwr-6 et Bwr-12 homozygotes. Parmi les marqueurs testés, mes collaborateurs ont trouvé que le marqueur Bwr3.2dCAPS situé sur le chromosome 3 était significativement corrélé à la résistance dans la population CLN4398. Ce marqueur correspond à un polymorphisme produit par la délétion de la 102e arginine, cette mutation conduisant à un changement de cadre de lecture dans le gène Asc, et donc ayant un fort impact. De plus, ce gène Asc (Solyc03g114600.4.1) a été identifié comme contribuant à la résistance à un pathogène fongique, Alternaria alternata f. sp. lycopersici (AAL). De façon intéressante, il a été démontré auparavant que H7996 a un QTL mineur, Bwr-3, pour la résistance à l'AAL sur le chromosome 3, et ce QTL, contenant Bwr3.2dCAPS, couvre une grande région dans l'extrémité distale du chromosome 3 (Wang et al., 2013 ; Carmeille et al., 2006 ; Hai, 2007). Cependant, ce gène n’explique pas complètement le niveau de résistance des plantes, donc des études supplémentaires doivent être menées pour identifier d’autres gènes conférant la résistance au BW.

Perspectives

Dans le premier projet, même si mon modèle peut reproduire une grande partie de la variation de la recombinaison expérimentale, il y a toujours des variations qui ne peuvent pas être prédites par le modèle, ce qui suggère que ce modèle a certaines limites. Premièrement, les 9 états de la chromatine ont

été construits en utilisant des cellules somatiques au lieu de cellules méiotiques. Cela peut donc entraîner un résultat biaisé, même si nous avons montré que les corrélations entre le taux de CO et différentes caractéristiques sont similaires qu'on utilise des tissus somatiques ou méiotiques. En outre, tous ces 9 états chromatiniens dépendent de l'écotype Col-0, et nous avons supposé que toutes les régions synténiques entre Col et Ler ont le même état chromatinien. Notre modèle pourrait bénéficier des données complètes des caractéristiques génomiques et épigénomiques des *deux* parents pour identifier les différents profils d'état chromatinien. De plus, les informations sur les structures en boucle des chromosomes méiotiques sont manquantes, et cette structuration, en termes de taille et de position, pourrait être utile pour améliorer la prédiction du taux de CO à petite échelle. Enfin, comme l'ensemble de données sur les CO que nous avons étudié provient d'une population F2, il ne reflète que la moyenne des taux de CO des mâles et des femelles. À l'heure actuelle, notre modèle n'a pas la capacité de prédire la différence entre les taux de recombinaison mâle et femelle. Dans l'ensemble, notre modèle fournit des informations utiles pour prédire les paysages de recombinaison, et il peut être étendu à d'autres croisements et espèces. Avec le développement de nouvelles technologies et de nouvelles données dans un futur proche, ce modèle pourra être amélioré pour mieux expliquer les variations des taux de recombinaison méiotique.

Dans les deuxième et troisième projets, je me suis appuyé sur les NGS pour fournir des outils moléculaires pour les futurs travaux de sélection. L'approche RAD basée sur 31 accessions d'arachide a montré que les accessions provenant de l'origine de l'arachide cultivée sont plus diversifiées génétiquement que celles provenant d'autres endroits, suggérant que les sélectionneurs taïwanais devraient introduire plus de diversité dans les programmes actuels de sélection d'arachide à Taïwan puisque les cultivars élites taïwanais sont très liés génétiquement. De plus, 14 marqueurs KASP développés dans cette étude ont été mis à disposition pour identifier la

structure de la population de 282 accessions de la collection nationale de germoplasme d'arachide. Ainsi, on peut s'appuyer sur les données de séquençage de 31 accessions pour développer davantage de marqueurs non basés sur les techniques de gel afin d'accélérer et d'améliorer le processus de sélection, par exemple lors de la sélection en fond et en premier plan. Dans le troisième et dernier projet, nous avons identifié 385 gènes candidats fortement influencés par 27 046 SNP et 5 975 indels spécifiquement identifiés dans des lignées résistantes de tomate. Seul Bwr3.2dCAPS, dans le gène Asc précédemment publié, se trouve être associé de manière statistiquement significative aux phénotypes d'une population F2. Dans cette étude, nous avons donc démontré que la comparaison par paire est utile pour identifier des QTL mineurs. Enfin, les populations F2 utilisées pour valider les gènes candidats ont été développées à partir de H7996 et d'une lignée de sélection avancée. Dans le futur, il faudrait développer plus de populations à partir des cinq autres lignées résistantes afin de valider plus de gènes candidats qui pourraient contribuer à la résistance à la BW.

En résumé, dans le premier projet, nous avons construit un modèle quantitatif qui prédit assez bien les paysages de recombinaison chez Arabidopsis thaliana. Ce travail pourrait être un outil efficace pour prédire les paysages de recombinaison dans d’autres espèces si les données liées à l'épigénomique le permettent. Dans le second projet, en mesurant la diversité génétique de l'arachide cultivée, nous avons fourni des informations utiles pour la future sélection de l'arachide à Taïwan, et les données de séquence de ce projet peuvent servir de base à la sélection moléculaire de l'arachide. Enfin, pour le troisième projet, les données sur les séquences et les gènes candidats de la tomate peuvent permettre aux sélectionneurs et généticiens de tomates de mieux concevoir des variétés de tomates présentant une résistance beaucoup plus durable à la maladie. Pris ensemble, les trois projets de ma thèse peuvent contribuer à améliorer la sélection végétale future par de nouveaux moyens méthodologiques permettant d’exploiter la diversité génétique.