

Full Paper

# Low-coverage resequencing detects meiotic recombination pattern and features in tomato RILs

Lars S. de Haas<sup>1</sup>, Roy Koopmans<sup>1</sup>, Cilia L. C. Lelivelt<sup>1</sup>, Remco Ursem<sup>1</sup>, Rob Dirks<sup>1,†</sup>, and Geo Velikkakam James<sup>1,\*</sup>

<sup>1</sup>Rijk Zwaan Breeding B.V., 4793 RS Fijnaart, The Netherlands

\*To whom correspondence should be addressed. Email: g.james@rijkszwaan.nl

<sup>†</sup>Present address: Department of Plant Production, Faculty of Bioscience Engineering, Ghent University, Coupure links 653, 9000 Gent, Belgium

Edited by Prof. Kazuhiro Sato

Received 10 December 2016; Editorial decision 5 May 2017; Accepted 18 May 2017

## Abstract

Traditional plant breeding relies on meiotic recombination for mixing of parental alleles to create novel allele combinations. Detailed analysis of recombination patterns in model organisms shows that recombination is tightly regulated within the genome, but frequencies vary extensively along chromosomes. Despite being a model organism for fruit developmental studies, high-resolution recombination patterns are lacking in tomato. In this study, we developed a novel methodology to use low-coverage resequencing to identify genome-wide recombination patterns and applied this methodology on 60 tomato Recombinant Inbred Lines (RILs). Our methodology identifies polymorphic markers from the low-coverage resequencing population data and utilizes the same data to locate the recombination breakpoints in individuals by using a variable sliding window. We identified 1,445 recombination sites comprising 112 recombination prone regions enriched for AT-rich DNA motifs. Furthermore, the recombination prone regions in tomato preferably occurred in gene promoters over intergenic regions, an observation consistent with *Arabidopsis thaliana*, *Zea mays* and *Mimulus guttatus*. Overall, our cost effective method and findings enhance the understanding of meiotic recombination in tomato and suggest evolutionarily conserved recombination associated genomic features.

**Key words:** meiotic recombination, recombination hotspot, genotyping by sequencing, DNA motifs, tomato

## 1. Introduction

Meiotic recombination plays a vital role in sexual reproduction by mixing parental alleles to create novel allele combinations. In general, both genetic mapping and traditional breeding depends upon the distribution of recombination along chromosomes. In plants, the meiotic recombination rate is tightly regulated within the genome, but varies between species.<sup>1</sup> Furthermore, the distribution of recombination events is not random along the chromosome. Regions that

are preferred and suppressed for recombination are known as recombination hotspots and coldspots, respectively. Various studies have shown that recombination prone regions vary in size and magnitude between species (*Zea mays*,<sup>2,3</sup> *Triticum aestivum*,<sup>4</sup> *Arabidopsis thaliana*<sup>5,6</sup> and *Medicago truncatula*<sup>7</sup>). Extended knowledge about genome-wide recombination patterns in a species would not only improve the understanding of recombination preferences but also accelerate fine mapping of genetic traits and breeding design of the crop.

Moreover, high-resolution recombination patterns provide the opportunity to study recombination associated genomic features, thereby the possibility to modulate recombination to accelerate breeding.

Advancement in sequencing technology has not only enabled rapid generation of draft genome assemblies for individual species but also identification of genome-wide ultra-dense markers in a population. Hence, whole genome resequencing has also enabled high resolution recombination pattern identification. Recent studies have produced base pair resolution recombination patterns in *Arabidopsis thaliana* and shown that the distribution of recombination is influenced by chromatin structure, histone variants and DNA methylation, among other factors.<sup>8,9</sup> Further analysis using next generation resequencing data revealed poly-A, CCN-repeat and CTT-repeat motifs associated with recombination prone regions in *Arabidopsis thaliana*.<sup>5,8,10</sup> The first motif was associated with nucleosome depleted regions, whereas CCN- and CTT-repeat motifs were enriched within genes.<sup>10,11</sup> However, lack of stringent resequencing data analysis led to false positive markers that inflate recombination rate.<sup>12</sup> Therefore, it is of the utmost importance to select an accurate marker-set while applying resequencing data for recombination identification.<sup>5,13</sup> This caveat is even bigger in crops with a large genome, or the ones possessing a high repeat content. Besides, an incomplete assembled reference genome representing only one haploid genome may hamper proper evaluation of recombination rates. Therefore, in crop plants, it is crucial to have a cost effective methodology that can identify genome-wide recombination patterns using low-coverage resequencing data. This is desired to reduce sequencing efforts to a minimum and at the same time maintain a low false-positive identification rate. In this study, we describe a novel methodology using meticulous filtering steps to conservatively select polymorphic markers from low-coverage resequencing population data to identify recombination sites. We applied our method to identify genome-wide recombination pattern on resequencing data from 60 tomato recombinant inbred lines (RILs) from an interspecific cross between *Solanum lycopersicum* cv. MoneyMaker and *Solanum pimpinellifolium* CGN 15528.

In crop plants, tomato is a model species for fruit developmental studies and interspecific crosses are primarily used for genetic mapping. *S. pimpinellifolium*, a wild but closely related species of *S. lycopersicum*, is a source for introgression breeding including various disease resistance, abiotic stress tolerance, and fruit quality improvements.<sup>14–18</sup> However, these interspecific crosses usually cause recombination suppression within the introgression region and require extensive additional breeding efforts for later improvements. For example, the tomato chromosomes 6 and 9 are known to contain introgression fragments for nematode and TMV resistance and conceive low or no recombination.<sup>19–21</sup> Despite the utility of local recombination pattern for an informed breeding design, there is no availability of genome-wide high-resolution interspecific recombination pattern in tomato. This is mainly due to the complexity and lack of methodology to apply resequencing to a relatively large genome for genome-wide recombination pattern identification. Like the majority of crop species, the tomato genome sequence is complex and incomplete. Current tomato reference genome has assembled 760 mega base pair (Mbp) from an estimated genome size of 900 Mbp.<sup>22</sup> This hinders the use of a resequencing approach for genome-wide recombination studies. Moreover, structural variations are reported between *S. lycopersicum* and *S. pimpinellifolium*, which impede the analysis of interspecific genomes even further.<sup>23</sup>

In this study, we applied our novel methodology to identify a reliable set of polymorphic markers from the resequencing data from an interspecific RIL population. By using allele segregation in the

population and resequencing data from one of the parents, we inferred the genotype of the second parent and estimated haplotypes for 60 RILs to identify recombination sites. We identified high-resolution genome-wide recombination patterns and identified several DNA motifs enriched at recombination prone regions, including two AT-rich motifs, which have not been reported previously in relation with recombination in plants. With this study, we have identified high-resolution recombination prone regions in tomato and genomic features associated with it.

## 2. Materials and methods

### 2.1. Data used and variant calling

We used low-coverage ( $\sim 6.3\times$ ) Illumina HiSeq 2000 resequencing data of 60 RILs (Supplementary Table S1). This F6 population was generated by an interspecific cross between *S. lycopersicum* (cv. MoneyMaker) and *S. pimpinellifolium* (CGN 15528, also referred to as CGN 14498).<sup>24</sup> Resequencing data from one of the parents, *S. lycopersicum*, was publicly available with an average genome coverage of  $\sim 38\times$ , whereas the sequencing data for *S. pimpinellifolium* was unavailable. Both RILs and *S. lycopersicum* were part of the 150 tomato Genome Resequencing Project (<http://www.tomato-genome.net/>).<sup>24</sup> We aligned the short reads from 60 RILs and *S. lycopersicum* separately to *S. lycopersicum* cv. Heinz 1706 reference genome (version SL2.40, <http://www.solgenomics.net>) using a modified version of BWA (based on version 0.5.9.) default settings.<sup>25</sup> For each sample read groups were corrected and duplicate reads were removed using PICARD tools (version 1.107, <http://broadinstitute.github.io/picard/>) with default settings. We improved the alignment by local realignment using GATK-lite (version 2.39) and these improved aligned resequencing data were subjected to variant calling by the UnifiedGenotyper of GATK-lite default settings.<sup>26</sup>

### 2.2. Population based marker-set identification

After variant calling, we strictly filtered single nucleotide variants (SNPs). To consider a genomic position a genuine marker, in contrast to sequencing and read alignment artefacts, characteristics of a SNP were needed to meet six criteria. These parameters were chosen based on the expected characteristics of the used RIL population.

By making sure that the same marker is reported in multiple plants, we made use of the recurrence of SNP information and took advantage of population's properties. Although it was expected to have approximately 30 homozygous alternative allele calls on a position, we incorporated a substantial margin to account for missing calls due to the low coverage data used and variation in the uniformity of the expected segregation pattern, considering the limited size of the population. For this study, at least 10 out of 60 RILs were required to report a homozygous alternative call on a single position.

Since the population is expected to be mostly homozygous, heterozygous SNPs have a higher chance of being either sequencing or read alignment artefacts. At the same time, it is harder to call heterozygous genotypes than it is to call homozygous genotypes in low coverage sequencing data. Therefore, we also required that the ratio of homozygous calls to the total number of calls had to be above 0.8. This way we avoided the selection of positions that were susceptible to show heterozygous genotypes throughout the majority of the population, which might for example have been caused by misaligned reads or paralogous sequences in the genome.

Meanwhile, the ratio of alternative calls to the total number of homozygous can assure the selection of markers that are evenly

segregating through the population. This ratio is expected to be equal between alternative and reference calls. However, this ratio is heavily influenced by the number of plants having, or missing, a genotype call on the respective position, which is expected to vary due to sequencing coverage. So, considering the limited population size and the amount of missing genotype calls, the ratio of alternative homozygous calls and total number of homozygous calls was set to be between 0.25 and 0.75. This was especially important in this particular study, because we were not only constructing a marker-set but also inferring the genotypes of the second parent. By confirming that both reference and alternative allele were segregating within the population, we confirmed that the parents have opposite genotypes on this position.

To correct for individual sequencing artefacts, only biallelic SNPs with a coverage of four or more for at least one of the alleles were considered. In total we yielded 4,463,846 markers after filtering and these markers were used for genotyping. A correlation between the abundance of available markers and the average coverage of a RIL was expected. The number of available markers and coverage per plant were visualized to estimate the strength of the expected correlation.

### 2.3. Inferring markers of second parent

Since resequencing data for *S. pimpinellifolium* was not available, we inferred the *S. pimpinellifolium* genotypes by assigning non-*S. lycopersicum* allele to previously identified population based markers. Markers with reference base calls reported in the *S. lycopersicum* accession were considered to have alternative base in *S. pimpinellifolium*. Out of 4,463,846 markers, 75,526 were inferred to be alternative base in *S. lycopersicum* and consequently reference base in *S. pimpinellifolium*. *S. lycopersicum* was expected to have less alternative base calls, because it is more closely related to the used reference genome as opposed to *S. pimpinellifolium*.<sup>24</sup> Using 130 Kompetitive Allele Specific PCR (KASP) assays,<sup>27</sup> we validated the inferred genotype of *S. pimpinellifolium* and estimated its accuracy.

### 2.4. Determining haplotype blocks and estimating recombination landscape

A variable sliding window approach was used to impute the haplotype blocks of RILs. Imputation of haplotypes enabled us to resolve conflicting genotypes as a result of erroneous parental markers, sequencing errors and read alignment artefacts. The used variable window was adjusted in size to the local density of available markers, with at least nine markers over a genomic distance of at least 10 kbp. Scoring markers in a variable window allowed the collective use of markers, correcting faulty individual markers and genotypes.

The use of a variable window also contributed to cope with missing markers due to low-coverage resequencing, without compromising the yielded resolution. The score of a window was calculated by adding up all genotypes (+1 for *S. lycopersicum*, 0 for heterozygous, -1 for *S. pimpinellifolium*) and dividing it by the number of markers in the window. When 50% or more of the markers in a window were heterozygous, the window was given an intermediate score of 0 (Supplementary Fig. S1: second panel).

Each time the score of a forthcoming window ( $w$ ) passed one of the thresholds  $w > 0.25$ ,  $-0.25 < w \leq 0.25$  or  $w < -0.25$  in respect to the previous window, the markers of both windows were merged and subjected to a second analysis. This was done to avoid false positive recombination events due to fluctuation in scores, caused by sequencing and read alignment artefacts. Again a variable window

approach was used, but over a greater distance of at least 250kbp with 50 markers both upwards and downwards of the presumed recombination region. This confirmed whether variation in scores was persistent, thus must have been caused by a recombination event. Finally, the inner unambiguous markers with contrasting parental genotypes were selected as borders of the recombination event.

When two or more haplotype changes were detected within 250kbp, this was considered to be most likely the result of fluctuation in scores caused by artefacts. In such cases, all involved events were merged and processed as a single event, which may or may not have resulted in the detection of a genuine recombination event. Our tool reports all the recombination sites per individual as well as the summarised recombination pattern per chromosome (Supplementary Fig. S1, first panel). This intermediate score and recombination landscape were visualised to ease manual inspection and optimization of the parameters described in “Population based marker-set identification” (Supplementary Fig. S1). In the population of 60 RILs, we calculated the average number of recombination per chromosome by considering all marker positions within the chromosome conceived with at least one recombination. The background recombination rate per chromosome was calculated by adding average recombination rate with one standard deviation. Regions having a recombination rate above the background rate were considered as recombination prone regions. The longest region of each chromosome without any recombination over all 60 RILs was considered recombination deprived region.

### 2.5. False-positive recombination events caused by the used reference genome assembly

Two versions (SI2.40 and SI2.50) of the *S. lycopersicum* cv. Heinz 1706 reference genome assembly were aligned to detect large inversions and relocations between the two reference genomes using MUMmer3's NUCmer.<sup>28</sup> For this, a minimum cluster size of 1kbp and all anchor matches regardless of their uniqueness were used. Large scale differences in alignment between these two reference assemblies indicated inconsistencies between assemblies. These differences were used to eliminate recombination prone regions overlapping with these regions to prevent false positives caused by the used assembly (version SI2.40).

### 2.6. Confirmation of inferred parental genotype and RIL haplotypes

In order to confirm both the inferred *S. pimpinellifolium* genotype and individual RIL haplotypes, we ordered the seeds of 60 RILs together with their corresponding parental lines from the 150 Tomato Project. The seeds were sown on wet tissue paper at 25 °C with a continuous exposure to light for 12 days. Due to poor seed germination, we proceeded with 55 RIL plants. Whole seedlings were harvested and outsourced to LGC Teddington, London, for genomic DNA extraction and KASP genotyping in duplo. Genotyping was done using LGC's KASP Master Mix, containing two universal cassettes (FAM and HEX), ROX<sup>™</sup> passive reference dye, Taq polymerase, free nucleotides and MgCl<sub>2</sub> in an optimized buffer solution. One hundred and fifty SNPs were selected for KASP assay-design of which 90 SNPs around recombination prone regions, 20 random, 24 in removed regions and 16 to confirm suspected false positives were indeed caused by sequencing and read alignment artefacts (Supplementary Table S2). Selected SNPs were subjected to LGC's Primer Picker software for an *in silico* examination of possible drop-outs prior to shipping.

We used an ad-hoc script to compare LGC's KASP genotypes with our predicted haplotype blocks for an exact match. Whenever one or both of the two KASP replicate assays had failed, or replicate assays did not match, the concerned assay was discarded in order to eliminate errors introduced by the used method of confirmation. This allowed us to validate polymorphic marker identification, as well as inferred genotypes of *S. pimpinellifolium* and our approach to estimate haplotype blocks for RILs.

### 2.7. Discovery of DNA motifs enriched with recombination prone regions

Identified recombination prone regions were assessed for over-representation of sequence motifs. Associated sequences were selected by our pipeline and subjected to *de novo* motif discovery by the MEME Suite.<sup>29</sup> To include sequence motifs regulating recombination upstream and downstream of the processed recombination prone regions, sets with various sized flanking regions (0, 1, 2.5 and 5kbp) were created and subjected to motif discovery by MEME.<sup>30</sup> The resulting collections of sequences had total sizes of 278kbp, 503kbp, 839kbp and 1,400kbp, respectively.

MEME was used for *de novo* 8–15bp DNA sequence motif discovery at above-mentioned collection of sequences using a Zero or One Occurrence Per Sequence model. The maximum allowed dataset size was increased to enable MEME to process the data. A control set of 1,000 random sequence collections was constructed, each containing 112 random *S. lycopersicum* cv. Heinz 1706 (reference version 2.40) sequences of lengths equal to the identified recombination prone regions together with their respective flanking regions. *De novo* discovered sequence motifs were analysed using default settings of AME<sup>31</sup> and the control set, to determine whether these motifs were over-enriched in recombination prone region compared with the random background.

### 2.8. Identification of genomic features associated with recombination prone regions

We annotated each base pair of recombination prone regions to identify genomic features, including 5'-UTR, gene start, gene, 3'-UTR, gene end and intergenic region using the ITAG2.3 annotation (<http://www.solgenomics.net>). We defined promoter and gene start features as 3kbp reverse and 200bp forward from the transcription start site, respectively. Similarly, gene end was defined as 200bp before transcription end site. To assess whether the distance of recombination events to genes is significantly lower compared with randomly sampled positions, the genomic distance between all 112 regions and their nearest gene was mapped using the ITAG version 2.3 annotation. Two sets of 1000 random genomic regions were created with and without pericentric heterochromatin. These pericentric heterochromatins were discarded in one of the two sets to exclude influence due to their low gene density.<sup>32</sup> Differences were visualized, showing the average of all randomly sampled positions in 1kbp windows, their corresponding standard deviation and observed distance to genes. We listed the annotated genes within the recombination prone regions for gene ontology (GO) enrichment analysis using the PANTHER classification system (<http://www.genontology.org/> (12 October 2016, date last accessed), which is up to date with GO annotations.<sup>33</sup> We also compared genes overlapping with recombination prone regions to nucleotide-binding site leucine-rich repeat (NB-LRR) resistance genes in tomato.<sup>34</sup>

## 3. Results

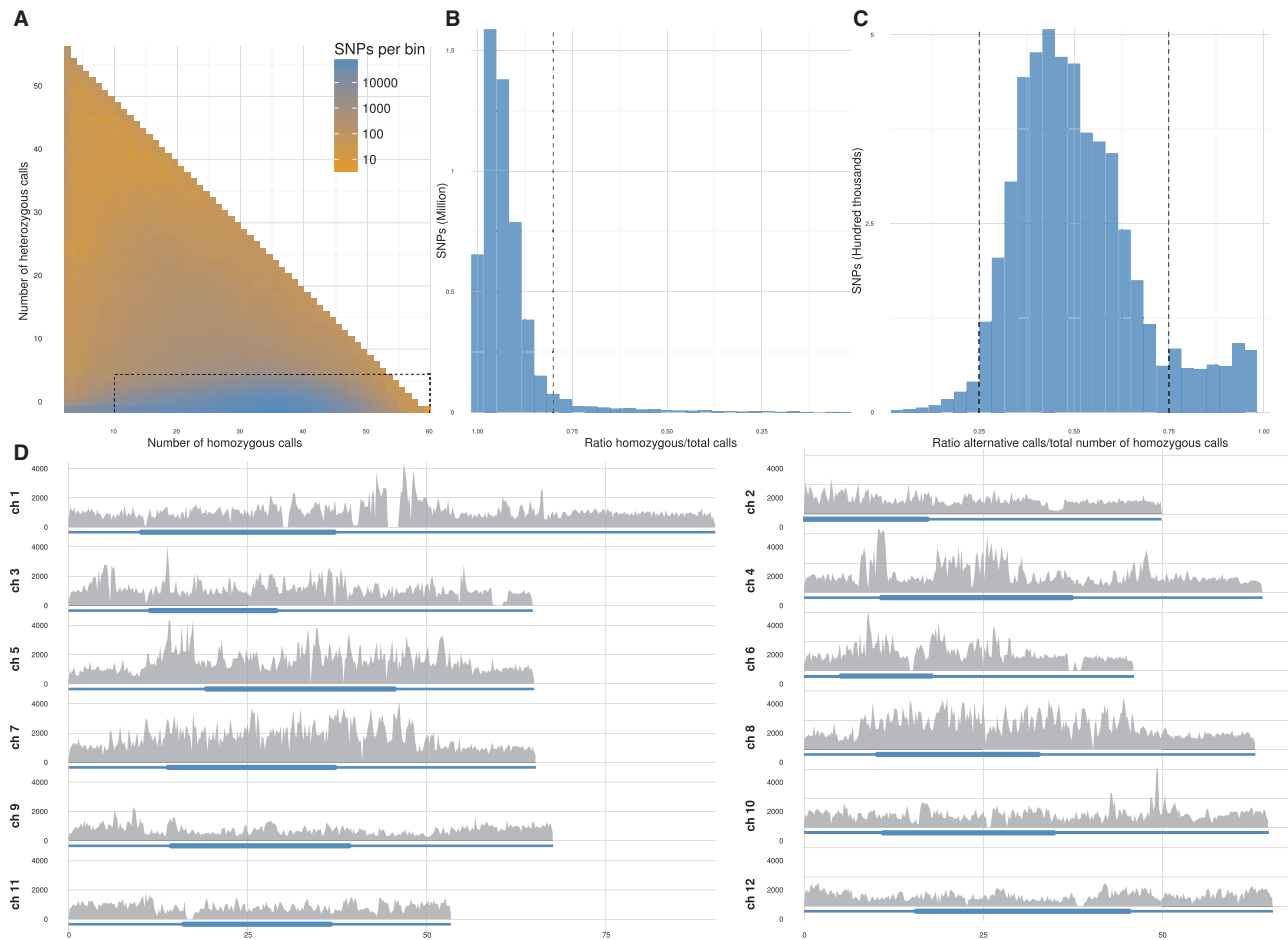
### 3.1. Identification of population based marker-set from low-coverage resequencing data

Genotyping by resequencing allows identification of recombination break points at a high-resolution, but non-allelic structural variations between parental lines with a high homology in sequence can lead to false-positive markers, especially when using short-read resequencing data.<sup>5,13</sup> Due to independent assortment of these loci, genotypes of these false positive markers will be random. Additionally, genotyping bias will increase if one of the parental lines is more closely related to the used reference genome. Together, these biases will produce false-positive markers and result in exaggerated recombination rates. Thus, it is crucial to filter out false-positive markers when estimating the recombination landscape, particularly while using low-coverage short-read resequencing data. We illustrated a novel methodology to conservatively select markers in the context to identify recombination sites by implementing fundamental population genetics principals with modifications to incorporate regions with low sequencing coverage.

Short reads from individual RILs were aligned to tomato reference genome to identify SNPs using GATK-lite pipeline. To increase the specificity of SNP identification, we applied three rounds of filters, addressing the independent allele segregation within population, fixation or near fixation of alleles and sufficient sequencing depth at SNP position in individuals as well as within the population (see section Materials and methods). As shown in Fig. 1A–C, the characteristics of most markers corresponded to what was expected from a RIL population. The majority of markers were predominantly homozygous calls, with relatively few heterozygous calls (Fig. 1A). The ratio between homozygous calls and the combined number of both homozygous and heterozygous genotypes was near to one, confirming the homozygous nature of a RIL population (Fig. 1B). We examined the segregation of reference alleles and alternative alleles and identified that the majority of markers were located around a ratio of 0.5 as expected from a biparental population (Fig. 1C). Confirmation of these characteristics justifies the reasoning behind our filtering strategy. In total 4,463,846 out of 7,979,788 unique genomic positions with reported SNPs passed all filtering steps. Of these positions, on average 2,548,113 markers were available per individual in the low-coverage sequencing data, resulting in over 150 million available markers across the population. Despite stringent filtering, marker distribution was well spread across the genome allowing us to identify recombination patterns with high-resolution (Fig. 1D). The lowest and the highest number of markers for a single chromosome of an individual were 31,707 and 563,533, respectively. The average genome coverage per individual played a crucial role in the number of available markers and showed a Pearson correlation coefficient of 0.92 (Fig. 2). Using the selected marker positions from the population, we used available parental genotype for haplotype phasing and inferred the missing parental genotype.

### 3.2. Estimated recombination rates and identified recombination prone regions

Using a variable sliding window approach to identify genotype breaks, we identified 1,445 recombination events in 60 RIL tomato plants with 123 recombination prone regions (see section Materials and Methods, Supplementary Tables S3 and S4 and Fig. S2). We identified recombination events with a median track length of 2,134bp that is comparable to *Arabidopsis thaliana*.<sup>5</sup> On an average, we identified 24 recombination events per RIL, which is in agreement with the expected number of observable recombinations.<sup>35</sup>



**Figure 1.** (A) Two-dimensional binning plot showing the number of homozygous and heterozygous calls per locus in the population. (B) Histogram showing the ratio between homozygous and the total number of calls. (C) Histogram showing the ratio between alternative and total number of homozygous calls. Cut offs are shown in dotted lines. (D) Genome-wide distribution of markers in 100kb bins (y-axis). Chromosomes are visualized in blue lines, pericentric heterochromatin are visualized in thick blue lines. (See online article for colour version of this figure).

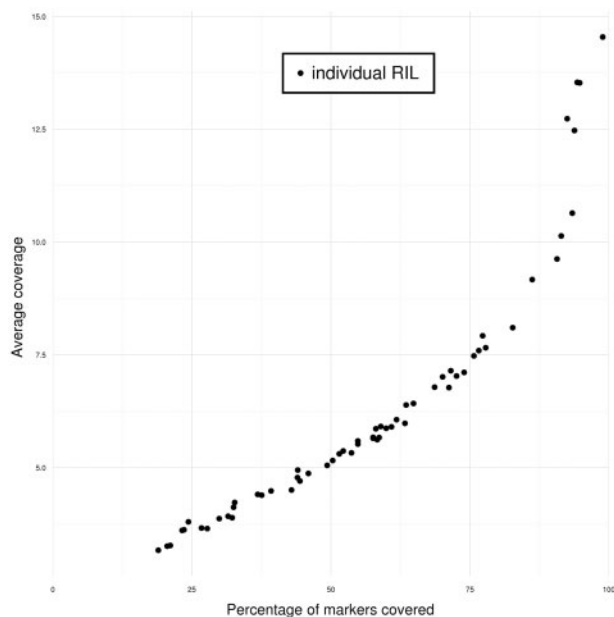
During genotyping, three genomic regions were shown to be polymorphic for individual markers, but haplotypes of these genomic regions could not clearly be assigned to one of the parental genomes. As shown in previous studies, abnormalities in reference genome could lead to such patterns.<sup>5,13</sup> Therefore, we removed recombination prone regions overlapping with these regions from further analysis. As a result, 0.32% of the markers were removed from further analysis. We further extended the above observation as large structural variation due to conflict in reference genome assembly, which can introduce false recombination events and consequently false recombination prone regions.<sup>23,36</sup> We aligned the two most recent *S. lycopersicum* cv. Heinz 1706 reference genomes (SL2.40 and SL2.50) and identified large inversions and rearrangements on chromosome 1, 3 and 12 (Supplementary Fig. S3). Structural variation was also seen in other chromosomes, but these cases did not interfere with recombination events. In total, 11 recombination prone regions (two from chromosome 1 and all from chromosomes 3 and 12) were attributed to structural variation in the used reference genome. This reduced set of 112 recombination prone regions with a median of 1,204bp, making up a total size of 278,937bp was used for identification of genomic features. Because of differences in number of available markers, due to variance in sequence coverage, recombination site resolution varied between individuals. However, by utilizing the shared regions in the individually observed recombination events to

define recombination prone regions allowed us to improve the resolution of these regions. These high-resolution recombination intervals in an interspecific tomato genome, allowed us to expand the knowledge of recombination patterns in tomato at a sequence level.

The low number of recombinations per chromosome hampered fine resolution identification of recombination deprived regions. However, regions on all chromosomes without any recombination event at all were excessively large, ranging from 17.8 to 60.0Mbp (Supplementary Table S5). Recombination prone and deprived regions were clearly overlapping with euchromatin and pericentric heterochromatin regions of the chromosome, respectively (Fig. 3).<sup>22,32</sup> This consistent pattern has been observed in other plant species indicating an evolutionarily conserved preference mechanism either due to local genomic features or open chromatin structure.<sup>1</sup> We used recombination prone regions to perform *de novo* DNA motif discovery, together with other genomic features identification.

### 3.3. Validated inferred parental genotype and RIL haplotype blocks

Validation of our genotyping approach for both parental lines and RILs was done using KASP assays. Five plants were excluded because either seed germination or DNA isolation was unsuccessful. An



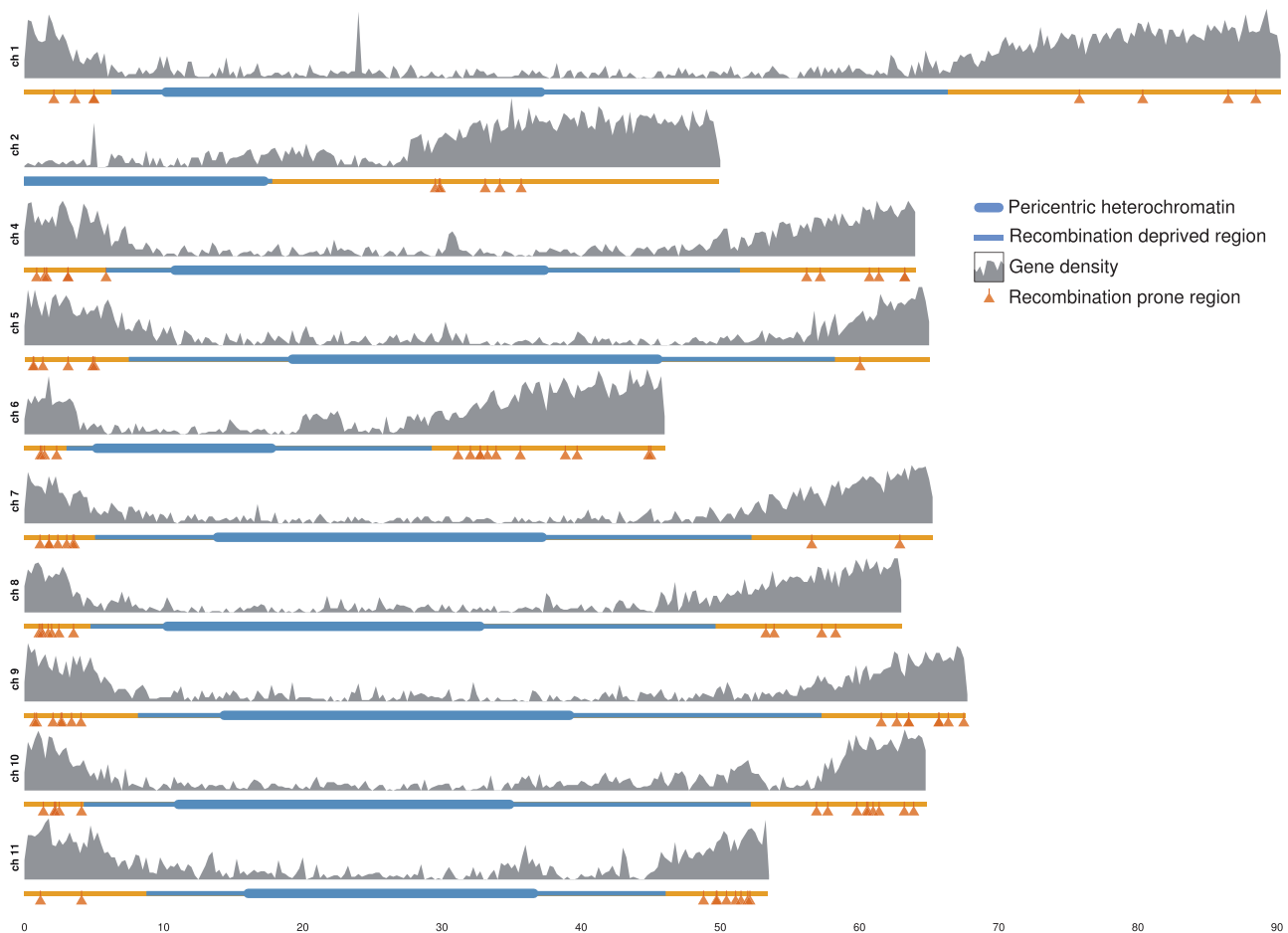
**Figure 2.** Correlation between abundance of available markers (x-axis) and average genome coverage of the each RIL (y-axis).

additional four plants were discarded after genotyping because of high levels of heterozygosity. Two plants showing random genotypes were also removed, as these plants were noticeably different from the corresponding resequenced RILs and were most likely the result of a mix up of supplied seeds, either before or during genotyping. During data analysis, three out of 133 successful KASP assays were discarded. Two of them failed for more than half of the population and one was non-polymorphic. All the remaining assays were polymorphic, confirming the identification of polymorphic markers from resequencing data.




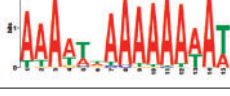


After above-mentioned filtering steps, we successfully genotyped 49 RILs using 130 KASP assays. RIL haplotypes estimated with low-coverage resequencing data were validated by KASP assays with an accuracy of 96.8%. This accuracy increased to 98.0% when only homozygous markers were considered. These high accuracies confirmed our methodology for both marker selection and haplotype block estimation using low-coverage resequencing data. Markers identified by resequencing of *S. lycopersicum* and inferred markers of *S. pimpinellifolium* matched with 99.2% of the assays, proving high accuracy of our methodology to infer markers for one of the parental lines.

### 3.4. AT-rich DNA motifs are enriched at recombination prone regions

Previous reports in *Arabidopsis thaliana* and *Citrus clementina* have identified DNA motifs associated with recombination sites.<sup>5,8,37</sup>



**Figure 3.** Summary plot of recombination prone and deprived regions along chromosomes 1 and 2 and 4–11. Pericentric heterochromatins are shown in thick blue lines.<sup>22</sup> Per chromosome, the biggest contiguous region completely deprived of recombination events are coloured blue, the remaining is coloured in yellow. Regions prone to recombination are indicated by orange triangles. (See online article for colour version of this figure).

Sequence motif	Occurrence in recombination prone regions	Adjusted p-value	Name of motif
	100%	4.47e-16	Poly-T/poly-A
	100%	5.13e-11	Poly-T
	41%	7.37e-11	TATA-repeat
	95%	6.69e-9	Poly-A
	100%	1.46e-3	Poly-A
	100%	2.17e-3	Poly-T

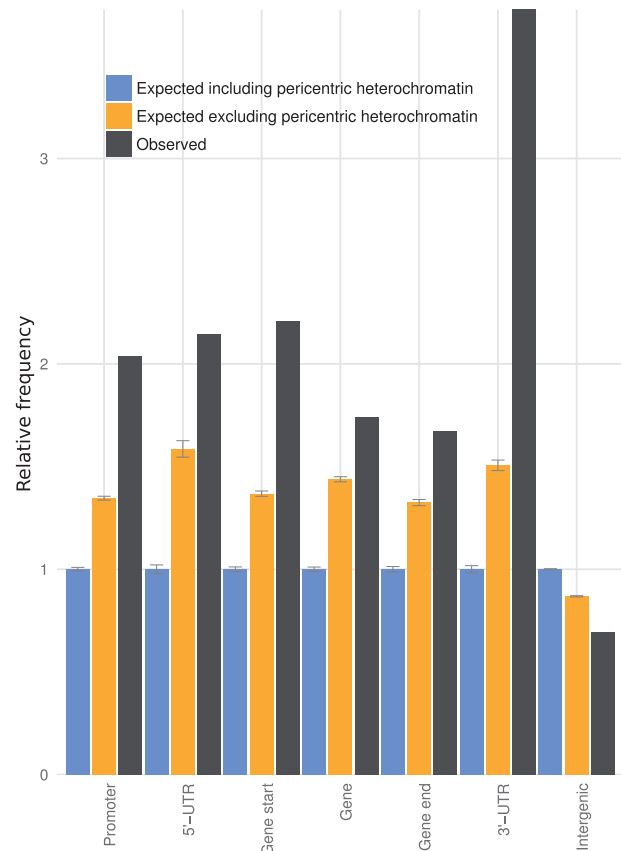
**Figure 4.** DNA sequence motifs enriched at recombination prone regions, including 1000bp flanking regions. (See online article for colour version of this figure).

In tomato, since the genome-wide high-resolution recombination pattern was unavailable, it was not possible to analyse recombination associated DNA sequence motifs so far. We carried out a *de novo* DNA motif discovery in recombination prone regions together with 0, 1, 2.5 and 5kbp flanking regions to discover sequence motifs in tomato. Several motifs were discovered out of which six motifs were found to be over-enriched compared to equivalent random sequences. These motifs were consistently enriched in recombination prone regions with 0–5kbp flanking regions (see section Materials and methods; Fig. 4).

Within enriched motifs, homopolymer poly-A and poly-T occurred in all recombination prone regions. Additionally, we identified a poly-T/poly-A motif in all recombination prone regions with the highest adjusted *P*-value compared to the whole tomato genome as background (Fig. 4). This motif is similar to motifs reported in earlier studies<sup>5,8</sup> and associated with nucleosome boundaries.<sup>38</sup> Besides the poly-T/poly-A motif, a 15bp TATA-repeat was discovered. Both the TATA-repeat and poly-T/poly-A motif were reported for the first time and further studies are required to establish the functional correlation. In contrast, the CCN-repeat motif previously identified in *Arabidopsis* was not identified in our study. Meanwhile, the CTT-repeat motif identified in *Arabidopsis* was detected, but was not enriched.

### 3.5. Recombination prone regions in tomato occurred at gene and promoter regions

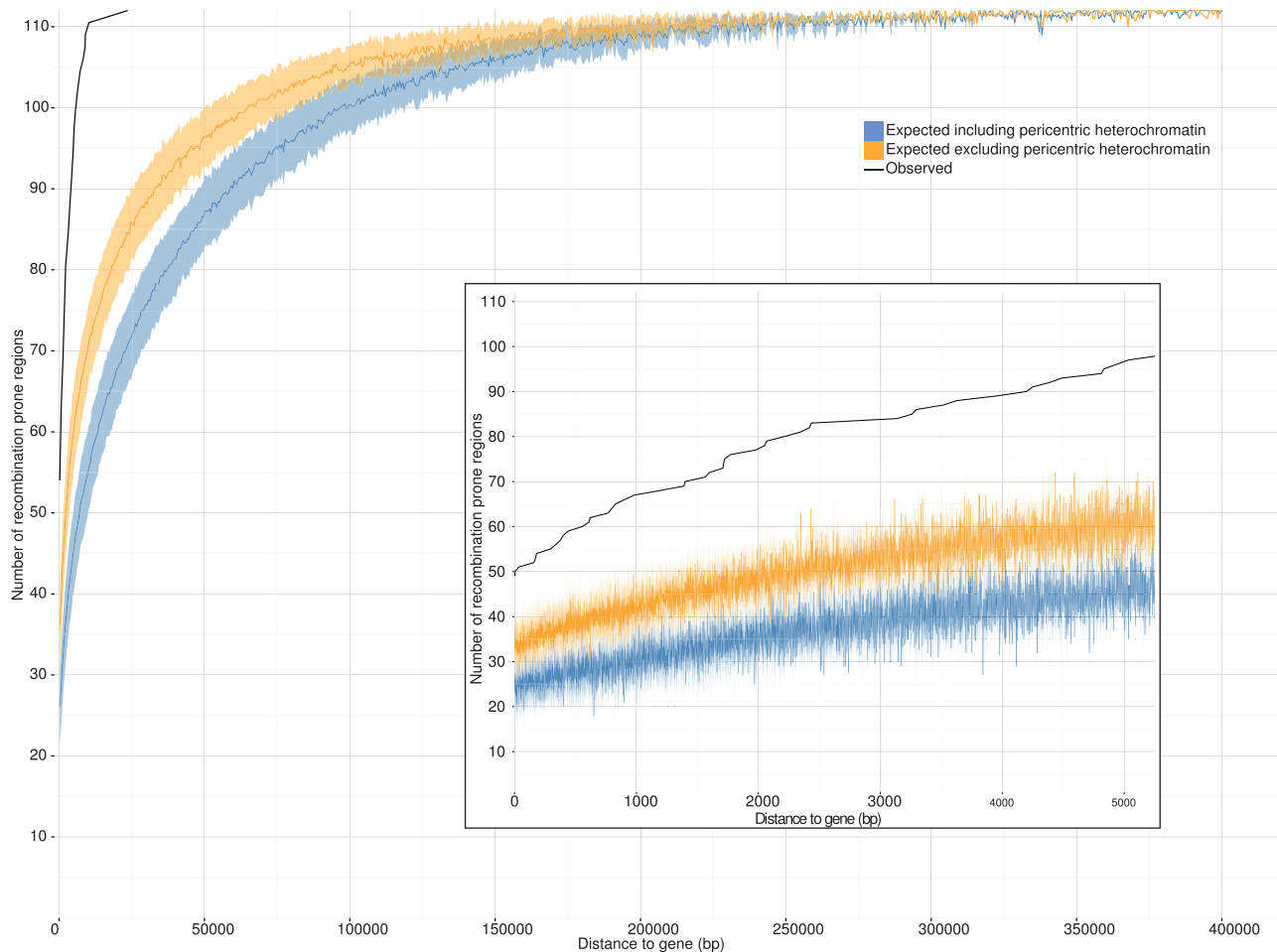
We further analysed other genomic features associated with recombination prone regions. All recombination prone regions were observed



**Figure 5.** Positions of recombination prone regions based on tomato genome annotation. Standard error is represented by grey error bars. (See online article for colour version of this figure).

within euchromatin regions. We examined the position of recombination prone regions in the genome using tomato gene annotation. Recombination prone regions showed a tendency to be at or near to genes. This is consistent to previous observations in *Arabidopsis*.<sup>5</sup> We have found that gene body was significantly over-represented in recombination prone regions (*P*-value 0.01 [permutation test]). Notably, features such as 3'-UTR and promoter showed significant enrichment in recombination prone regions (*P*-value  $1.403 \times 10^{-6}$  and 0.00043, respectively [permutation test]). Forty-eight recombination prone regions overlapped with at least one annotated gene. We repeated the analysis by excluding pericentric heterochromatin from the expected probability analysis. This stringent analysis also showed consistent enrichment (Fig. 5). However, for genes overlapping with recombination prone regions, no enrichment was found for either any gene ontology category or NB-LRR resistance genes.

We visualized the genomic distance from recombination prone regions to the nearest gene and observed that the distance to the nearest gene was lower when starting from a recombination prone region, than when starting from a randomly sampled region (Fig. 6). We observed that a remarkable 74% of recombination prone regions were at gene or promoter regions, compared with the random average expected rate of 32%. Furthermore, this pattern was persistent despite removing pericentric heterochromatin, known for low gene density, from the analysis. Similar observations were made in *Arabidopsis thaliana* where recombination targets AT-rich promoter sites.<sup>5</sup> Although tomato has lower gene density compared with *Arabidopsis thaliana*, strikingly, it has over-enrichment of



**Figure 6.** Genomic distance from recombination prone regions to the nearest annotated gene in 500bp bins. Black line shows the observed distance from recombination prone region to the nearest annotated gene, whereas the expected background mean and standard deviation are shown in blue and orange. Magnified version of first 5kbp is shown for visual easiness without using bins. (See online article for colour version of this figure).

recombination prone regions at gene body. This might be due to longer tomato genes and the genomic composition of the tomato genome. These observations indicate that gene or promoter associated genomic features, such as DNA motifs and chromatin stage, may influence recombination sites.

#### 4. Discussion

Marker identification and selection are crucial when markers are identified using low-coverage resequencing data. Therefore, we developed a population-based methodology to compile marker sets, using various criteria to ensure selected markers were genuine markers rather than sequencing or reference genome artefacts. The methodology is based on only selecting genomic positions that have multiple unequivocal called SNPs that match with the genetic characteristics of those expected from a RIL population. By selecting these positions, the sequencing and read alignment artefacts introduced with low-coverage sequencing were reduced to less than one per cent. In order to achieve this, markers for a RIL population were filtered to contain high percentages of homozygous calls and an equal ratio between reference and alternative calls. Different parameters will be desired when using other population types and sizes. Also, the quality

of the used sequencing technique and the average depth strongly influence the desired parameters. We observed that in tomato the added value of increased genome coverage is depreciating above a coverage of 10, similar to *Arabidopsis thaliana* resequencing recommendations.<sup>39</sup> Therefore, we suggest a minimum coverage of 10 for projects with similar experimental setup and aim.

We confirmed the efficiency of our approach to estimate haplotypes by KASP validation, and identified a specificity of 97.96% for homozygous haplotypes. The same KASP validation was used to show the efficiency and specificity of our inferred *S. pimpinellifolium* genotypes and validated 99.63% of these inferred genotypes. Only one assay was found to be monomorphic between parental lines. However, within the RIL population, this locus had 18 reference calls, 24 alternative calls and 1 heterozygous call, pointing to the likelihood that this error was introduced by the used KASP assay.

Despite conservative selection of markers, we were able to use 4,463,846 markers for genotyping. This in combination with the variable sliding window approach we developed, allowed us to identify recombination patterns in tomato at an unprecedented resolution, with a median size of 2,134bp. By accumulating the identified recombination events, we managed to increase the resolution of recombination prone regions to a median size of 1,204bp. The achieved resolution is the highest so far in tomato, improving the



accuracy of subsequent identification of associated genomic features. Although the false-negative rate of our variable sliding window approach is unknown, we deliberately selected parameters and elected specificity over sensitivity in order to accurately predict genomic features associated with recombination prone regions. Genomes with high repeat content and incomplete reference genome would benefit the most from our marker selection and methodology to estimate haplotype blocks.

We have identified 112 recombination prone regions from 60 interspecific tomato RILs in which 74% of recombination prone regions overlapped with gene or promoter regions. This strong preference is consistent with previous reports from *Arabidopsis thaliana*,<sup>5,10</sup> *Zea mays*<sup>40</sup> and *Mimulus guttatus*.<sup>41</sup> Contradictory to *Arabidopsis thaliana* and *Oryza sativa*, where high recombination rate in resistance genes was reported, our recombination prone regions showed no enrichment for these type of genes.<sup>42,43</sup> This might be due to heterogeneity in resistance genes or diversity in used parental lines. In all chromosomes, we identified recombination deprived regions to be centric to centromere. This might be due to the extended pericentric heterochromatin in tomato.<sup>32</sup>

Previous studies, particularly those in *Arabidopsis thaliana*, have found poly-A and CTT-repeat motifs to be enriched in hotspot sequences. Recombination prone regions described in this study were found to be enriched with AT-rich motifs. The presence of poly-A motif in all recombination prone regions was prominent in our analysis, whereas CTT-repeats were not enriched. However, lately it has been shown in *Arabidopsis* that CCN repeats were enriched at crossover regions though they were preferentially associated with genes but not with promoters.<sup>10</sup> These observations indicate that although similarities in DNA motif and genomic features are found across plant species, there are differences too. Differences between eukaryotic species in mechanistic performances also are likely to occur, when looking at preferred recombination sites in human, mice, plants and yeast.<sup>1,44</sup>

Besides previously reported poly-A motif, we identified a novel TATA-repeat and poly-T/poly-A motif. According to the literature, these AT-rich motifs are associated with strong nucleosome free regions, of which poly-T/A results in the most robust nucleosome-free region.<sup>38</sup> Furthermore, the flanking DNA on both sides of the motif is considered to be nucleosome depleted as well, and this effect would remain significant over ~100–150bp.<sup>45</sup> In yeast, nucleosome depletion is in fact especially pronounced around AT-rich repeats and nucleosomes containing the histone variant H2A.Z tend to border tandem repeats.<sup>46,47</sup> Additionally, H2A.Z is associated with transcriptional regulation, anti-silencing, silencing and genome stability suggesting that H2A.Z and recombination rates could have several other biological relevancies.<sup>48</sup> Hence the presence of poly-T to poly-A tract, among the other enriched motifs in recombination-related sequences in tomato, infers the possibility of nucleosome depletion at recombination prone region. These consistent and emergent observations about the association between recombination prone region, gene promoter region, nucleosome depletion and AT-rich DNA motifs indicate an evolutionary conserved mechanism within eukaryotes and demands further causality studies. Better understanding of the effect of these genetic and genomics features on recombination will provide tools to modulate recombination for the advancement of plant breeding.

## Data availability

This tool is available for academic research. Contact corresponding author to request a copy. Nucleotide sequences are deposited to the

European Nucleotide Archive by the Plant Research International. The accession number for this study is PRJEB6659. The filtered marker-set can be downloaded from [ftp://ftp.solgenomics.net/manuals/Geo\\_2017/Supplementary\\_table\\_S2\\_Marker-set.txt](ftp://ftp.solgenomics.net/manuals/Geo_2017/Supplementary_table_S2_Marker-set.txt)

## Acknowledgements

We are grateful to all the members of 150 tomato genome project consortium and in particular to Dr. Richard Finkers for sharing the RIL seeds. Also, we would like to thank Ms. Eva Paulina Verhofstad and Dr. Shushimita Shushimita for their detailed comments in the final stage of our manuscript. Finally, we would like to thank Martijn van Elk for his assistance during read alignment and variant calling.

## Conflict of interest

None declared.

## Supplementary data

Supplementary data are available at DNARES online.

## Funding

No external funding.

## References

- Mercier, R., Mézard, C., Jenczewski, E., Macaisne, N. and Grelon, M. 2015, The molecular biology of meiosis in plants, *Annu. Rev. Plant Biol.*, **66**, 297–327.
- Bauer, E., Falque, M., Walter, H., et al. 2013, Intraspecific variation of recombination rate in maize, *Genome Biol.*, **14**, R103.
- Schnable, P.S., Ware, D., Fulton, R.S., et al. 2009, The B73 maize genome: complexity, diversity, and dynamics, *Science*, **326**, 1112–5.
- Saintenac, C., Faure, S., Remay, A., et al. 2011, Variation in crossover rates across a 3-Mb contig of bread wheat (*Triticum aestivum*) reveals the presence of a meiotic recombination hotspot, *Chromosoma*, **120**, 185–98.
- Wijnker, E., James, G.V., Ding, J., et al. 2013, The genomic landscape of meiotic crossovers and gene conversions in *Arabidopsis thaliana*, *eLife*, **2**, e01426.
- Salomé, P.A., Bomblies, K., Fitz, J., et al. 2012, The recombination landscape in *Arabidopsis thaliana* F2 populations, *Heredity*, **108**, 447–55.
- Paape, T., Zhou, P., Branca, A., Briskine, R., Young, N. and Tiffin, P. 2012, Fine-scale population recombination rates, hotspots, and correlates of recombination in the *Medicago truncatula* genome, *Genome Biol. Evol.*, **4**, 726–37.
- Choi, K., Zhao, X., Kelly, K.A., et al. 2013, *Arabidopsis* meiotic crossover hot spots overlap with H2A.Z nucleosomes at gene promoters, *Nat. Genet.*, **45**, 1327–36.
- Zhang, X., Bernatavichute, Y.V., Cokus, S., Pellegrini, M. and Jacobsen, S.E. 2009, Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in *Arabidopsis thaliana*, *Genome Biol.*, **10**, R62.
- Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N. and Levy, A.A. 2015, DNA crossover motifs associated with epigenetic modifications delineate open chromatin regions in *Arabidopsis*, *Plant Cell*, **27**, 2427–36.
- Melamed-Bessudo, C., Shilo, S. and Levy, A.A. 2016, Meiotic recombination and genome evolution in plants, *Curr. Opin. Plant Biol.*, **30**, 82–7.
- Yang, S., Yuan, Y., Wang, L., et al. 2012, Great majority of recombination events in *Arabidopsis* are gene conversion events, *Proc. Natl. Acad. Sci. U. S. A.*, **109**, 20992–7.
- Qi, J., Chen, Y., Copenhaver, G.P. and Ma, H. 2014, Detection of genomic variations and DNA polymorphisms and impact on analysis of

- meiotic recombination and genetic mapping, *Proc. Natl. Acad. Sci.*, **111**, 10007–12.
14. Grandillo, S., Ku, H.M. and Tanksley, S.D. 1999, Identifying the loci responsible for natural variation in fruit size and shape in tomato, *Theor. Appl. Genet.*, **99**, 978–87.
  15. Foolad, M.R., Chen, F.Q. and Lin, G.Y. 1998, RFLP mapping of QTLs conferring cold tolerance during seed germination in an interspecific cross of tomato, *Mol. Breed.*, **4**, 519–29.
  16. Chen, F.Q., Foolad, M.R., Hyman, J., Clair, D.A.S. and Beelman, R.B. 1999, Mapping of QTLs for lycopene and other fruit traits in a *Lycopersicon esculentum* × *L. pimpinellifolium* cross and comparison of QTLs across tomato species, *Mol. Breed.*, **5**, 283–99.
  17. Ashrafi, H., Kinkade, M. and Foolad, M.R. 2009, A new genetic linkage map of tomato based on a *Solanum lycopersicum* × *S. pimpinellifolium* RIL population displaying locations of candidate pathogen response genes, *Genome Natl. Res. Coun. Can. Génome Cons. Natl. Rech. Can.*, **52**, 935–56.
  18. Viquez-Zamora, M., Caro, M., Finkers, R., et al. 2014, Mapping in the era of sequencing: high density genotyping and its application for mapping TYLCV resistance in *Solanum pimpinellifolium*, *BMC Genomics*, **15**, 1152.
  19. Ganai, M.W. and Tanksley, S.D. 1996, Recombination around the Tm2a and Mi resistance genes in different crosses of *Lycopersicon peruvianum*, *Theor. Appl. Genet.*, **92**, 101–8.
  20. Foolad, M.R. 2007, Genome mapping and molecular breeding of tomato, *Int. J. Plant Genomics*, 2007.
  21. Liharska, T., Wordragen, M., Kammen, A., Zabel, P. and Koornneef, M. 1996, Tomato chromosome 6: effect of alien chromosomal segments on recombinant frequencies, *Genome*, **39**, 485–91.
  22. Sato, S., Tabata, S., Hirakawa, H., et al. 2012, The tomato genome sequence provides insights into fleshy fruit evolution, *Nature*, **485**, 635–41.
  23. Anderson, L.K., Covey, P.A., Larsen, L.R., Bedinger, P. and Stack, S.M. 2010, Structural differences in chromosomes distinguish species in the tomato clade, *Cytogenet. Genome Res.*, **129**, 24–34.
  24. The 100 Tomato Genome Sequencing Consortium, Aflitos, S., Schijlen, E., et al. 2014, Exploring genetic variation in the tomato (*Solanum section Lycopersicon*) clade by whole-genome sequencing, *Plant J.*, **80**, 136–48.
  25. Li, H. and Durbin, R. 2009, Fast and accurate short read alignment with Burrows–Wheeler transform, *Bioinformatics*, **25**, 1754–60.
  26. DePristo, M.A., Banks, E., Poplin, R., et al. 2011, A framework for variation discovery and genotyping using next-generation DNA sequencing data, *Nat. Genet.*, **43**, 491–8.
  27. Semagn, K., Babu, R., Hearne, S. and Olsen, M. 2014, Single nucleotide polymorphism genotyping using Kompetitive Allele Specific PCR (KASP): overview of the technology and its application in crop improvement, *Mol. Breed.*, **33**, 1–14.
  28. Kurtz, S., Phillippy, A., Delcher, A.L., et al. 2004, Versatile and open software for comparing large genomes, *Genome Biol.*, **5**, R12.
  29. Bailey, T.L., Boden, M., Buske, F.A., et al. 2009, MEME Suite: tools for motif discovery and searching, *Nucleic Acids Res.*, **37**, W202–8.
  30. Bailey, T.L. and Elkan, C. 1994, Fitting a mixture model by expectation maximization to discover motifs in biopolymers, *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, **2**, 28–36.
  31. McLeay, R.C. and Bailey, T.L. 2010, Motif Enrichment Analysis: a unified framework and an evaluation on ChIP data, *BMC Bioinformatics*, **11**, 165.
  32. Wang, Y., Tang, X., Cheng, Z., Mueller, L., Giovannoni, J. and Tanksley, S.D. 2006, Euchromatin and pericentromeric heterochromatin: comparative composition in the tomato genome, *Genetics*, **172**, 2529–40.
  33. Mi, H., Muruganujan, A., Casagrande, J.T. and Thomas, P.D. 2013, Large-scale gene function analysis with the PANTHER classification system, *Nat. Protoc.*, **8**, 1551–66.
  34. Andolfo, G., Jupe, F., Witek, K., Etherington, G.J., Ercolano, M.R. and Jones, J.D.G. 2014, Defining the full tomato NB-LRR resistance gene repertoire using genomic and cDNA RenSeq, *BMC Plant Biol.*, **14**, 120.
  35. Zheng, C., P Boer, M. and van Eeuwijk, F.A. 2014, A general modeling framework for genome ancestral origins in multiparental populations, *Genetics*, **198**, 87–101.
  36. Shearer, L.A., Anderson, L.K., de Jong, H., et al. 2014, Fluorescence in situ hybridization and optical mapping to correct scaffold arrangement in the tomato genome, *G3 GenesGenomesGenetics*, **4**, 1395–405.
  37. Terol, J., Ibañez, V., Carbonell, J., et al. 2015, Involvement of a citrus meiotic recombination TTC-repeat motif in the formation of gross deletions generated by ionizing radiation and MULE activation, *BMC Genomics*, **16**, 69.
  38. de Boer, C.G. and Hughes, T.R. 2014, Poly-dA:dT tracts form an in vivo nucleosomal turnstile, *PLoS One*, **9**, e110479.
  39. James, G.V., Patel, V., Nordström, K.J., et al. 2013, User guide for mapping-by-sequencing in Arabidopsis, *Genome Biol.*, **14**, R61.
  40. Li, X., Li, L. and Yan, J. 2015, Dissecting meiotic recombination based on tetrad analysis by single-microspore sequencing in maize, *Nat. Commun.*, **6**, 6648.
  41. Hellsten, U., Wright, K.M., Jenkins, J., et al. 2013, Fine-scale variation in meiotic recombination in *Mimulus* inferred from population shotgun sequencing, *Proc. Natl. Acad. Sci.*, **110**, 19478–82.
  42. Si, W., Yuan, Y., Huang, J., et al. 2015, Widely distributed hot and cold spots in meiotic recombination as shown by the sequencing of rice F2 plants, *New Phytol.*, **206**, 1491–502.
  43. Choi, K., Reinhard, C., Serra, H., et al. 2016, Recombination rate heterogeneity within Arabidopsis disease resistance genes, *PLoS Genet.*, **12**, e1006179.
  44. de Massy, B. 2013, Initiation of meiotic recombination: how and where? Conservation and specificities among eukaryotes, *Annu. Rev. Genet.*, **47**, 563–99.
  45. Segal, E. and Widom, J. 2009, Poly(dA:dT) Tracts: major determinants of nucleosome organization, *Curr. Opin. Struct. Biol.*, **19**, 65–71.
  46. Pan, J., Sasaki, M., Kniewel, R., et al. 2011, A hierarchical combination of factors shapes the genome-wide topography of yeast meiotic recombination initiation, *Cell*, **144**, 719–31.
  47. Vices, M.D., Legendre, M., Caldara, M., Hagihara, M. and Verstrepen, K.J. 2009, Unstable tandem repeats in promoters confer transcriptional evolvability, *Science*, **324**, 1213–6.
  48. Guillemette, B., Bataille, A.R., Gévry, N., et al. 2005, Variant histone H2A.Z is globally localized to the promoters of inactive yeast genes and regulates nucleosome positioning, *PLoS Biol.*, **3**.