Linked-read sequencing of gametes allows efficient genome-wide analysis of meiotic recombination

Hequan Sun[1,†], Beth A. Rowan[2,†,*], Pádraic J. Flood[1], Ronny Brandt[3], Janina Fuss[3], Angela M. Hancock[1], Richard W. Michelmore[2], Bruno Huettel[3], Korbinian Schneeberger[1,*]

[1]Department of Plant Developmental Biology, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany [2]The Genome Center and Department of Plant Sciences, University of California, Davis, 451 East Health Sciences Drive, Davis, California 95616, USA [3]Max Planck-Genome-center Cologne, Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, 50829 Cologne, Germany

[†]These authors contributed equally to this work.

*Correspondence:

Korbinian Schneeberger (schneeberger@mpipz.mpg.de)

Beth A. Rowan (browan@ucdavis.edu)

Key words: meiotic recombination, crossover detection, single molecule, linked-read sequencing, *Arabidopsis thaliana*

ABSTRACT

Meiotic crossovers (COs) ensure proper chromosome segregation and redistribute the genetic variation that is transmitted to the next generation. Existing methods for CO identification are challenged by large populations and the demand for genome-wide and fine-scale resolution. Taking advantage of linked-read sequencing, we developed a highly efficient method for genome-wide identification of COs at kilobase resolution in pooled recombinants. We first tested this method using a pool of Arabidopsis F_2 recombinants, and obtained results that recapitulated those identified from the same plants using individual whole-genome sequencing. By applying this method to a pool of pollen DNA from a single F_1 plant, we established a highly accurate CO landscape without generating or sequencing a single recombinant plant. The simplicity of this approach now enables the simultaneous generation and analysis of multiple CO landscapes and thereby allows for efficient comparison of genotypic and environmental effects on recombination, accelerating the pace at which the mechanisms for the regulation of recombination can be elucidated.

INTRODUCTION

During meiosis, existing genetic variation is reshuffled and passed down to offspring through COs that exchange portions of homologous chromosomes. This results in new combinations of alleles in the next generation that can generate novel phenotypic variation, which is the raw material for both natural and artificial selection[1-3]. In most organisms, the locations of COs along chromosomes do not form a random distribution[4-11]. Thus, the local CO rate governs the types of allelic combinations that can arise through sexual reproduction. Both environmental and genetic factors have been shown to affect CO rates and distributions; some examples are the local sequence divergence or rearrangements between the homologous chromosomes[12-14], the chromatin context[15-18], variation in DNA repair mechanisms[19-25], and environmental stress[26,27].

Despite a large body of research on CO formation, our knowledge of what determines where and how often COs occur is still incomplete, in part because the time, effort, and resources needed to study this phenomenon have been limiting. Genotyping recombinant individuals, either by classical methods[28], reduced representation sequencing[29,30], or whole genome sequencing[31] and performing cytological analysis of meiotic cells[24] represent the common methods for determining CO locations and frequencies. However, none of these methods is easily suited for high-throughput analysis of thousands of COs in parallel. The use of recombination reporters[32,33] and pollen-typing[34] enable rapid screening, but can only assess differences in CO frequency in a specific region of the genome. The availability of an efficient method to assess the genome-wide distribution and frequency of COs at a high resolution would greatly enhance our understanding of the processes that govern meiotic recombination.

Here we investigated the use of linked-read sequencing of bulk recombinants for high-throughput genome-wide determination of COs in *Arabidopsis thaliana.* We first developed and assessed this approach using bulked F_2 individuals within known recombination sites and then

applied this method to hybrid pollen to generate a genome-wide CO map with a single sequencing experiment and without even growing a single recombinant plant. These results show that the time and effort needed to generate genome-wide CO maps can be reduced to sequencing and analyzing a single DNA library, making it feasible to compare multiple CO maps and thereby determine the effect of genetic and environmental factors within a single study. We believe that this method will be widely applicable to different organisms and will have a huge impact on the design of experiments aiming to decipher how meiotic recombination is regulated.

RESULTS

CO breakpoint detection from bulk recombinants

To establish a set of COs for verifying our method, we first performed whole-genome sequencing of 50 individual F_2 plants derived from a cross of two of the best-studied inbred lab strains of Arabidopsis, Col-0 and L*er* and used the haplotype reconstruction software TIGER[31] to determine a benchmark set of 400 COs across all 50 genomes (Fig. 1; Materials and Methods; Supplementary Tables 1 and 2).

We then bulked the identical 50 F_2 plants by pooling individual leaves of comparable size and extracting high molecular weight (HMW) DNA[35] (Fig. 1). After size selection and quality control (Supplementary Fig. 1), we loaded 0.25 ng DNA into a 10X Genomics Chromium Controller. The Chromium Controller encapsulates millions of gel beads as GEMs (Gel bead in EMulsion), each of which is loaded with a small number of long DNA molecules. These long molecules are fragmented and ligated with GEM-specific DNA barcodes to generate a 10X library suitable for Illumina sequencing. This library, which we called P50L25, was whole-genome sequenced with 84 million 151 bp-read pairs (Supplementary Table 1).

After aligning the reads against the Col-0 reference sequence[36] using *longranger* (v2.2.2, 10X Genomics), we recovered 3.6 million molecules (≥1 kb) including 116 million reads (Fig. 2a; Materials and Methods). On average, these molecules were ~45 kb in size and were covered by ~21 read pairs, leading to a molecule base coverage of ~0.16x (Fig. 2b-d; Supplementary Table 1). To avoid chimeras resulting from the accidental co-occurrence of two independent, but closely-spaced molecules with identical barcodes, we selected molecules which were smaller than 65 kb that had fewer than 55 reads and no heterozygous genotypes. These thresholds were based on the distributions in molecule size and read number per molecule (Fig. 2b-d; Materials and Methods; Supplementary Note 1). Overall, 2.7 million molecules passed all filtering.

Initially, we genotyped these molecules at the ~660,000 SNP markers predicted by *longranger*. If an individual molecule was composed of two distinct clusters of different parental alleles and fulfilled additional criteria regarding length and marker distribution (Fig. 2a; Supplementary Table 3; Supplementary Note 1), the molecule was considered as a recombinant molecule revealing a CO breakpoint. Overall, we predicted 1,786 recombinant molecules with a median CO breakpoint resolution (distance between the two flanking markers) of 6.7 kb. However, a comparison with the 400 benchmark COs revealed that only 674 of the molecules overlapped with verified COs, while the remaining 1,112 were putative false positives (FP). Many of the FPs appeared close or within structural rearrangements between the parental genomes, suggesting that the molecule reconstruction using linked-read alignments is vulnerable to unrecognized structural differences between the parental genomes and thereby leads to false predictions of recombinant molecules (Supplementary Note 1).

Making use of chromosome-level assemblies of both parental genomes[36,37], we defined a new marker set composed of ~500,000 SNP markers, which only included SNPs in non-rearranged, co-linear regions between the parental genomes (Supplementary Table 4). We repeated the

analysis with this new marker set and filtered for molecules that appeared recombinant independent of which parental genome was used as reference sequence. Overall, we recovered 558 recombinant molecules, of which 475 (85.1%) CO breakpoints (median resolution 5.3 kb) overlapped with the benchmark COs. The remaining 83 (14.9%) COs were putative FPs most likely due to the co-occurrence of different molecules with the same barcode (Supplementary Note 1). However, we cannot rule out that some of them might have been true recombination events that were missed in the benchmark set (e.g. two closely spaced COs, gene conversion events[31,38] or mitotic COs, which were only present in some cells of the sequenced material).

Increasing the number of molecules per library

To test whether the number of (recombinant) molecules per library would be increased by increasing the DNA loading, we generated two additional 10X libraries from the same DNA pool by loading 0.40 ng (P50L40) and 0.75 ng (P50L75) into the Chromium Controller (Fig. 1; Materials and Methods). We sequenced these libraries with 104.8 and 212.3 million read pairs in anticipation of an increased number of molecules within both libraries (Fig. 2b-d). Following the same analysis as for the first library, the higher DNA loading drastically increased the number of recovered molecules. As compared to the 2.7 million molecules for P50L25, we now found 4.7 and 10.5 million molecules for P50L40 and P50L75 after filtering (Table 1). This also increased the number of recombinant molecules to 1,012 in P50L40 and 2,519 in P50L75 as compared to 558 in P50L25. This, however, had the cost of also increasing the FP rate by 5% in P50L40 and 10% in P50L75 (Table 1).

To investigate the effect of the FPs, we compared the distribution (i.e. recombination landscape) of all 2,519 COs in the library with the highest FP rate, P50L75, with two other landscapes: one calculated only from 1,874 true COs (those overlapping with the 400 benchmark COs) and one calculated only from the remaining 645 false recombinant molecules (Materials and Methods). The recombination landscape calculated for all COs (TP+FP) was nearly identical to the

landscape generated from true COs only (TP), while the distribution of the FPs was highly random (Fig. 3a; Supplementary Fig. 2: *K-S* test, *p*-value=9.6e-01), suggesting that FPs hardly obscure true recombination landscapes. In fact, correlating the sliding window-values of each of the three CO landscapes revealed that the landscape calculated from all CO sites was almost perfectly correlated to the one generated from real COs (Fig. 3b: Pearson's *r* 0.97, *p*-value<2.2e-16), while the frequency of FP along the chromosomes was not correlated to the frequency of real COs and was only marginally correlated to landscape of all COs.

To test the association of COs with genomic features, we checked all of the CO sites for their annotation in the Col-0 reference sequence (Fig. 3c). In comparison to randomly placed CO sites, the 1,874 true COs sites were significantly enriched in promoter regions (permutation test: *p*-value 5.0e-03) and intergenic regions (*p*-value 1.4e-03) and were significantly depleted in gene bodies (*p*-value 6.0e-04) and transposable elements (*p*-value 9.0e-04) recapitulating regional preferences which have been described before[38,39]. When additionally including the FP, the CO set showed the same significant regional associations with one exception: COs were significantly enriched in gene ends. Though this was not true when considering only the true COs, gene ends represented only a minor fraction of all COs. To examine CO associations with genomic features at a finer scale, we analyzed the regional preferences of the CO sites in individual transposable element super families because these were found to have strong differences in CO rates[40]. Though this only included 331 putative CO sites, we found that COs were significantly depleted within *LTR-Gypsy* and *En-Spm* super families and enriched within *Helitron* and *LINE* elements, as has been previously shown[40], whether or not FP were included (Supplementary Figure 3). In addition, the presence of FP did not affect the previously reported relationships between COs and GC content or DNA methylation[38]. The two datasets showed nearly identical correlations between CO frequency and GC content (Fig. 3d: Pearson's *r* -0.46 and -0.45, both *p*-value<2.2e-16). Similarly, COs in both datasets were found in regions with low levels of methylation (Fig. 3e).

Together this suggests that our method is not only effective when facing an increasing amount of FP in libraries with a large number of molecules, but that it is also powerful enough to accurately identify chromosome-wide and local CO patterns.

Increasing the number of genomes per library

Another technical consideration is that many of the recombinant molecules were derived from the same CO breakpoints, given the limited number of distinct COs within the 50-individual pools. For instance, there were only 363 distinct COs recovered by the 1,874 recombinant molecules in P50L75 (Table 1). Even though identifying a single CO breakpoint multiple times can help to increase its resolution and reliability, it does reduce the overall number of distinct COs that can be found with one library. Though the inclusion of more plants (i.e. increasing the number of genomes in the pool) does not increase the number of recombinant molecules, it does increase the number of *independent* recombinant molecules and thereby maximizes the number of distinct CO events found with one library.

To test the effect of genome number on the number of distinct COs, we applied our method to pools of 200 and 1,250 F_2 plants. We generated two individual 10X libraries each for a pool of 200 (P200R1, P200R2) and a pool of 1250 (P1250R1 and P1250R2) different F_2 plants (Fig. 1; Supplementary Figures 1 and 4; Materials and Methods). These libraries were sequenced with 168.2 – 194.4 million read pairs (Supplementary Table 1) and revealed 12.3 – 16.0 million molecules. For P200R1 and P200R2, we found 2,538 and 2,334 recombinant molecules, which identified 1,779 (69.5%) and 1,606 (68.8%) distinct CO sites. For the pool of 1,250 plants, there were 2,880 and 3,430 recombinant molecules, which uncovered 2,386 (82.8%) and 2,788 (81.3%) distinct CO sites. Comparison of these COs with 1,840 COs determined using individual whole genome sequencing for an independent set of Col x L*er* F_2s[20] showed consistent genome-wide patterns (Supplementary Figure 5). While the larger pools revealed more distinct CO sites, there

were still nearly 20% COs overlapping, but it does not necessarily imply that these are not unique CO events. As the genome size of *A. thaliana* is relatively small (~135 Mb[41]), independent COs have an unneglectable probability to overlap (Supplementary Note 2; Supplementary Table 5). According to simulations, given a pool of 1,250 F_2s, ~15% of the independent CO events are expected to overlap (Supplementary Table 5 and 6) making up for large parts of the overlapping recombinant molecules in the 1,250 pools.

Estimating relative recombination frequency

Since the probability to identify a CO breakpoint within a molecule depends on the average recombination frequency among the pooled genomes, it might be possible to estimate the average recombination frequency from the fraction of observed recombinant molecules. However, the probability to identify a CO breakpoint also depends on the length and sequencing coverage of the molecules within a library. It is therefore only meaningful to calculate an average recombination frequency that is relative to the underlying molecule characteristics (*relative recombination frequency*). To test for significant differences in relative recombination frequencies between libraries, it is essential that they have close-to-identical molecule characteristics. For libraries with differences in molecule length and sequencing coverage distributions, subsampling can be used to generate identical distributions. Once the distributions are similar, the relative recombination frequency can be determined as the number of recombinant molecules per million molecules (C^M). Repeated subsampling even allows for the calculation of confidence intervals for C^M values and thereby allows for testing significant differences between the average recombination frequencies of different pools.

First, we compared the CO frequencies of P200R1 and P200R2, which were two independent libraries generated from the same DNA (Fig. 1; Fig. 2b-d; Materials and Methods). After subsampling (Supplementary Figure 6), there was no significant difference in C^M values (26.7±0.4 and 26.8±0.4), as expected for these libraries (Table 2). We repeated this comparison for

P1250R1 and P1250R2, which differed greatly in their original molecule characteristics (Fig. 2b-d). After subsampling to the same molecule distributions as for the smaller pools (Supplementary Figure 6), the C^M values of 27.2±0.6 and 27.1±0.7 were also not significantly different from each other (Table 2). Moreover, when comparing the 200 and 1,250 recombinant pools, we also found no significant difference between any of the pools (all confidence intervals overlapped), which is expected, given that all libraries were generated from individuals of the same F_2 population. Together, these results show that C^M values are stable against differences in the molecule characteristics and pool sizes and allow for determining and comparing average recombination frequencies between samples.

Estimating recombination frequency and landscapes from gametes

Crosses between divergent strains provide the simplest opportunity for determining COs, but typically require the generation of a recombinant population after a single round of meiosis in a hybrid context. Observing COs directly in the products of meiosis (gametes) would greatly expedite the study of recombination, especially in inbred species with long generation times. To test how our method performs on recombinant gametes, we extracted HMW DNA from Col x L*er* F_1 hybrid pollen, created a 10X library (P8000) and sequenced it with 319.9 million read pairs (Fig.1; Supplementary Figure 1f; Supplementary Table 1). Following the same analysis as for the F_2 pools, we identified 20 million molecules with an average molecule base coverage of 0.26x (Fig. 2b-d). Among those, there were 3,246 recombinant molecules with a median CO breakpoint resolution of 8.0 kb.

We compared relative recombination frequencies in pollen with the P1250R1 and P1250R2 pools after re-sampling molecules to obtain comparable characteristics (Supplementary Figure 6; Materials and Methods). The CO frequency estimate (C^M) for pollen was significantly higher than for F_2s (Table 2), consistent with the higher male recombination rate in this species[42].

The distributions of the CO breakpoints in pollen and F_2s were highly correlated in most regions of the genome (Pearson's *r* 0.77~0.83, both *p*-value<2.2e-16), but exhibited a few local differences (Fig. 4). The increased COs in pollen therefore followed a similar genome-wide distribution as in F_2s, as the regions with locally enriched COs in pollen accounted for only a minor fraction of the all COs.

We further compared the recombination landscape of our pollen sample to a previously published recombination landscape from a Col-0 x L*er* backcross population, where all COs were derived from male meiosis[42]. This data set, consisting of 7,418 COs at an average resolution of ~320 kb, was generated from 1,505 individuals genotyped with 380 SNP markers that were evenly spaced throughout the genomes. After binning our CO data into the same windows, we found that the recombination landscapes were broadly similar (Supplementary Figure 7: Pearson's *r* 0.41, *p*-value<2.2e-16). Unexpected differences, however, were observed mostly at the end of chromosomes. COs formed by male meiosis and identified in recombinant BCF_1 plants showed significant increases at the end of some chromosomes[42], while most of these increases were not observed in pollen data.

In general, this shows that a single sequencing library is sufficient to derive genome-wide CO patterns from pooled gametes without the need for producing any recombinant individuals. Such CO maps have the advantage to measure sex-specific meiotic recombination without doing any backcrosses and are less influenced by post-meiotic biases.

DISCUSSION

For the past century, the study of recombination has relied on inferring or determining the genotypes at (limited) marker positions along the chromosomes of recombinant individuals. Here, we developed an efficient and accurate method for detecting recombination breakpoints in pooled DNA using linked-read sequencing. This method can be applied directly to gametes, avoiding the

need for generating or genotyping recombinant populations. In turn, this simplification allows for the production of multiple different genome-wide CO maps with relatively little effort. Thus, it is now feasible to directly compare CO maps generated from many different genotypes or environments. Long-read sequencing or single-cell DNA-seq of individual pollen would be two alternative methods, but the current costs of library preparation and sequencing are prohibitive for comparing multiple samples.

The small genome size of the species we used in this study, *A. thaliana,* was potentially an obstacle to accurately scoring CO events using linked-read sequencing, as this technology assigns the same barcode to a small number of individual DNA molecules that are tens of kb in length. Since this process is random, two molecules that originate from different genomes in the pool that happen to align near each other can receive the same barcode. The probability of these “collision” events is much higher in small genomes and can mimic COs if the two molecules have different parental genotypes, giving rise to false positives (FP). However, the FP had little effect on the overall accuracy of the CO predictions (Fig. 3). Even though the FP rate did not pose a substantial problem, making several libraries using a reduced input would result in high molecule numbers at the lower FP rate. This would incur higher costs of library production while maintaining the same sequencing cost. Moreover, when applying this method to larger genomes, the chance of “collision” of two molecules with the same barcode would be lower, leading to a lower likelihood of FPs.

The spatial distributions of COs identified with our method recapitulated known recombination landscapes and precisely co-localized with genomic and epigenetic features that have been reported to be associated with meiotic recombination[34,37,39]. The resolution of the CO breakpoints was very precise, with the median CO interval less than 10 kb in each of the samples. This accuracy, however, could only be achieved by filtering the marker list for co-linear regions between the parental genomes as genomic rearrangements between the parents were hotspots for FPs.

The careful validation studies of our method using F_2s provided the basis for determining the COs directly in gametes. Linked-read sequencing of bulk F_1 pollen led to the discovery of ~3,500 CO events that showed a similar distribution as was recently published for male meiosis (Supplementary Figure 7). Dreissig et al. previously assessed the recombination rate in barley pollen using whole-genome sequencing of individual pollen nuclei[46]. While this allowed for the analysis of CO interference, their method employed a more technically challenging library preparation that does not scale to high numbers and achieved a much lower CO resolution.

Interrogating COs in gametes (pollen, in the case of plants or sperm in animals) has many advantages over current approaches. It avoids the laborious and time-consuming step of growing/rearing recombinant populations. It reduces the number of generations required, greatly facilitating recombination studies in inbred organisms with long generation times. Because entire recombination landscapes can be generated from single libraries, a recombination landscape can now be studied as a single trait. Multiple CO landscapes that can be replicated either in the same or different genetic backgrounds and environments and compared in a single study. In consequence, this now allows for a more complex and sophisticated experimental design to test hypotheses regarding the regulation of recombination.

Materials and Methods

F_2 DNA extraction and library preparation

F_2 seeds from Col-0 x L*er*-0 were stratified for 7 days at 4°C, sown on soil in 24-pot trays, and grown under 16 h light, 8 h dark cycles at 20°C for three weeks. DNA pools of up to 1,250 F_2 individuals were constructed (Fig. 1). HMW DNA was extracted from pools with 50 distinct F_2 plants using a published protocol[35]. One 50-F_2 pool was selected for validating the method, for which DNA was also extracted from each individual ($a_{1\text{-}50}$) and WGS libraries were prepared using the Illumina

298 DNA TruSeq protocol. The DNA of four 50-F_2 pools with similar fragment size distribution
299 (Supplementary Figure 4a) were merged based on equal concentration, from which replicates
300 P200R1 and P200R2 were obtained. In addition, 25 50-F_2 pools with divergent fragment size
301 distributions (Supplementary Figure 4b) were merged based on equal molarity of molecules
302 between 42-70 kb according to FEMTOpulse, AATI genomic DNA quality check, and the resultant
303 DNA was used for replicates P1250R1 and P1250R2.

304 Size selection was performed on each DNA pool using the Sage Science BluePippin high
305 pass protocol (i.e., 0.75% Agarose Dye-Free / 0.75% DF Marker U1 high-pass 30-40 kb vs3) with
306 starting point at 40 kb to ensure that most molecules were over 40 kb (Supplementary Figure 1a-e).
307 For P200R1/2 and P1250R1/2, the size-selected DNA was evaluated using the Qubit fluorometer
308 and TapeStation analyzer. After selection, for each library 0.75 ng DNA were loaded into the 10X
309 Chromium controller (Supplementary Table 1). P50 was subjected to the same quality control
310 measures. For this library three different amounts of DNA (0.25 ng, 0.40 ng and 0.75 ng) were
311 loaded into the 10X Chromium Controller, generating the P50L25, P50L40, P50L75 libraries.

312 **Pollen DNA extraction and library preparation**

313 Col *CEN3 qrt 420* and L*er* seeds were stratified for 4-7 days at 4°C before sowing on soil in
314 18-pot trays and growing under 16 h light, 8 h dark cycles at 20°C until flowering. Pollen from a
315 single L*er* plant was used to pollinate a single Col *CEN3 qrt 420* stigma to generate F_1 plants.
316 Inflorescences were collected from a single F_1 plant and ground in 1 mL of 10% sucrose using a
317 mortar and pestle. 9 mL of 10% were added to the mortar slurry and the resulting homogenate was
318 mixed by pipetting up and down with a wide bore 1 mL pipet tip and filtered through an 80-μM nylon
319 mesh before centrifugation at 350 x g for 10 min at 4°C. The supernatant was discarded and the
320 pollen pellet was washed two times with 10% sucrose. The pellet was resuspended in four volumes
321 of lysis buffer (100 mM NaCl, 50 mM Tris–HCl (pH 8)), 1 mM EDTA, 1% SDS and proteinase K was

added to achieve a concentration of 20 µg/mL. Five to ten 2-mm glass beads were added to the sample and it was vortexed at full speed for 30 s. To check for pollen disruption, a 1-µL sample was removed and mixed with 10 µL of lysis buffer and examined under a microscope. The sample was then vortexed two more times for 30s until 90% of the pollen grains had been disrupted. An equal volume of Tris-saturated phenol was added and the sample was placed on a rotating wheel at room temperature for 30 min. After centrifuging at 15,000 x g for 10 min, the supernatant was transferred to a new tube and mixed with an equal volume of 24:1 chloroform:isoamylalcohol and homogenized by shaking the tube. The tube was centrifuged again at 15,000 x g for 10 min and the supernatant was mixed with 0.7 volumes isopropanol in a new tube and inverted gently. After another centrifugation step at 15,000 x g for 10 min, the supernatant was discarded and the pellet was washed with 1 mL of 70% EtOH. After a final centrifugation at 15,000 x g for 2 min, the supernatant was discarded and the DNA pellet was allowed to dry at room temperature. The pellet was resuspended in 50 µL of 10 mM Tris-HCl, pH 8 with 0.1 mM EDTA and stored at 4°C.

DNA was analyzed by field inversion gel electrophoresis to confirm that most molecules were around 48 kb (Supplementary Figure 1f) before preparing the linked-read library (load: 1.00 ng) using the Chromium Genome Reagent Kit, the Chromium Genome Library Kit & Gel Bead Kit, Chromium Genome Chip Kit v2, and Chromium i7 Multiplex Kit according to the manufacturer's instructions. As 1.00 ng is equivalent to 10^6 Mb, the haploid genome size is ~135 Mb[41], and each pollen grain is composed of three cells, there were ~3,000 pollen grains expected with unique recombinant genomes.

In total, eight 10X and 50 standard Illumina libraries were sequenced on HiSeq 3000/4000 with 151 bp paired-end reads, where the individual F_2s were at ~5x and the pools of F_2s and pollen nuclei were at 173-659x (Supplementary Table 2).

Molecule recovery using linked-read alignments

Col-0 and L*er* reference genomes were indexed using the function *mkref* of *longranger* (v2.2.2 10X Genomics). Linked-reads of each sample were aligned against the Col-0 and L*er* reference genomes separately using *longranger wgs* with options: --id=ALIGN-ID --reference=MKREF-INDX-FOLDER --fastqs=READ-PATH --sample=READ-ID --localcores=40 --localmem=192 --noloupe --sex=male --vcmode=freebayes --library=LIBRARY-ID and all other options under default settings (note that option --sex was set as “male”, which is required by the tool but not affecting the alignment here). Read alignments with the same read barcodes were clustered to molecules according to their genomic location ensuring that neighboring read alignments were not more than 25 kb away from each other. The resultant molecules (over 1 kb) were further filtered for falsely merged molecules by removing very long or densely covered molecules (Supplementary Note 1).

Generating the CO benchmark set

For each individually sequenced F_2, read alignments against the Arabidopsis Col-0 reference sequence[36] and variant calling were performed using *Bowtie2*[43] (version 2.2.8) and *SAMtools/BCFtools*[44] (version 1.3.1). *TIGER*[31] was used to reconstruct the parental haplotypes using SNPs in co-linear regions between the Col-0 and L*er*[37] genomes. We determined an average of 8.3 COs per diploid genome, totaling 415 COs with a median breakpoint resolution of 0.5 kb (Supplementary Table 2). Of those, 15 COs were located in regions with an inter-marker distance of more than 10 kb. As local CO breakpoint identification relies on markers near the breakpoint, we excluded these COs, generating a final benchmark set of 400 COs.

Estimating chromosomal CO distributions (recombination landscape)

Recombination landscapes were estimated using sliding windows along each chromosome (window size 1 Mb, step size 50 kb). Within a given window, the CO frequency f was calculated by $f = n / t_i$, where n is the number of COs in the window and t_i is the total CO number within the respective chromosome i across each of the focal libraries. This ensures that different CO landscapes can be compared to each other within the given chromosome.

Relative CO frequency estimation

For frequency estimation, we first subsampled molecules from the pools of interest by intersecting their molecule characteristic distributions. Using the molecule length distributions as the basis (Supplementary Figure 6a), the molecules were randomly sampled down within 1 kb large bins. For each bin, the number of randomly selected molecules per pool was equal to the lowest number of molecules across all pools within this specific bin. To increase the subsampling rate, 80% of the sampled molecules were further randomly selected within each bin. After this second subsampling, the read and base coverage distributions were also highly similar (Supplementary Figure 6b-c). Final molecule filtering was applied following the same strategy as used for the complete sets of molecules, by applying 30 kb as size and 24 as read number thresholds.

Then, COs per million molecules (C^M) was calculated for each pool. The process of molecule sub-sampling, CO identification and C^M calculation was repeated fifty times for each pool. The means (μ) of all 50 C^M values and their 95% confidence intervals ($\mu \pm s$, where s is 2.0096*σ/50^0.5 with σ being the standard deviation) were used to assess significant differences between samples.

DNA methylation level estimation

Within a recent study, we have assessed DNA methylation for *A. thaliana* Col-0[45]. Methylation level M for a CO or a random interval was calculated by $M = N_{met} / N$, where N_{met} is the number of

reads supporting methylated cytosines at all C and G sites, while *N* is the total number of reads at these sites.

Data availability

Read data of all eight 10X linked-read libraries (ERS2851779-ERS2851786) and 50 whole genome sequencing libraries (ERS2851943-ERS2851992) are available in BAM format from European Nucleotide Archive under accession number ERP11158.

Acknowledgements

The authors would like to thank Ian R. Henderson (Department of Plant Sciences, University of Cambridge) for providing the CO breakpoint lists, Erik Wijnker (Wageningen University) for providing seeds and Ulrike Hümann, Manish Goel, Wen-Biao Jiao, Vidya Oruganti, and Onur Dogan (Max Planck Institute for Plant Breeding Research) for their help in the greenhouse. We also would like to thank 10X Genomics for their help on setting up the *longranger* software and kind donation of library reagents to support the development of recombination identification.

Author contributions

H.S., B.A.R. and K.S. designed the project. H.S. and B.A.R. performed all analysis. B.A.R., H.S. and P.J.F. prepared the samples. B.A.R., R.B., J.F. and B.H. performed DNA extraction, quality control, library preparation and sequencing. K.S., B.A.R., A.M.H. and R.M.W. supervised the project. H.S., B.A.R. and K.S. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was supported by a Max Planck Society postdoctoral fellowship and a UC Davis Genome Center Pilot Project grant.

Competing Interests

None.

References

1. Barton, N. H. & Charlesworth, B. Why sex and recombination? *Science* **281,** 1986-1990 (1998).
2. Rice, W. R. & Chippindale, A. K. Sexual recombination and the power of natural selection. *Science* **294,** 555-559 (2001).
3. McDonald, M. J., Rice, D. P. & Desai, M. M. Sex speeds adaptation by altering the dynamics of molecular evolution. *Nature* **531,** 233-236 (2016).
4. Mancera, E., Bourgon, R., Brozzi, A., Huber, W. & Steinmetz, L. M. High-resolution mapping of meiotic crossovers and non-crossovers in *yeast*. *Nature* **454,** 479-485 (2008).
5. Kauppi, L., Jeffreys, A. J. & Keeney, S. Where the crossovers are: recombination distributions in mammals. *Nat. Rev. Genet.* **5,** 413-424 (2004).
6. Yamada, S. *et al.* Genomic and chromatin features shaping meiotic double-strand break formation and repair in mice. *Cell Cycle* **16,** 1870-1884 (2017).
7. Singhal, S. *et al.* Stable recombination hotspots in birds. *Science* **350,** 928-932 (2015).
8. Miller, D. E. *et al.* Whole-Genome Analysis of Individual Meiotic Events in Drosophila melanogaster Reveals That Noncrossover Gene Conversions Are Insensitive to Interference and the Centromere Effect. *Genetics* **203,** 159-171 (2016).
9. Kianian, P. M. A. *et al.* High-resolution crossover mapping reveals similarities and differences of male and female recombination in *maize*. *Nat. Commun.* **9,** 2370 (2018).
10. Salomé, P. A. *et al.* The recombination landscape in *Arabidopsis thaliana* F2 populations. *Heredity* **108,** 447-455 (2012).
11. Higgins, J. D., Osman, K., Jones, G. H. & Franklin, F. C. H. Factors underlying restricted crossover localization in barley meiosis. *Annu. Rev. Genet.* **48,** 29-47 (2014).

12. Sturtevant, A. H. A crossover reducer in *Drosophila melanogaster* due to inversion of a section of the third chromosome. *Biol. Zent. Bl.* **46,** 697-702 (1926).
13. Dobzhansky, T. The Decrease of Crossing-Over Observed in Translocations, and Its Probable Explanation. *Am. Nat.* **65,** 214–232 (1931).
14. Chakraborty, U. & Alani, E. Understanding how mismatch repair proteins participate in the repair/anti-recombination decision. *FEMS Yeast Res.* **16,** (2016).
15. Choi, K. *et al. Arabidopsis* meiotic crossover hot spots overlap with *H2A.Z* nucleosomes at gene promoters. *Nat. Genet.* **45,** 1327-1336 (2013).
16. Underwood, C. J. *et al.* Epigenetic activation of meiotic recombination near *Arabidopsis thaliana* centromeres via loss of H3K9me2 and non-CG DNA methylation. *Genome Research* (2018). doi:10.1101/gr.227116.117.
17. Yelina, N. E. *et al.* Epigenetic remodeling of meiotic crossover frequency in *Arabidopsis thaliana* DNA methyltransferase mutants. *PLoS Genet.* **8,** e1002844 (2012).
18. Marand, A. P. *et al.* Meiotic crossovers are associated with open chromatin and enriched with Stowaway transposons in potato. *Genome Biol.* **18,** 203 (2017).
19. Ziolkowski, P. A. *et al.* Natural variation and dosage of the HEI10 meiotic E3 ligase control *Arabidopsis* crossover recombination. *Genes Dev.* **31,** 306-317 (2017).
20. Serra, H., Lambing, C., Griffin, C.H., Topp, S.D., Nageswaran D.C., Underwood C.J., Ziolkowski, P.A., Séguéla-Arnaud, M., Fernandes, J.B., Mercier R. and Henderson I. Massive crossover elevation via combination of *HEI10* and *recq4a recq4b* during *Arabidopsis* meiosis. *PNAS* **115**, 2437-2442 (2018).
21. Fernandes, J. B., Seguéla-Arnaud, M., Larchevêque, C., Lloyd, A. H. & Mercier, R. Unleashing meiotic crossovers in hybrid plants. *PNAS* **115**, 2431-2436 (2017).
22. Girard, C. *et al.* FANCM-associated proteins *MHF1* and *MHF2*, but not the other Fanconi anemia factors, limit meiotic crossovers. *Nucleic Acids Res.* **42,** 9087-9095 (2014).

23. Crismani, W. *et al.* FANCM limits meiotic crossovers. *Science* **336,** 1588–1590 (2012).
24. Higgins, J. D. *et al. AtMSH5* partners *AtMSH4* in the class I meiotic crossover pathway in *Arabidopsis thaliana*, but is not required for synapsis. *Plant J.* **55,** 28-39 (2008).
25. Berchowitz, L. E., Francis, K. E., Bey, A. L. & Copenhaver, G. P. The role of *AtMUS81* in interference-insensitive crossovers in *A. thaliana*. *PLoS Genet.* **3,** e132 (2007).
26. De Storme, N. & Geelen, D. The impact of environmental stress on male reproductive development in plants: biological processes and molecular mechanisms. *Plant Cell Environ.* **37,** 1-18 (2014).
27. Jackson, S., Nielsen, D. M. & Singh, N. D. Increased exposure to acute thermal stress is associated with a non-linear increase in recombination frequency and an independent linear decrease in fitness in *Drosophila*. *BMC Evol. Biol.* **15,** 175 (2015).
28. Konieczny, A., and F. M. Ausubel. A procedure for mapping Arabidopsis mutations using co-dominant ecotype-specific PCR-based markers. *Plant J.* **4**, 403-410 (1993).
29. Baird, N.A., Etter, P.D., Atwood, T.S. *et al.* Rapid SNP discovery and genetic mapping using sequenced RAD markers. *PLoS One* **3**, e3376 (2008).
30. Rowan, B.A., Danelle, K.S., Eunyoung, C., Derek. S.L. and D. Weigel. Methods for Genotyping-by-Sequencing. *Methods in Molecular Biology* **1492**, 221-42 (2017).
31. Rowan, B. A., Patel, V., Weigel, D. & Schneeberger, K. Rapid and inexpensive whole-genome genotyping-by-sequencing for crossover localization and fine-scale genetic mapping. *G3* **5,** 385-398 (2015).
32. Melamed-Bessudo, C., Yehuda, E., Stuitje, A. R. & Levy, A. A. A new seed-based assay for meiotic recombination in *Arabidopsis thaliana*. *Plant J.* **43,** 458-466 (2005).
33. Francis, K. E. *et al.* Pollen tetrad-based visual assay for meiotic recombination in *Arabidopsis*. *PNAS* **104,** 3913-3918 (2007).
34. Choi, K., Yelina, N. E., Serra, H. & Henderson, I. R. Quantification and Sequencing of Crossover

Recombinant Molecules from *Arabidopsis* Pollen DNA. in *Haplotyping: Methods and Protocols* (eds. Tiemann-Boege, I. & Betancourt, A.) **1551**, 23-57 (Springer New York, 2017).

35. Mayjonade, B., Gouzy, J., Donnadieu, C., Pouilly, N., Marande, W., Callot, C., Langlade, N., Muños, S. Extraction of high-molecular-weight genomic DNA for long-read sequencing of single molecules. *Biotechniques* **61**, 203-205 (2016).
36. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* **408**, 796-815 (2000).
37. Zapata, L., Ding, J., Willing, E.M., Hartwig, B., Bezdan, D., Jiao, W.B., Patel, V., Velikkakam, J.G., Koornneef, M., Ossowski, S., Schneeberger, K. Chromosome-level assembly of *Arabidopsis thaliana* L*er* reveals the extent of translocation and inversion polymorphisms. *PNAS* **113**, E4052-60 (2016).
38. Wijnker, E., James, G.V., Ding, J., Becker, F., Klasen, J.K., Rawat, V., Rowan, B.A., de Jong, D.F., de Snoo, C.B., and Zapata, L. et al. The Genomic Landscape of Meiotic Crossovers and Gene Conversions in *Arabidopsis Thaliana*. *Elife* **2**, e01426 (2013).
39. Shilo, S., Melamed-Bessudo, C., Dorone, Y., Barkai, N., Levy, AA. DNA Crossover Motifs Associated with Epigenetic Modifications Delineate Open Chromatin Regions in *Arabidopsis*. *Plant Cell* **27**, 2427-36 (2015).
40. Choi, K., Zhao, X., Tock, A. J., Lambing, C., Underwood, C.J., Hardcastle, T.J., Serra, H., Kim, J., Cho, H. S., Kim, J. et al. Nucleosomes and DNA Methylation Shape Meiotic, DSB Frequency in *Arabidopsis Thaliana* Transposons and Gene Regulatory Regions. *Genome Research* **28**, 532-546 (2018).
41. Sun, H., Ding, J., Piednoël, M., Schneeberger, K. *findGSE*: estimating genome size variation within human and Arabidopsis using *k*-mer frequencies. *Bioinformatics* **34**, 550-557 (2018).
42. Giraut, L., Falque, M., Drouaud, J., Pereira, L., Martin, O.C., Mézard, C. Genome-wide crossover distribution in *Arabidopsis thaliana* meiosis reveals sex-specific patterns along

chromosomes. *PLoS Genet.* **7**, e1002354 (2011).

43. Langmead, B. and Salzberg, S. L. Fast gapped-read alignment with *Bowtie* 2. *Nat Methods* **9**, 357-9, (2012).
44. Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, H., Marth, G., Abecasis, G., Durbin, R., 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and *SAMtools*. *Bioinformatics* **25**, 2078-9 (2009).
45. Willing, E. M., Rawat, V. , Mandáková, T., Maumus, F., James, J. V., Nordström, K. J., Becker, C., Warthmann, N., Chica, C., Szarzynska, B. et al. Genome expansion of *Arabis alpina* linked with retrotransposition and reduced symmetric DNA methylation. *Nature plants* **1**, 14023 (2015). doi: 10.1038/nplants.2014.23.
46. Dreissig, S., Fuchs, J., Himmelbach, A., Mascher, M. and Houben, A. Sequencing of Single Pollen Nuclei Reveals Meiotic Recombination Events at Megabase Resolution and Circumvents Segregation Distortion Caused by Postmeiotic Processes. *Front Plant Sci.*. **8**, 1620 (2017).

522 **Table 1. Effect of increasing the molecule number of increased DNA loading**

Pool	P50L25	P50L40	P50L75
Raw molecules (≥1kb)[C]	3,577,104	5,804,559	12,551,833
Raw molecules (≥1kb)[L]	3,462,796	5,631,294	12,240,762
Filtered molecules[C]	2,664,977	4,783,828	10,651,147
Filtered molecules[L]	2,565,403	4,642,515	10,417,489
Total CO molecules[CL]	558	1,012	2,519
TP CO molecules[CL]	475 (85.1%)	804 (79.4%)	1,874 (74.4%)
TP unique (non-redundant) COs[CL]	254 (63.5%)	325 (81.3%)	363 (90.8%)
FP molecules[CL]	83 (14.9%)	208 (20.6%)	645 (25.6%)

523 *TP, true positive; FP, false positive
524 *C: number based on Col-0 reference genome[36]
525 *L: number based on L*er* reference genome[37]
526 *CL: number based on intersection of two sets of predictions (of C and L)

Table 2 Relative recombination frequency in F_2 and pollen pools

Pool		Molecules[a]	CO molecules[b]	C^M
F_2	P200R1	4,845,588±1,556	129±7	26.7±0.4
	P200R2	4,856,886±1,371	130±7	26.8±0.4
	P1250R1	4,650,420±1,662	126±9	27.2±0.6
	P1250R2	4,309,936±1,454	117±10	27.1±0.7
Pollen	P8000	4,902,994±1,722	212±10	43.2±0.6

*Note: for ***a*** and ***b***, the values are $\mu \pm \sigma$, where μ is the mean and σ is the standard deviation of the molecule numbers in 50 random sub-samplings. For C^M, the values are $\mu \pm s$ giving 95% confidence intervals for the mean μ of C^M, where s is 2.0096*σ/50^0.5 with σ being the standard deviation.

Figure legends

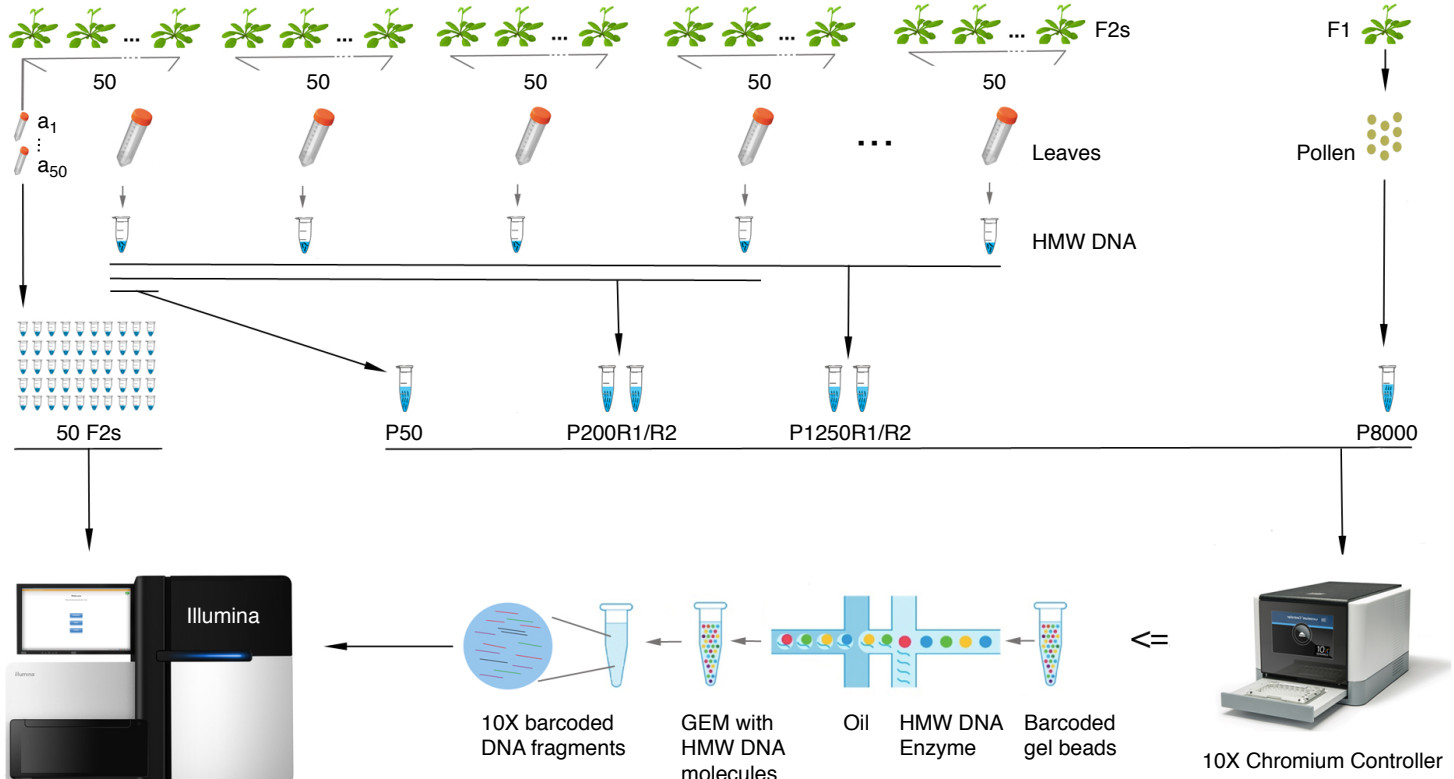
Figure 1. Experimental design and CO detection using linked-reads.

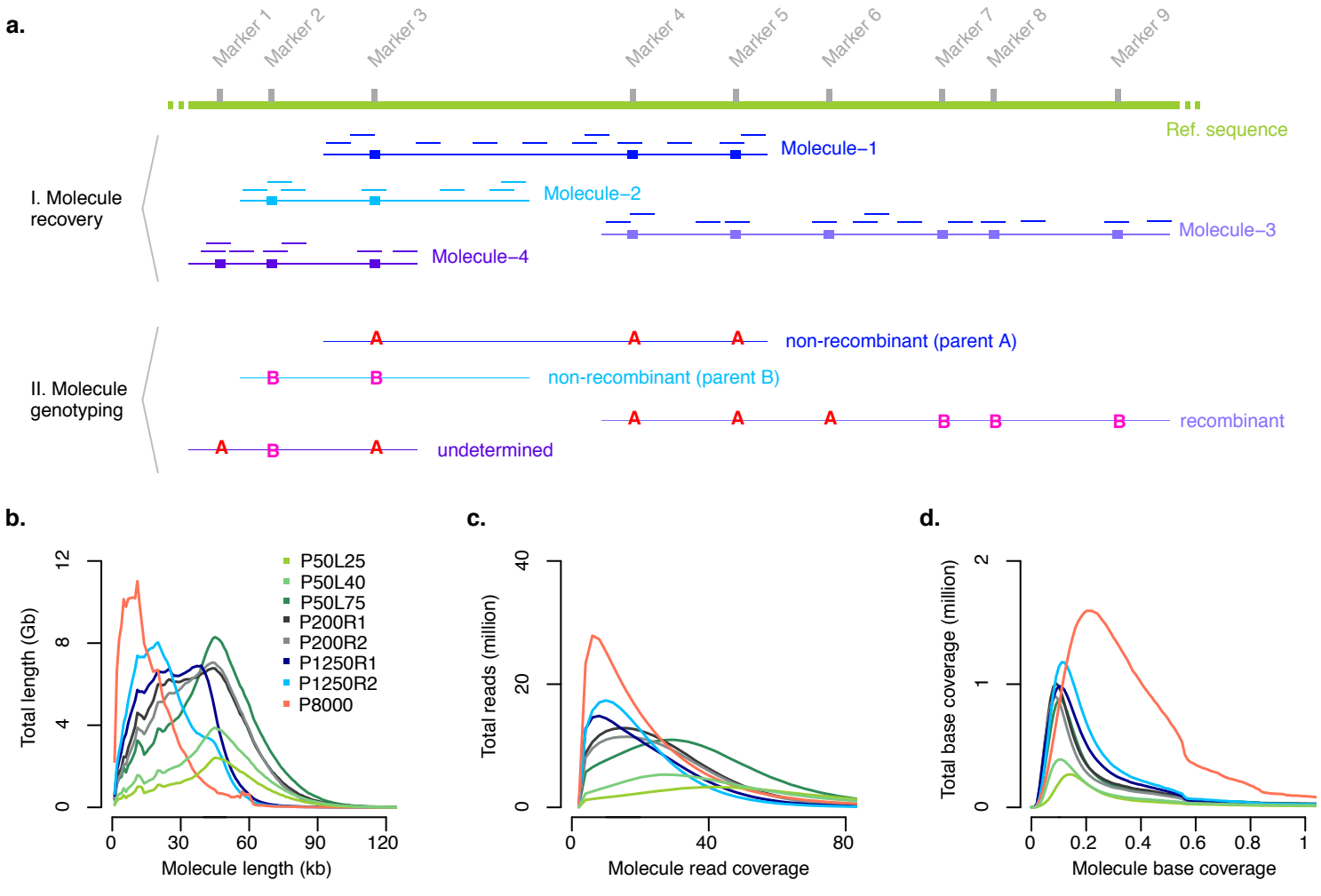
F_1 and F_2 plants were derived from crosses of two divergent *A. thaliana* accessions. *Leaf sampling*: Leaves of 50 selected F_2s were individually sampled ($a_{1\sim50}$). In addition, the leaves of all other plants were pooled in batches of 50 F_2s (for further merging), where the 50 individually sampled plants formed one of the pools. Also, pollen from a single F_1 plant was sampled. *DNA extraction*: DNA from individual the samples $a_{1\sim50}$ were extracted for Illumina whole-genome sequencing, while the DNA of the 50-F_2 and pollen samples were extracted with a protocol for high molecular weight DNA (see Materials and Methods and Mayjonade *et al.* (2016)[35]). *Pooling*: Four and twenty-five 50-F_2 samples were merged leading to 200-F_2 and 1,250-F_2 pools (P200 and P1250), each with two replicates (R1 and R2). The 50-F_2 pool composed of the 50 F_2s that were individually sequenced was labelled as P50, while the pollen pool was labelled as P8000. *Library preparation*: 50 individual F_2 DNA samples were used for preparing Illumina DNA TruSeq libraries. The P50 DNA sample was loaded into a 10X Chromium Controller (illustration modified from 10X Genomics) with three different amounts of DNA including 0.25 ng, 0.40 ng and 0.75 ng (P50L25, P50L40 and P50L75). P200R1/R2 and P1250R1/R2 were loaded using 0.75 ng and P8000 was loaded using 1.00 ng. *Sequencing*: All libraries were sequenced on Illumina HiSeq3000/4000 sequencers.

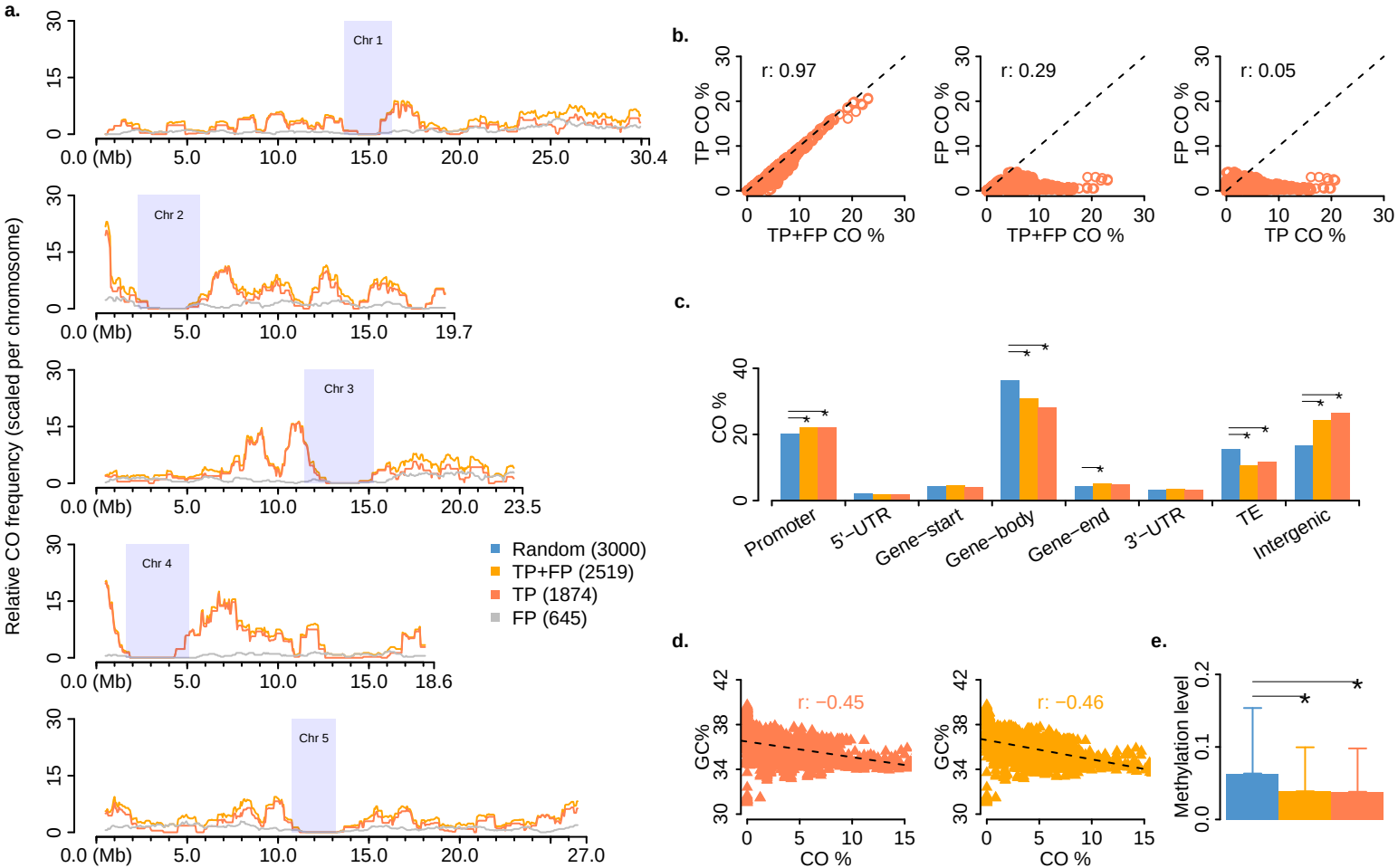
Figure 2. Molecule recovery and genotyping, and real molecule characteristics. *a*. Reads with the same barcode (represented by short lines in matching colors) that were aligned within close proximity (less than 25 kb apart) have been connected to recover molecules (M-1 to M-4). The genotypes that can be assessed with the read alignments of each recovered molecule fall into three major categories, non-recombinant (M-1, M-2), recombinant (M-3) and undetermined (M-4). ***b.*** Length of recovered molecules for each of the 10X libraries. ***c.*** Number of reads per recovered molecule. ***d.*** Recovered molecule base coverage.

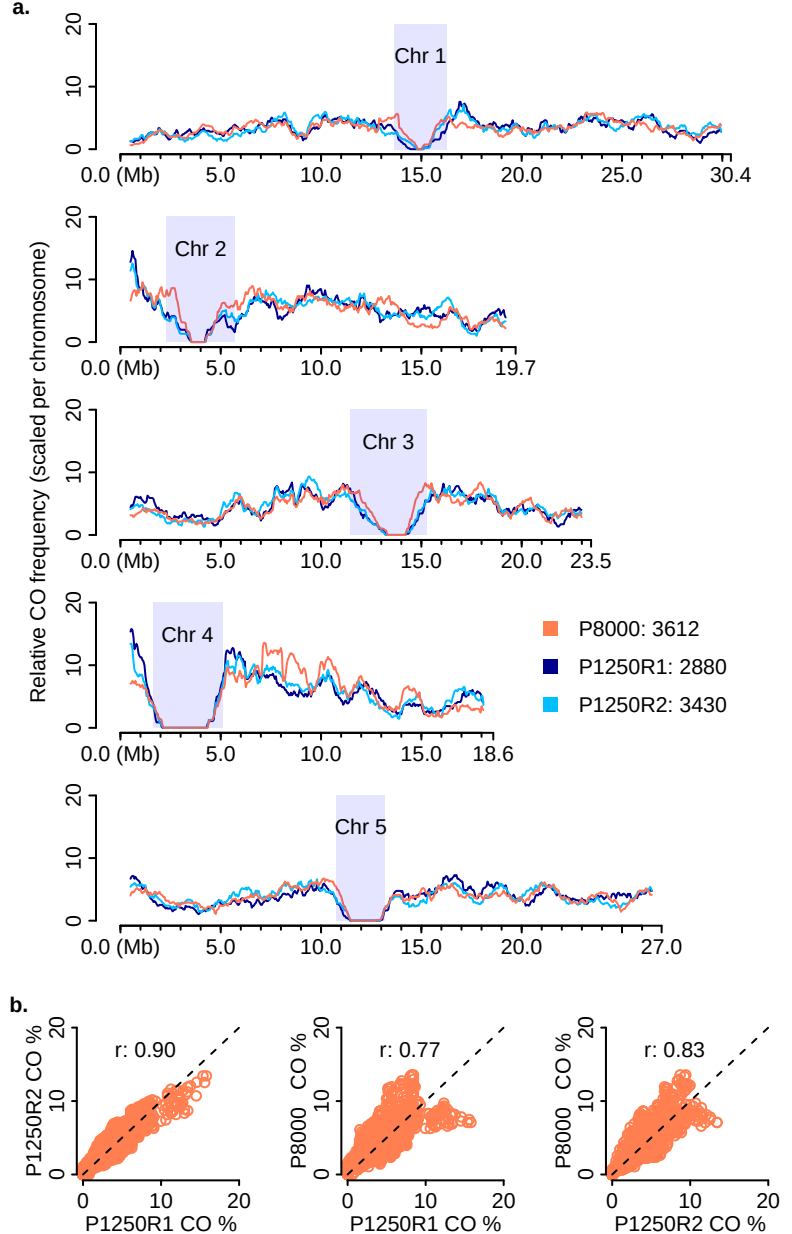
Figure 3. Genome-wide CO landscape formation and feature association. ***a.*** Sliding window-based (window size 1 Mb, step size 50 kb) recombination landscapes calculated for true positives (TP), false positives (FP) and the combined set of TP+FP recombinant molecules of P50L75. ***b.*** Correlation of regional CO frequencies comparing TP *v.s.* TP+FP, FP *v.s.* TP+FP and FP *v.s.* TP. ***c.*** Association of the TP and TP+FP sets with different genomic features in contrast to the random expectation. ***d.*** Correlation of regional GC-content with CO frequency within the TP and TP+FP sets. ***e.*** Association of TP or TP+FP sets with DNA methylation, compared with a random expectation (Materials and Methods). Error bars reflect the standard deviation with respective to the mean of methylation of a set of CO intervals. Note: in *c* and *e*, the expectations were obtained based on randomly sampled intervals within the reference genome excluding heterochromatic regions. The middle position of a CO interval was used for associations with genomic features in *c*, where the promoter of a gene was defined as the 1,000 bp region upstream of the transcription start site, and the gene start/end as the first/last 200 bp of a gene. For each feature, a permutation test (*oneway_test* in *R* "coin" package) was performed between the TP set (or TP+FP set) with 1,000 sets of randomly sampled 3,000 intervals. The asterisks indicate the observed values (either from the TP or the TP+FP set) were significantly different from a random expectation.

Figure 4. Genome-wide recombination landscape in pollen. ***a.*** Comparison of sliding window-based (window size 1 Mb, step size 50 kb) recombination landscapes in pollen and F_2 populations (P1250R1/R2). ***b.*** Correlation of genome-wide CO frequencies in pollen and F_2s (Pearson's *r* 0.77~0.83, both *p*-value<2.2e-16).


F2s
F1
50
50
50
50
50
a_1
a_{50}
Leaves
Pollen
HMW DNA
50 F2s
P50
P200R1/R2
P1250R1/R2
P8000
Illumina
10X barcoded DNA fragments
GEM with HMW DNA molecules
Oil
HMW DNA Enzyme
Barcoded gel beads
<=
10X Chromium Controller


a.
Marker 1
Marker 2
Marker 3
Marker 4
Marker 5
Marker 6
Marker 7
Marker 8
Marker 9
Ref. sequence
I. Molecule recovery
Molecule−1
Molecule−2
Molecule−3
Molecule−4
II. Molecule genotyping
non-recombinant (parent A)
non-recombinant (parent B)
recombinant
undetermined
b.
Total length (Gb)
Molecule length (kb)
P50L25
P50L40
P50L75
P200R1
P200R2
P1250R1
P1250R2
P8000
c.
Total reads (million)
Molecule read coverage
d.
Total base coverage (million)
Molecule base coverage


a.
Relative CO frequency (scaled per chromosome)
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
0.0 (Mb)
5.0
10.0
15.0
20.0
25.0
30.4
19.7
23.5
18.6
27.0
Random (3000)
TP+FP (2519)
TP (1874)
FP (645)
b.
r: 0.97
TP CO %
TP+FP CO %
r: 0.29
FP CO %
TP+FP CO %
r: 0.05
FP CO %
TP CO %
c.
CO %
Promoter
5'-UTR
Gene-start
Gene-body
Gene-end
3'-UTR
TE
Intergenic
d.
r: −0.45
GC%
CO %
r: −0.46
GC%
CO %
e.
Methylation level


a.
Chr 1
Chr 2
Chr 3
Chr 4
Chr 5
Relative CO frequency (scaled per chromosome)
0.0 (Mb)
5.0
10.0
15.0
20.0
25.0
30.4
19.7
23.5
18.6
27.0
P8000: 3612
P1250R1: 2880
P1250R2: 3430
b.
r: 0.90
r: 0.77
r: 0.83
P1250R2 CO %
P8000 CO %
P1250R1 CO %