Supporting Information

The great majority of recombination events in *Arabidopsis* are gene conversion events

Sihai Yang [a,1], Yang Yuan [a,1], Long Wang [a,1], Jing Li [a], Wen Wang [b], Haoxuan Liu [a], Jian-Qun Chen [a,2], Laurence D. Hurst[c,2] and Dacheng Tian [a,2]

[a] State Key Laboratory of Pharmaceutical Biotechnology, Department of Biology, Nanjing University, Nanjing, China; [b] State Key Laboratory of Genetic Resources and Evolution, Kunming Institute of Zoology, Kunming, China; [c] Department of Biology and Biochemistry, University of Bath, Bath, U.K, BA2 7AY

Table S1. Summary statistics for sequencing data.

Sample	1st sequencing		2nd sequencing		3rd sequencing		Overlapped
	Mean depth	Genome coverage	Mean depth	Genome coverage	Mean depth	Genome coverage	Genome coverage
Sample_5	20.0	97.6%	23.2	98.7%	—	—	97.5%
Sample_6	20.0	99.1%	19.3	98.5%	—	—	98.3%
Sample_7	20.0	97.5%	23.2	98.7%	—	—	97.3%
Sample_14	19.2	97.2%	22.9	97.9%	—	—	96.9%
Sample_c41	19.4	97.7%	23.3	98.9%	—	—	97.6%
Sample_c42	19.2	98.4%	23.3	99.1%	—	—	98.3%
Sample_c45	19.3	98.8%	23.3	99.1%	—	—	98.7%
Sample_c48	19.6	99.3%	23.3	99.4%	—	—	99.1%
Sample_c51	19.6	98.1%	23.3	98.1%	—	—	97.8%
Sample_c52	19.6	97.8%	23.3	98.2%	—	—	97.6%
Sample_c54	19.6	99.2%	23.3	99.4%	—	—	99.0%
Sample_c57	19.2	98.3%	23.3	99.1%	—	—	98.2%
Sample_c61	19.4	99.4%	23.3	99.6%	—	—	99.2%
Sample_c62	19.5	98.3%	23.3	98.9%	—	—	98.1%
Sample_c63	19.2	95.7%	23.3	96.4%	—	—	95.3%
Sample_c64	19.3	99.3%	23.3	99.4%	—	—	99.1%
Sample_c65	19.2	98.1%	23.3	99.1%	—	—	98.0%
Sample_c66	19.4	98.9%	23.3	99.6%	—	—	98.8%

Sample_c73	19.5	99.2%	19.5	99.0%	—	—	98.8%
Sample_c81	19.6	96.0%	23.3	95.9%	—	—	95.5%
Sample_c82	19.4	96.2%	23.3	96.5%	—	—	95.8%
Sample_c83	19.7	98.6%	23.3	99.5%	—	—	98.5%
Sample_c84	19.8	97.1%	23.3	96.8%	—	—	96.3%
Sample_c85	19.7	98.8%	23.3	99.6%	—	—	98.8%
Sample_c87	19.7	97.8%	21.8	98.4%	—	—	97.6%
Sample_c88	19.8	98.9%	21.6	99.2%	—	—	98.8%
Sample_c89	19.3	96.6%	23.3	97.4%	—	—	96.3%
Sample_c90	19.5	98.1%	22.5	98.7%	—	—	98.0%
Sample_c91	19.4	96.5%	23.3	96.6%	—	—	95.8%
Sample_c92	19.5	97.1%	23.3	97.9%	—	—	96.9%
Sample_c93	19.6	99.1%	23.3	99.5%	—	—	99.1%
Sample_c94	33.0	95.5%	34.0	95.5%	32.4	95.4%	94.6%
Sample_c95	31.4	98.0%	31.2	98.0%	32.2	98.0%	97.8%
Sample_4	20.0	96.5%	—	—	—	—	—
Sample_8	20.0	98.0%	—	—	—	—	—
Sample_18	20.0	98.2%	—	—	—	—	—
Sample_19	20.0	97.3%	—	—	—	—	—
Sample_20	20.5	97.7%	—	—	—	—	—
Sample_21	20.0	96.3%	—	—	—	—	—
Sample_c47	19.4	98.2%	—	—	—	—	—
Col3	28.3	99.9%	28.1	99.9%	28.6	99.9%	99.9%
Ler4	31.2	93.6%	31.0	93.6%	31.5	93.6%	93.1%
Sample_c1c2	23.3	99.0%	—	—	—	—	—
Sample_c2l1	23.3	99.1%	—	—	—	—	—
Sample_l3c1	23.3	99.5%	—	—	—	—	—
Sample_l1l2	23.3	93.4%	—	—	—	—	—
Sample_l2l3	23.3	93.2%	—	—	—	—	—

The *Arabidopsis thaliana* accession "Columbia-0" (Col) and "Landsberg *erecta*" (Ler) were obtained from Joy Bergelson (University of Chicago). In total, there are 75 sequenced F_2 genomes with 1649×coverage. In addition, one of Ler (Ler_4) and Col (Col_3) were independently sequenced thrice with ~29.8×coverage in each sequencing. The other 3 Ler (Ler_1, Ler_2 and Ler_3) and 2 Col plants (Col_1 and Col_2) were sequenced with ~21.2× coverage. In these raw reads, more than 90% bases have the phred quality greater than 29 in the raw reads, while 30 in the mapping reads, indicating high sequencing quality. In raw reads, more than 99% are of the length of 100; in mapping reads, more than 90% reads have the length greater than 96. This Table also showed that the average depth was 22.2X, ranging from 19.2 to 33.9X; the mean coverage mapping to the reference genome was 98.2%, ranging from 93.2 to 99.9%. These statistical analyses have shown a high sequencing quality and depth.

Table S2. Comparison of sequencing strategy and quality between Lu's (Lu et al. 2012) and this study

Items	Sequencing rounds for each sample	Strategy of sequencing	Length of the reads	Coverage for each sequencing
Lu et al. 2011	Once	Single-end reads	46.5 bp (35-70bp)	12.8× (8.2 to 16.6×) in each sample
		Paired-end reads in 200bp inserts	2×40 bp	
This study	Twice or thrice, both independent library constructing and re-sequencing	Paired-end reads in 500bp inserts	2×100 bp	22.2 × (19.3 to 33.9×) in each sequencing; 42.4× in twice sequencing samples and 97.0× in trice sequencing samples

In this study, the paired-end reads with 2×100 bp in average 500bp long inserts should be good for mapping correctly in the small Arabidopsis genome with small proportion of repeat sequences. This expectation was showed in the results of zero controls (Table S4), where no error was observed in the >10kb blocks in our sequences and by the methods of block identification.

Table S3-1. The possible events of GCs identified in 4 samples of Lu's study by the criteria used in this study for different sets of gene conversions

Samples	Col→Ler	Ler→Col	Total
The Sample 1A			
GCs Set 3	3211	9808	13019
GCs Set 2	421	2036	2457
GCs Set 1	3900	5504	9404
2-10kb blocks	502	320	822
The Sample 1B			
GCs Set 3	559	24738	25297
GCs Set 2	100	6223	6323
GCs Set 1	801	8542	9343
2-10kb blocks	157	390	547
The Sample 1C			
GCs Set 3	88	3988	4076
GCs Set 2	13	1033	1046
GCs Set 1	59	1253	1312
2-10kb blocks	9	71	80
The Sample 1D			
GCs Set 3	4885	31530	36415
GCs Set 2	784	7298	8082
GCs Set 1	6369	15509	21878
2-10kb blocks	702	668	1370
Average	2185.8	17516	19702

All the 4 samples were from backcrossing with Col, so GCs should be generated from homozygous to heterozygous backgrounds. Different criteria are used for each of three sets of gene conversions in this study: in the first set, each GC event must contain ≥2 continuous markers in two or three rounds of sequencings and with GC length 20-2000 bp; in the second set, the quality control is the same as in the first set but with 2-19 bp between two border markers; in the third set, 1 or 2 gold standard markers in two or three rounds of sequencings were required. For the third set, the less-reliable SNPs could not be used to increase the reliability of GCs identified, because only one marker was required in either or both independent sequencings.

Table S3-2. Comparison of numbers and types of gene conversions between the two independent sequencings in 31 F_2 plants

Samples	Col→Het	Ler→Het	Het→Col	Het→Ler	Total
First sequencing					
GCs Set 3	172.2	3794.5	9691.5	845.2	14503.4
GCs Set 2	19.5	746.9	1930.9	119.5	2816.9
GCs Set 1	91.2	697.2	3397.3	112.7	4298.5
2-10kb blocks	14.4	24.6	112.9	6.4	158.3
Second sequencing					
GCs Set 3	171.9	3258.8	7268.9	486.0	11185.6
GCs Set 2	19.0	688.3	1491.0	69.9	2268.2
GCs Set 1	77.0	438.6	2017.3	51.3	2584.1
2-10kb blocks	14.0	13.8	66.3	3.4	97.5
Average for two samples	289.6	4831.4	12988.1	847.2	18956.3
Overlapped by two sequencings					
GCs Set 3	169.9	1118.4	774.0	245.2	2307.5
GCs Set 2	18.5	243.8	27.1	31.5	320.9
GCs Set 1	72.3	127.5	46.1	19.5	265.3
2-10kb blocks	13.9	8.2	4.6	3.5	30.2
Total	274.6	1497.9	851.8	299.6	2923.9
Relative rate to one sequencing					
GCs Set 3	0.98	0.32	**0.09**	0.37	0.18
GCs Set 2	0.96	0.34	**0.02**	0.33	0.13
GCs Set 1	**0.86**	**0.22**	0.02	0.24	0.08
2-10kb blocks	**0.97**	**0.43**	0.05	0.71	0.24
Relative rate to Lu' data for Set 1					
For one sequencing	**0.03**	**0.03**	/	/	/
For two sequencings	**0.03**	**0.02**	/	/	/

The number of GCs or blocks is an average for 31 plants, either one or two rounds of sequencings. The same criteria were used as those in Table 2 and Table S5 and S8. The relative rates were calculated as the number identified by two rounds of sequencings divided by the average number in one rounds of sequencings (the first and the second). The relative rates to Lu et al's data were calculated as the total numbers of GCs set 1 in one or two rounds of sequencings divided by the average corresponding numbers in Table S3-1.

The first rates showed that the errors could be reduced to only 2-9% for set 2-3 when the numbers were too high to be true in the first round of sequencing. However, when the low numbers were observed in first round, supposed to be more reliable, the rates were less reduced, e.g., for the set 1 and small blocks. The second rates showed that the rates in the samples with higher quality and coverage were reduced to 2-3%, compared with those in Lu's data. The reduced numbers should be errors most-likely, and the remained should contain the correct ones, almost 100% of which were confirmed by PCR and Sanger sequencing in Table 2. This indicates almost 100% of errors were reduced.

Table S4. The error numbers and types of gene conversions in the mixed samples of Col and Ler DNA.

Samples	Col→Het	Ler→Het	Het→Col	Het→Ler	Total
CK 1 (C2L1)					
GCs Set 3	0	7	20712	16579	37298
GCs Set 2	1	0	288	2604	2893
GCs Set 1	0	1	712	4837	5550
2-10kb blocks	0	0	6	53	59
10-500 kb blocks	—	—	—	—	0
					45800
CK 2 (L3C1)					
GCs Set 3	1	1	12287	6491	18780
GCs Set 2	0	0	329	971	1300
GCs Set 1	2	0	465	1113	1580
2-10kb blocks	0	0	2	12	14
10-500 kb blocks	—	—	—	—	0
					21674
CK 1 + CK 2					
GCs Set 3	0	0	1744	1771	3515
GCs Set 2	0	0	44	187	231
GCs Set 1	0	0	11	151	162
2-10kb blocks	0	0	0	0	0
10-500 kb blocks	—	—	—	—	0
					3908
Relative Error rate					
GCs Set 3	0.00	0.00	0.11	0.15	0.13
GCs Set 2	0.00	/	0.14	0.10	0.11
GCs Set 1	/	0.00	**0.02**	**0.05**	0.05
2-10kb blocks	/	/	**0.00**	**0.00**	0.00
Total					0.12

The criteria for the identification of set 1 - 3 of GCs and 2-500kb blocks were the same as for those in Table 1-2, Table S5 and S8. The relative rates were calculated as the number identified by two rounds of sequencings (CK 1 + CK 2) divided by the average number by one rounds of sequencings (CK 1 or CK 2). The two samples of Col and Ler DNA mixtures, C_2L_1 and L_3C_1, were used as zero controls, because no COs and GCs could be generated in these plants. Indeed by two rounds of independent sequencings, the errors of 2-10kb and <2kb blocks are reduced to 0% and 2-5%, respectively. These results showed that: 1) the two rounds of sequencings are critical to reduce the error rate of GCs in Table 2 and of small blocks; 2) all >10 kb blocks identified in F_2 plants are expected to be correct, because no single block with >10kb was detected in either CK1 or CK2; 3) the higher error rates in heterozygous regions, particularly from Het→Ler, are probably due to the local coverage variation of sequence reads. When missing one or more reads containing two markers in two rounds of sequencing, a false GC event will be present. This result is consistent to the PCR results in Table 2.

Table S5. **The second set of gene conversions in 33 F2 plants.**

Sample	Col→Hete	Hete→Col	Hete→Ler	Ler→Hete	Total
Sample_14	18	49	81	198	346
Sample_5	17	47	61	151	276
Sample_6	7	34	47	261	349
Sample_7	26	34	56	220	336
Sample_c42	15	1	1	90	107
Sample_c45	18	6	0	121	145
Sample_c48	15	16	13	198	242
Sample_c51	17	36	15	293	361
Sample_c52	22	14	17	172	225
Sample_c57	2	31	16	150	199
Sample_c61	22	23	26	44	115
Sample_c62	28	41	44	189	302
Sample_c63	13	19	18	584	634
Sample_c64	17	18	7	122	164
Sample_c65	21	14	16	215	266
Sample_c66	16	13	25	67	121
Sample_c73	18	42	52	59	171
Sample_c81	7	69	79	500	655
Sample_c82	22	50	75	526	673
Sample_c83	22	7	13	161	203
Sample_c84	2	13	28	307	350
Sample_c85	25	33	43	67	168
Sample_c87	25	39	46	403	513
Sample_c88	12	10	15	283	320
Sample_c89	22	15	12	225	274
Sample_c90	32	25	20	190	267
Sample_c91	25	63	83	632	803
Sample_c92	14	26	19	388	447
Sample_c93	31	11	11	133	186
Sample_c94	16	21	29	435	501
Sample_c95	28	20	8	174	230
*Sample_c41	35	92	16	22	165
*Sample_c54	9	35	12	231	287
Average	18.5	27.1	31.5	243.8	320.9

Each GC must contain ≥2 continuous markers in two rounds of sequencings. However, the distance is only 2-19 bp long between two border markers (the farthest distance between two or more markers). This set of data has also been modified via adding less-reliable 212,617 SNPs. If the GC events disrupted by the added markers, these candidate GCs were discarded. Owing to poor sequencing quality, ~50 and 60 MB regions were wiped out in Sample c41 and c54, respectively. Therefore, these two samples were excluded from the calculation of average data. PCR amplification and Sanger sequencing confirmed 100% of 5 samples in homozygous and 30% of 24 sampled in heterozygous regions. Because 100% of GCs in Col background were confirmed, the estimated GCs per genome are 73, based on the equation in Table 2.

Table S6. Identified GCs in non-repeat and repeat regions

GCs	Markers of GCs involved in repeat or non-repeat regions			Total
	All markers in non-repeat regions	All markers in repeat regions	Parts of markers in non-repeat regions	
GCs per plant	116.7	103.5	45.1	265.3
Percent (%)	44.0	39.0	17.0	100

The repeat and non-repeat sequences were grouped by both annotated TEs and RepeatMasker regions for *Arabidopsis* (http://www.repeatmasker.org/) and calculated separately for recombination events to avoid the possible assembly problems in repeat regions. In total, 26.7 and 92.3 Mb regions were identified as repeat and non-repeat sequences in Arabidopsis. Then the first set of 265.3 GCs identified in Table 2 could be categorized as all markers of a GC in repeat, all in non-repeat and parts of markers in non-repeat regions, which resulted in 103.5 (39.0%), 116.7 (44.0%) and 45.1 (17.0%), respectively for the three categories of GCs. In total, 61.0% of GCs were located in or partly in non-repeat regions. These results showed that majority of GCs identified cannot be mapping errors because the sequences around markers in non-repeat regions are unique in Arabidopsis genome (this can be further confirmed by Blast search).

Table S7. PCR confirmation in non-repeat and repeat regions

GCs	Confirmed GCs by PCR and sequencing	PCR numbers in involved in repeats or non-repeats			Total
		All markers in non-repeat regions	All markers in repeat region	Parts of markers in non-repeat regions	
Set 1	True	48	35	21	104
	False	7	12	3	22
Set 2	True	0	6	6	12
	False	7	5	5	17
Set 3	True	7	6	0	13
	False	12	12	0	24
2-10kb	True	6	1	3	10
	False	0	0	0	0
Total	True	61	48	30	139
	False	26	29	8	63

All PCR pairs of primers used for the confirmation of GCs were unique pairs of sequences in Arabidopsis genome. Therefore, these PCR results should be reliable for both repeat and non-repeat regions to confirm whether the GCs identified are true or false. In fact, the confirmed rate in non-repeat (61/87) is similar to that of repeat regions (48/77).

Table S8. **The third set of gene conversions in 33 F2 plants.**

Sample	Col→Hete	Hete→Col	Hete→Ler	Ler→Hete	Total
Sample_14	154	1741	734	1008	3637
Sample_5	186	1494	545	674	2899
Sample_6	80	974	334	1160	2548
Sample_7	123	1244	459	852	2678
Sample_c42	139	169	0	470	778
Sample_c45	196	413	0	544	1153
Sample_c48	145	388	104	658	1295
Sample_c51	179	571	146	1280	2176
Sample_c52	211	559	135	716	1621
Sample_c57	19	502	140	595	1256
Sample_c61	221	690	209	147	1267
Sample_c62	365	1068	352	1106	2891
Sample_c63	211	589	168	2482	3450
Sample_c64	204	421	92	393	1110
Sample_c65	143	435	149	805	1532
Sample_c66	115	575	177	201	1068
Sample_c73	160	1124	323	160	1767
Sample_c81	104	1674	623	2199	4600
Sample_c82	156	1627	631	2055	4469
Sample_c83	163	353	97	474	1087
Sample_c84	33	516	149	1328	2026
Sample_c85	336	992	248	261	1837
Sample_c87	226	1063	372	1392	3053
Sample_c88	119	362	73	850	1404
Sample_c89	193	435	105	1235	1968
Sample_c90	233	583	118	846	1780
Sample_c91	163	1418	541	2750	4872
Sample_c92	159	558	175	1475	2367
Sample_c93	303	312	87	491	1193
Sample_c94	169	53	19	6633	6874
Sample_c95	413	144	13	2480	3050
*Sample_c41	302	1514	121	89	2026
*Sample_c54	33	619	77	801	1530
Average	181.3	743.5	236.1	1216.8	2377.6

Each GC must contains ≥1 marker in two rounds of sequencings. When a GC has only one marker, its track length is unavailable. Attributed to the sequencing quality, ~50 and 60 MB regions were wiped out in Sample c41 and c54, respectively. Therefore, these two samples were excluded from the calculation of average data. As above (Table S5), Samples c41 and c54 were excluded from the calculation of average data. PCR amplification and Sanger sequencing confirmed 100% of 9 samples in homozygous and 14.3% of 28 sampled in heterozygous regions.

Table S9. Estimating the GC rate in the genome.

	Average GC tract length (bp)	The rate of GC per site in the genome (the total length of all GC tracks divided by the genome length)
Observed	335	0.0007
Simulated	704	0.0015

In this study, we assume that 265 GC events are detected in each sample and employ these to estimate the mean tract length and the proportion of the genome subject to tracts including missed events. For each GC event, we use parameter i to stand for the start mark and j to represent the last mark. m_i and m_j describe the genome position of the mark i and mark j.

It is known that there are two breakpoints in each GC event, one of which is between mark i-1 and mark i and the other is between mark j and mark j+1. We assume x_1 was one base between mark i-1 and mark i, and x_2 was between mark j and mark j+1. As in model 1 described in Genes 2: 313-331, the elongation of converted tracts in the two directions independently follows an exponential distribution:

$$\mathrm{P}(x_1, x_2 \,|T) = \frac{4}{T^2} e^{-\frac{2(x_2 - x_1)}{T}}$$

Here T stands for the expected GC length. And then the probability that the GC event described by i, j is

$$\mathrm{Prob(i, j|T)} = \int_{m_{i-1}}^{m_i} \int_{m_j}^{m_{j+1}} P(x_1, x_2|T)\, dx_2 dx_1$$

From here we can obtain the relationship between probability of a GC event and the T. For these GC events, m_{i-1}, m_i, m_j and m_{j+1} are constants, and for each T value between (m_j-m_i) and (m_{j+1}-m_{i-1}), we can get a probability of the GC events. When the probability reaches the maximum value, T indicates the most probable GC length. This suggests the average expected GC track is 704bp and the gene conversion rate of all the genome is 0.0015 per site. If without using this model, the average length of each GC track is 335bp and the GC rate in the genome is about 0.0007 per site.

Reference

Mansai SP, Kado T, Innan H. (2011) The rate and tract length of gene conversion between duplicated genes. Genes 2: 313-33

Table S10. Confirmation of 10 candidate GCs near centromeric regions by a single-stranded cloning strategy.

Loci	Samples	Chr.	Orientation of GCs	Start position of GCs	End position of GCs	Distance from Centromeres	PCR Regions	Non-phased genotypes	Genotypes of the sequencing results
1	c90	1	Col->Hete	13037739	13037841	262159	13037535-13037878	CCCHHC	CCCCCC
									CCCLLC
2	c90	1	Col->Hete	16675486	16675554	675486	16674803-16675572	CCCHHHHHC	CCCCCCCCC
									CCCLLLLLC
3	c52	2	Col->Hete	1840318	1841004	1258111	1839746-1841889	CCHHC	CCCCC
									CCLLC
4	c52	2	Col->Hete	3037888	3038286	61596	3037766-3038404	CCHHHCC	CCCCCCC
									CCLLLCC
5	c93	3	Col->Hete	11676329	11676361	223639	11676200-11676615	CCHHCC	CCCCCC
									CCLLCC
6	c93	3	Col->Hete	15265697	15265744	165616	15265616-15265849	CCCHHCC	CCCCCCC
									CCCLLCC
	c62	4	Col->Hete	1653179	1653277	1246401	1653108-1653599	CHHCC	CCCCC
									CLLCC
8	c62	4	Col->Hete	5560971	5561132	460494	5560494-5562044	CCHHHHC	CCCCCCC
									CCLLLLC
9	c66	5	Col->Hete	9952328	9953414	46586	9952265-9953448	CCHHHHHC	CCCCCCCC
									CCLLLLLC
10	c66	5	Col->Hete	14116137	14116311	616137	14115529-14116592	CCHHC	CCCCC
									CCLLC

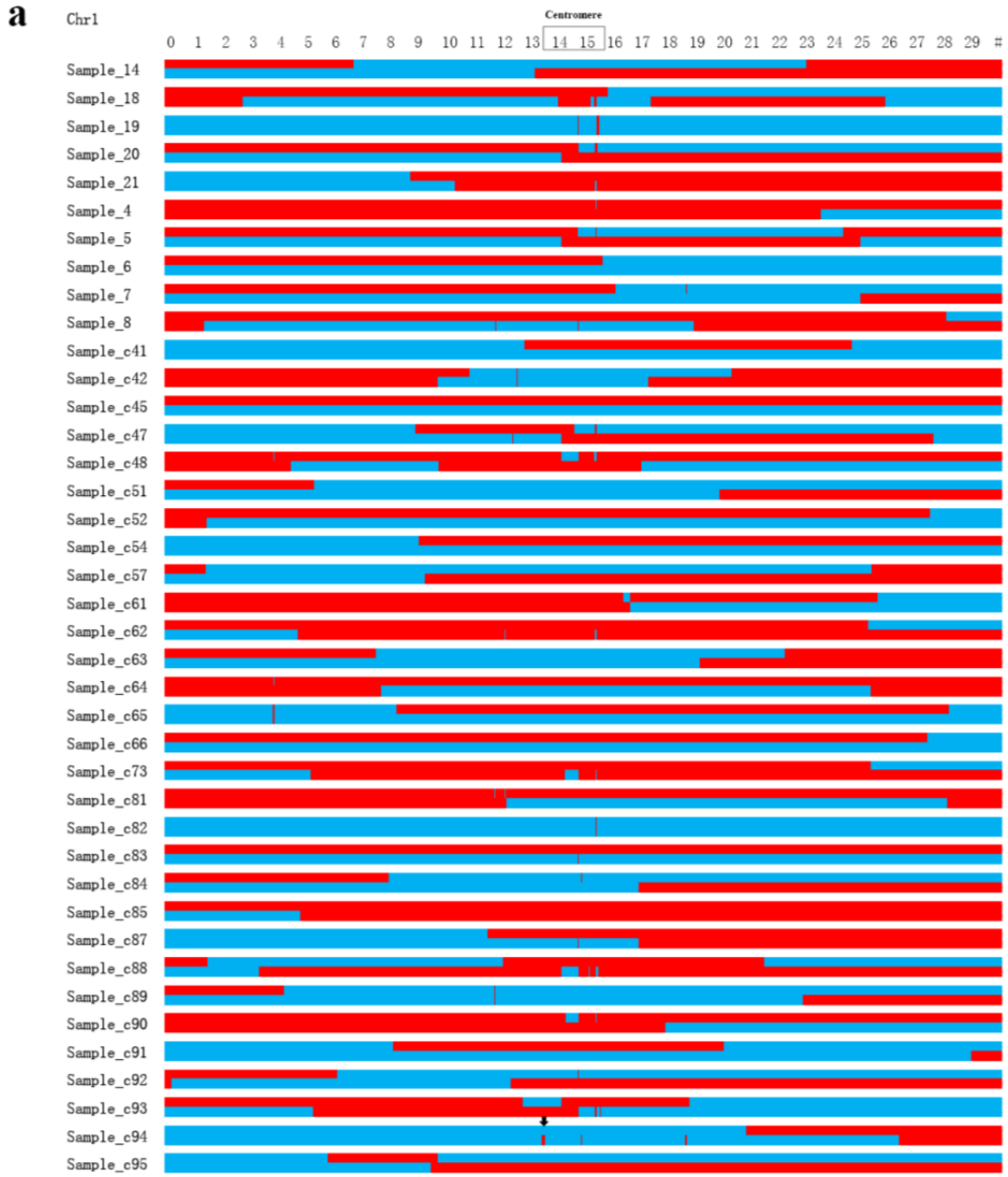
A single-stranded cloning strategy was employed. This can unambiguously tell the exact sequence for each sister chromosome at the same region. Two criteria were used to choose the candidate GCs: 1) the background of these candidate GCs should be pure Col-0, where the exact genome sequences are available to determine the chromosome positions of GCs, particularly for those near centromeres. This requires that the direction of GCs should be Col -> Ler and the non-phased genotype of the chromosome pair of the GCs should be “CCCCCCCHHHCCCCCCC” (C, pure Col genotype markers; H, Heterozygous genotype; Fig. S11); 2) the primers must be designed on the pure Col regions and cover all of the heterozygous tract. Then, the PCR products were cloned into pGEM-T Easy vectors (Promega) and at least five clones were selected to sequence (Sanger sequencing). Finally, 10 candidate GCs, putatively located in pericentromeric regions, have been analysied by this strategy. Two haplotypes, “CCCCCCCLLLCCCCCCC” (L, pure Ler genotype markers;) and “CCCCCCCCCCCCCCCCC”, were detected in all of 10 candidate GCs, indicating that these GCs are in the pericentromeric regions.

Table S11. **MEME motifs identified surrounding or within shared COs or GCs.**

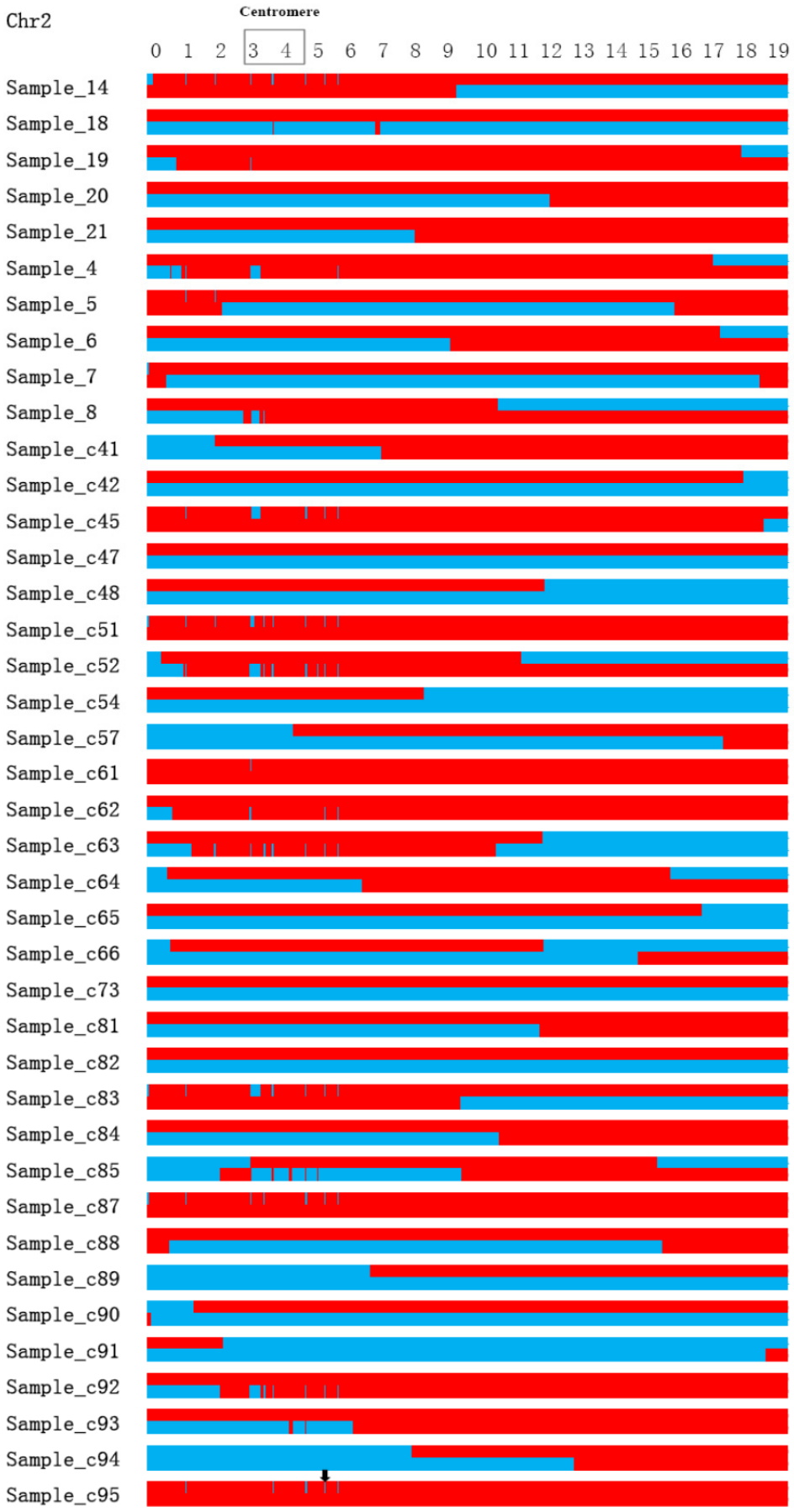
Unique motifs (only present in certain loci of GCs and COs)			
Width	**Regular expression**	**Sites**	**E-value**
SHARED_BP	942		
15	G[N]C[N][N]C[N]G[N]AG[N][N]GC	61	1.4e-053
15	[N]AA[N]AA[N]AA[N][N][N][N]AA	99	1.6e-048
14	[N][N][N]GA[N]C[N]GG[N][N]GC	97	6.9e-041
14	[N][N]C[N]GG[N]G[N]CTTC[N]	53	1.1e-038
SHARED_ INTER	797		
14	G[N][N][N][N]TGGTGG[N]GG	30	3.4e-012
13	A[N][N][N]A[N]A[N]A[N]A[N]A	98	1.7e-010
General motifs (present in group 1 & 2 sequences and NONE_GC group:			
Width	**Regular expression**	**Sites**	**E-value**
SHARED_BP	942		
15	AGA[N][N]AA[N]A[N][N][N]A[N]A	>=100	2.2e-146
14	[N][N]G[N][N]GG[N]GGAG[N][N]	>=100	4.0e-083
14	AA[N]AAAAAA[N]AAAA	>=100	1.6e-112
13	C[N]TC[N]TCTTCT[N]C	99	3.3e-070
15	[N]T[N]T[N][N][N]T[N]T[N][N][N]T[N]	>=100	1.3e-059
14	[N][N]AGA[N][N][N]AA[N]A[N][N]	98	8.8e-030
SHARED_INTER	797		
14	AAAA[AG]AAAAAAAAA	99	5.2e-128
15	[N]T[N]T[N][N][N]TCT[N][N][N]T[N]	>=100	9.6e-088
11	AAAAAAAAAAA	98	1.2e-072
11	GA[N]GA[N]GA[N]GA	>=100	2.0e-029
15	A[N]A[N][N][N][N][N]A[N]AGA[N]A	92	1.1e-020
15	G[N]A[N]GAG[N][N][N]GA[N][N]A	48	2.7e-003
10	TTTTT[N]TTT[N]	89	2.7e-003

Several groups of sequences were chosen to search candidate motifs by using MEME software (http://meme.nbcr.net). 1) SHARED_BP: The 100-1000bp flanking regions of shared GCs and COs. A total of 942 sequences (471 loci) were chosen for this group. 2) SHARED_INTER: Converted regions of shared GCs and tracts of COs, with shared individual frequency ≥2. Each of sequences contains repeat masked regions no more than 50%. This group contains 797 sequences. (3) NONE_GC: Regions where no GC event found, with length <= 2000bp, this group contains 1000 sequences chosen by random. This group was used as a control to compare whether there are unique motifs in other groups. All repetitive regions were masked with Ns use RepeatMasker (with RM database version 20110920), and all TE regions annotated by TAIR9 were also masked with Ns with a local PERL script. Several criteria were applied to keep a reasonable size of each dataset in considering of the speed of MEME. The MEME software was run locally with options: -nmotifs = 10, -evt = 0.01, -minw = 10, -maxw = 15, -maxsites = 100, -mod = anr. The MEME results showed that some of shared loci have a specific motif(s), and some have a different one(s) either surrounding or within shared GCs (or COs). However, the others have no conserved ones.

Fig. S1. The distribution of COs on five pairs of chromosomes in 40 F_2 plants. 415,357 reliable markers were used for COs identification. Based on this amount of markers and a random occurrence model, every long CO (>500 kb) should have a specific tract length, a piece between two COs. The probability of cross-over occurs at the same locus in two independent meioses is expected to be almost zero. For example, with dense and physically unlinked markers (e.g., >300000) and sparse recombination events (e.g., <1000) in a diploid plant, assuming a random occurrence model, the probability of two events occurring between the same two markers is roughly equal to $1000/300000^2 = 1.1 \times 10^{-8}$.

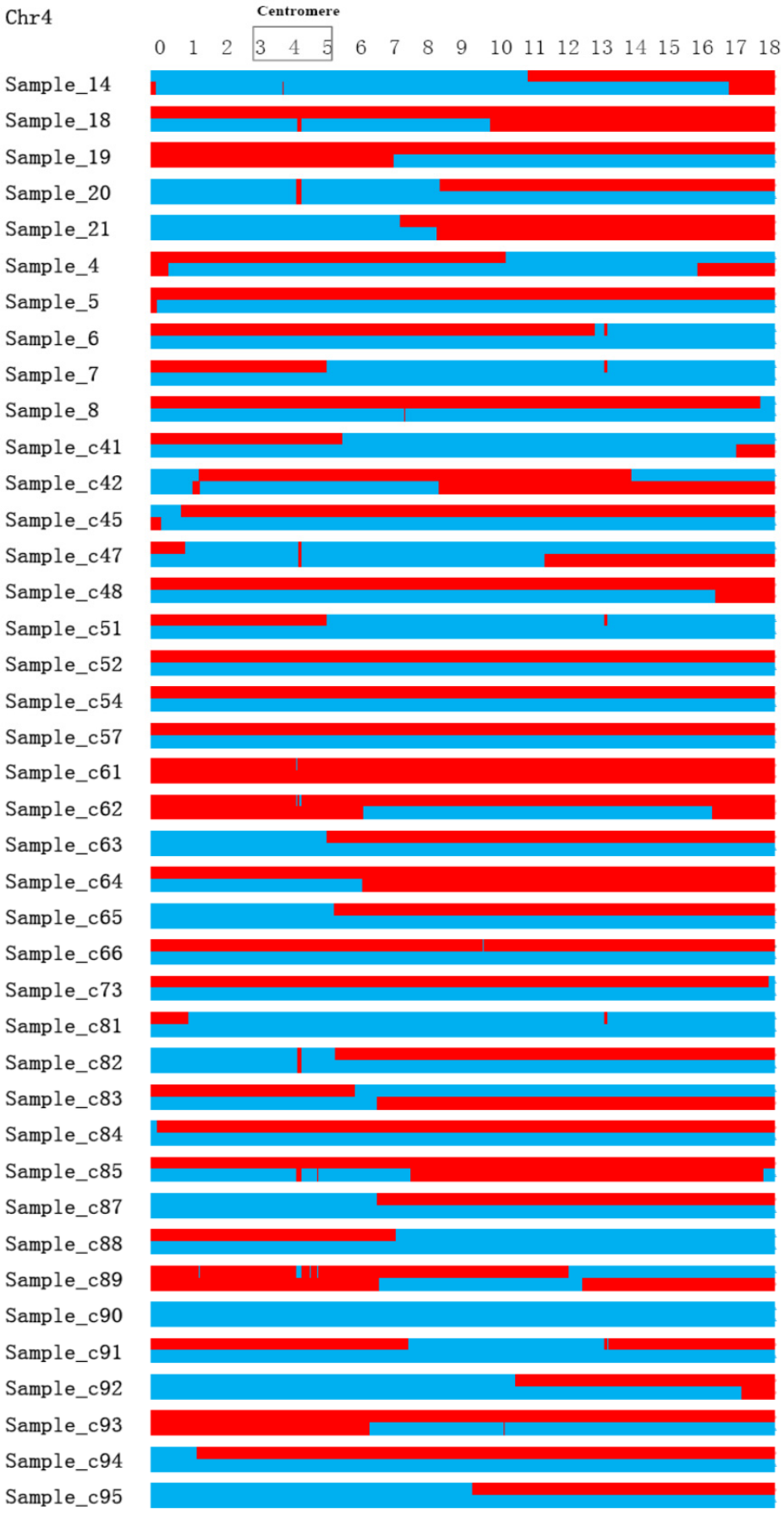


b


Chr2
Centromere
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19
Sample_14
Sample_18
Sample_19
Sample_20
Sample_21
Sample_4
Sample_5
Sample_6
Sample_7
Sample_8
Sample_c41
Sample_c42
Sample_c45
Sample_c47
Sample_c48
Sample_c51
Sample_c52
Sample_c54
Sample_c57
Sample_c61
Sample_c62
Sample_c63
Sample_c64
Sample_c65
Sample_c66
Sample_c73
Sample_c81
Sample_c82
Sample_c83
Sample_c84
Sample_c85
Sample_c87
Sample_c88
Sample_c89
Sample_c90
Sample_c91
Sample_c92
Sample_c93
Sample_c94
Sample_c95


c
Chr3
Centromere
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23
Sample_14
Sample_18
Sample_19
Sample_20
Sample_21
Sample_4
Sample_5
Sample_6
Sample_7
Sample_8
Sample_c41
Sample_c42
Sample_c45
Sample_c47
Sample_c48
Sample_c51
Sample_c52
Sample_c54
Sample_c57
Sample_c61
Sample_c62
Sample_c63
Sample_c64
Sample_c65
Sample_c66
Sample_c73
Sample_c81
Sample_c82
Sample_c83
Sample_c84
Sample_c85
Sample_c87
Sample_c88
Sample_c89
Sample_c90
Sample_c91
Sample_c92
Sample_c93
Sample_c94
Sample_c95

d


Chr4
Centromere
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Sample_14
Sample_18
Sample_19
Sample_20
Sample_21
Sample_4
Sample_5
Sample_6
Sample_7
Sample_8
Sample_c41
Sample_c42
Sample_c45
Sample_c47
Sample_c48
Sample_c51
Sample_c52
Sample_c54
Sample_c57
Sample_c61
Sample_c62
Sample_c63
Sample_c64
Sample_c65
Sample_c66
Sample_c73
Sample_c81
Sample_c82
Sample_c83
Sample_c84
Sample_c85
Sample_c87
Sample_c88
Sample_c89
Sample_c90
Sample_c91
Sample_c92
Sample_c93
Sample_c94
Sample_c95


e
Chr5
Centromere
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
Sample_14
Sample_18
Sample_19
Sample_20
Sample_21
Sample_4
Sample_5
Sample_6
Sample_7
Sample_8
Sample_c41
Sample_c42
Sample_c45
Sample_c47
Sample_c48
Sample_c51
Sample_c52
Sample_c54
Sample_c57
Sample_c61
Sample_c62
Sample_c63
Sample_c64
Sample_c65
Sample_c66
Sample_c73
Sample_c81
Sample_c82
Sample_c83
Sample_c84
Sample_c85
Sample_c87
Sample_c88
Sample_c89
Sample_c90
Sample_c91
Sample_c92
Sample_c93
Sample_c94
Sample_c95

Fig. S2. The relationship between marker number in both GCs and small part of small COs and their occurrence. The Y-axis is the number of events of GCs and some small COs per plant. The X-axis is the number of markers in each GC or small CO event. The formula is calculated based on the black dots which are the observed data. The three red dots are expected for the numbers of GCs with 1-3 markers, based on this formula. The expected numbers (1718, 797 and 370) are actually less than those (5240, 1491 and 578 for 1-3 markers identified, respectively).

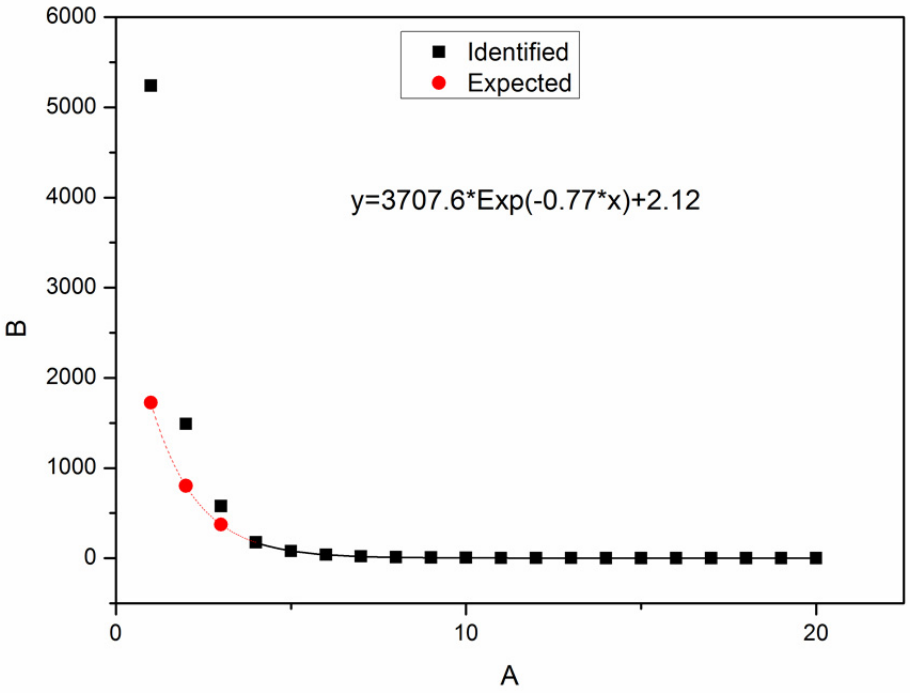
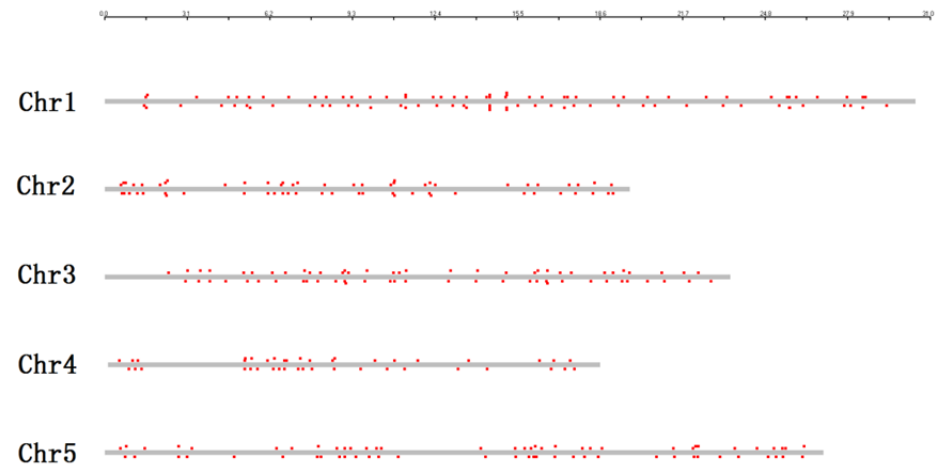
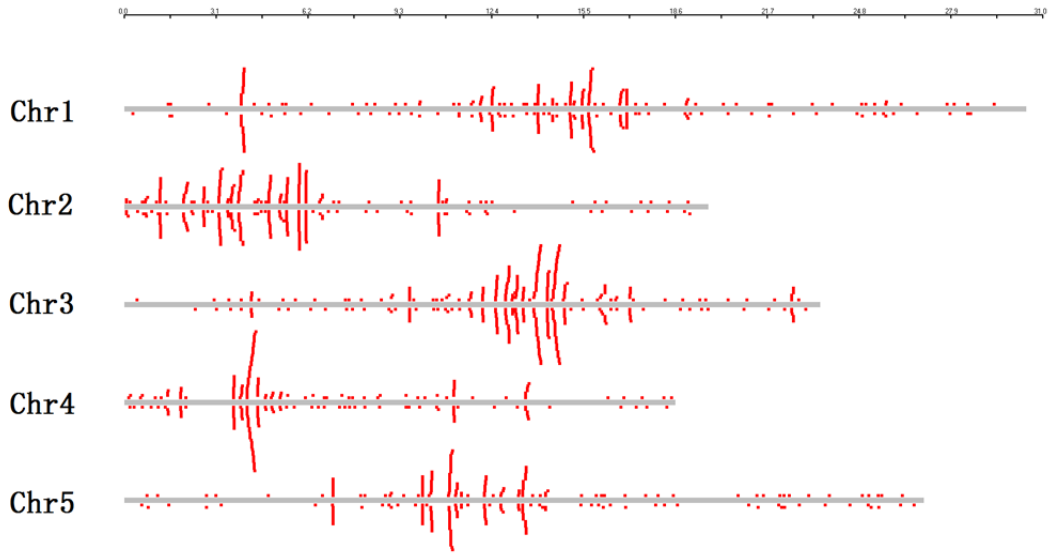


Fig. S3. Distribution of COs (a and b) and GCs (c) on five chromosomes. The long (>500Kb, a) and small COs (10-500 Kb, b) are shown separately. Shared and non-shared GCs (20bp to 10 Kb, c) were demonstrated by different lines. The centromere regions were represented as grey bars in c. **a.** Distribution of the long COs (>500Kb) on chromosomes. **b.** Distribution of the small COs (10-500 Kb) on chromosomes, where ~72.6% of small COs were located in centromeres or 2 Mb regions around, the hot-spots distribution. On average, 17.0 out of 120 Mb (14.2%) of the 2-Mb regions are converted during one meiosise. **c.** Distribution of GCs on chromosome 3, 4 and 5. See Fig 2b for chromosomes 1 and 2. Shared and non-shared GCs were demonstrated by black and red lines, respectively. The centromere regions were represented as grey bars. The total GC numbers were calculated for every Mb of 31 F_2 individuals along a pair of chromosomes. Only these GCs in Table 2 were used.

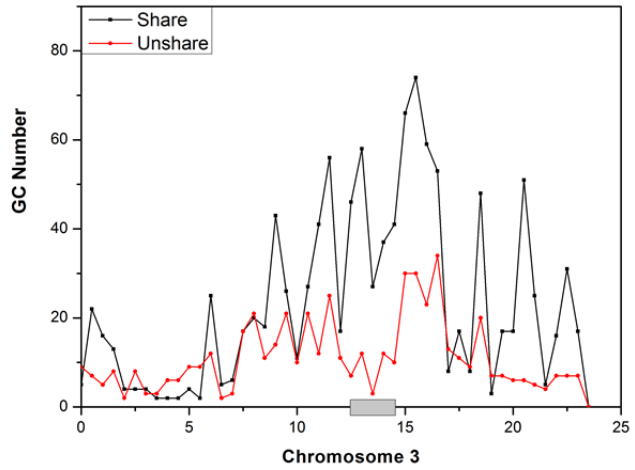
a

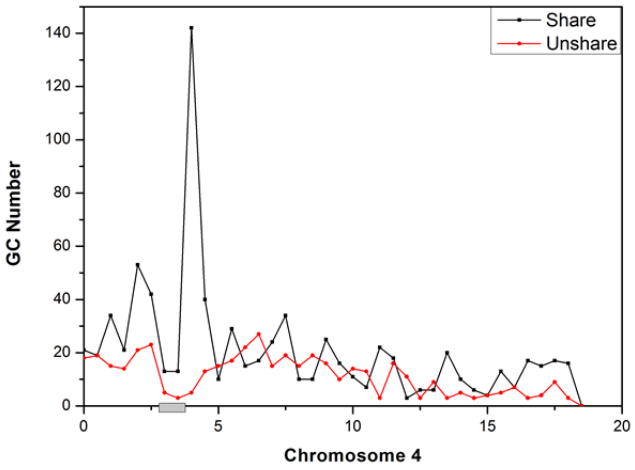


b



c


Share
Unshare
GC Number
Chromosome 3


Share
Unshare
GC Number
Chromosome 4

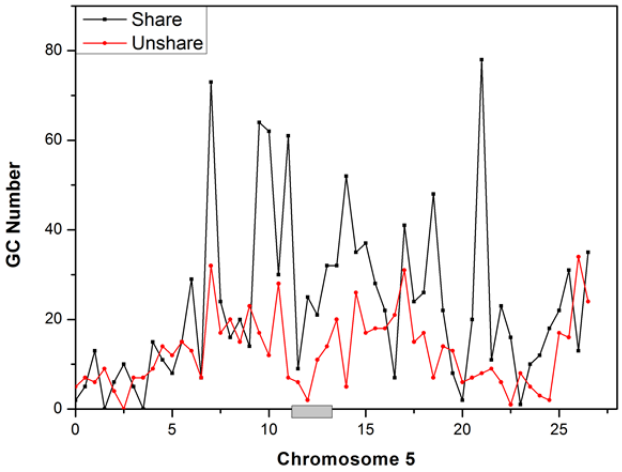
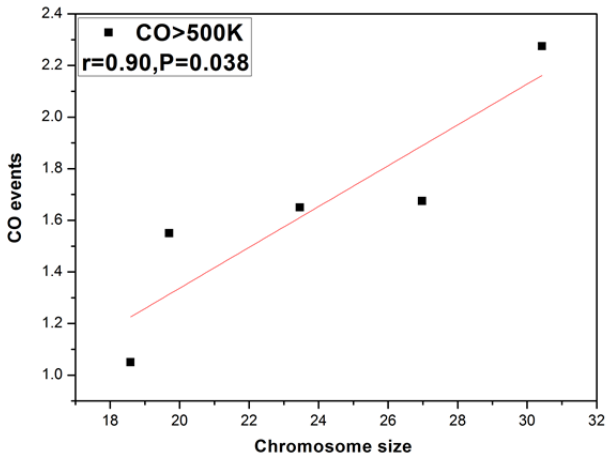

Share
Unshare
GC Number
Chromosome 5

Fig. S4. Correlation between CO sizes and chromosome lengths. a. When the CO size is ≥500kb, the average CO number is correlated with chromosome physical length, which is consistent with previous studies (Salome et al. 2012). b. When the CO size is >10 kb, the positive correlation was not detected (r=-0.078, *P*=0.90)

a



b

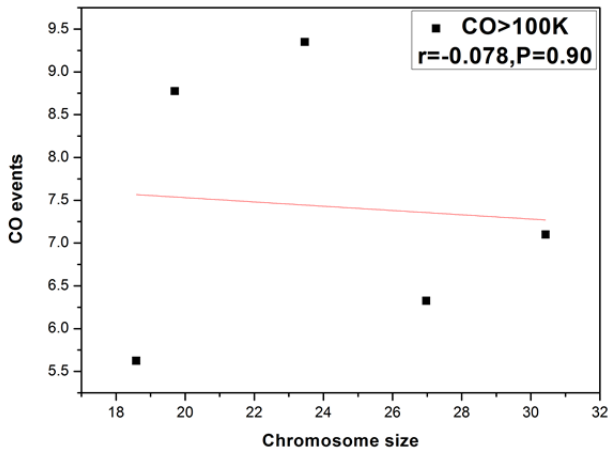


Fig. S5. The heterogeneous paired-ends the same insert at the CO transition bolder. There are two examples to display the mapped reads in breakpoints of CO events by using software inGAP-sv (http://ingap.sourceforge.net/) at the position 13949040 on chromosome 1 in Fig. 1b (pointed by an arrow) and 4925084 on chromosome 2. In the top figure, the genotype in this region (the top bars) should be LLLL-HH, the H-genotype within L background. The breakpoint is marked with an arrow. The red and blue vertical lines denote L-homozygosity and heterozygosity (H) markers, respectively. The blue squares stand for the reads at the left side of paired ends and the red ones at right side. The grey lines are un-sequenced region of an insert between the two sequenced ends as a pair. The stars (*) indicate those reads that have Het marker(s) on the left sides but Ler marker(s) on the right for the same paired ends. Actually, the reads around breakpoint contains heterogeneous markers in themselves. The bottom figure is another example of a breakpoint in the end.

In total, 12 COs in Sample_c94 and c95 (each with 97× coverage) meet our criteria. 8 of them are confirmed to be real and the others can be still true because of the drawbacks in this analysis (see Methods for explanation). These examples, the changed state at the border of CO transition, show that the COs are not from building problems and that high mapping quality is obtained by the long paired ends (2×100bp) in a long insert (500bp)

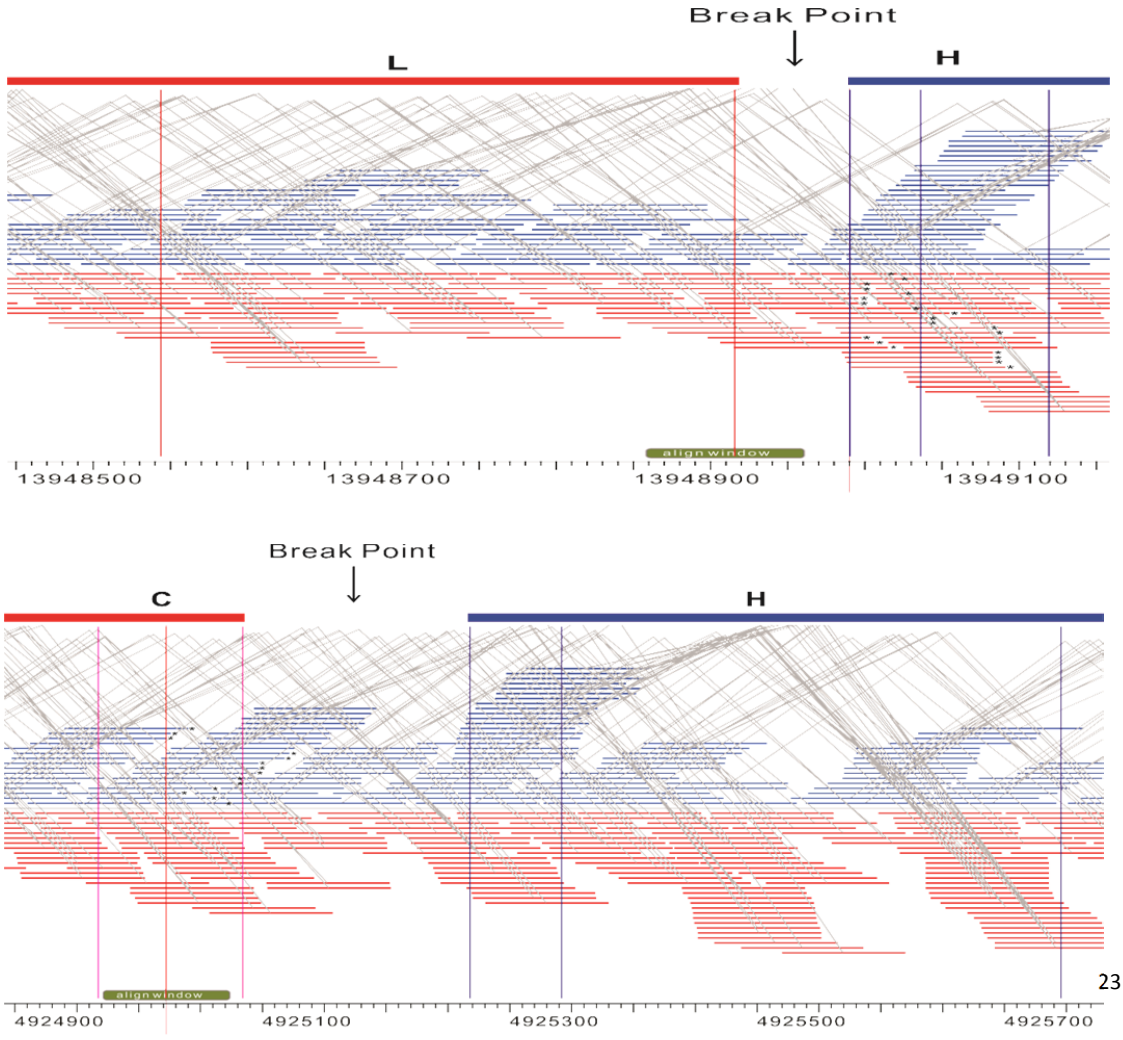
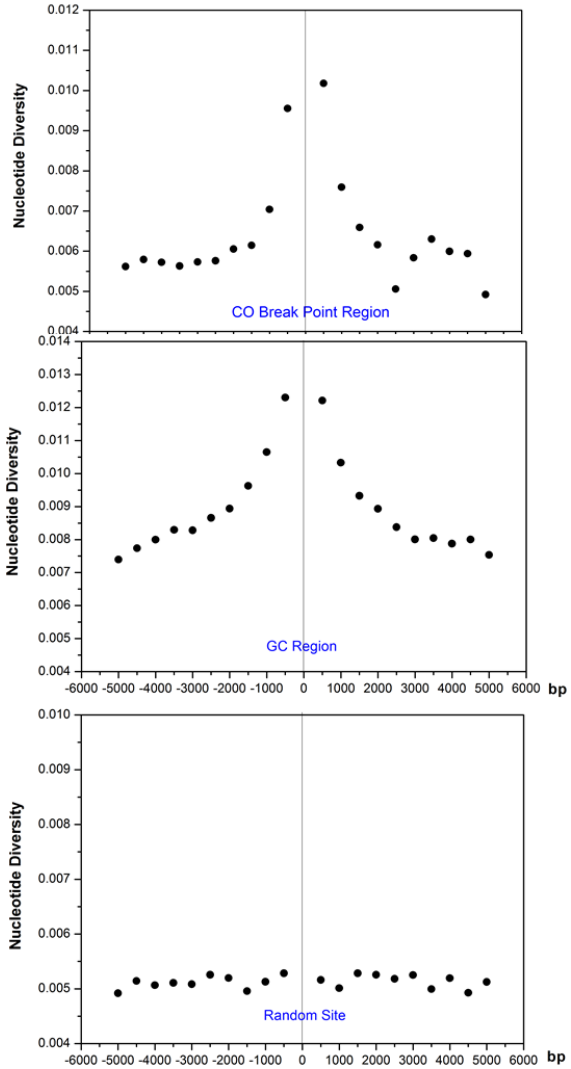
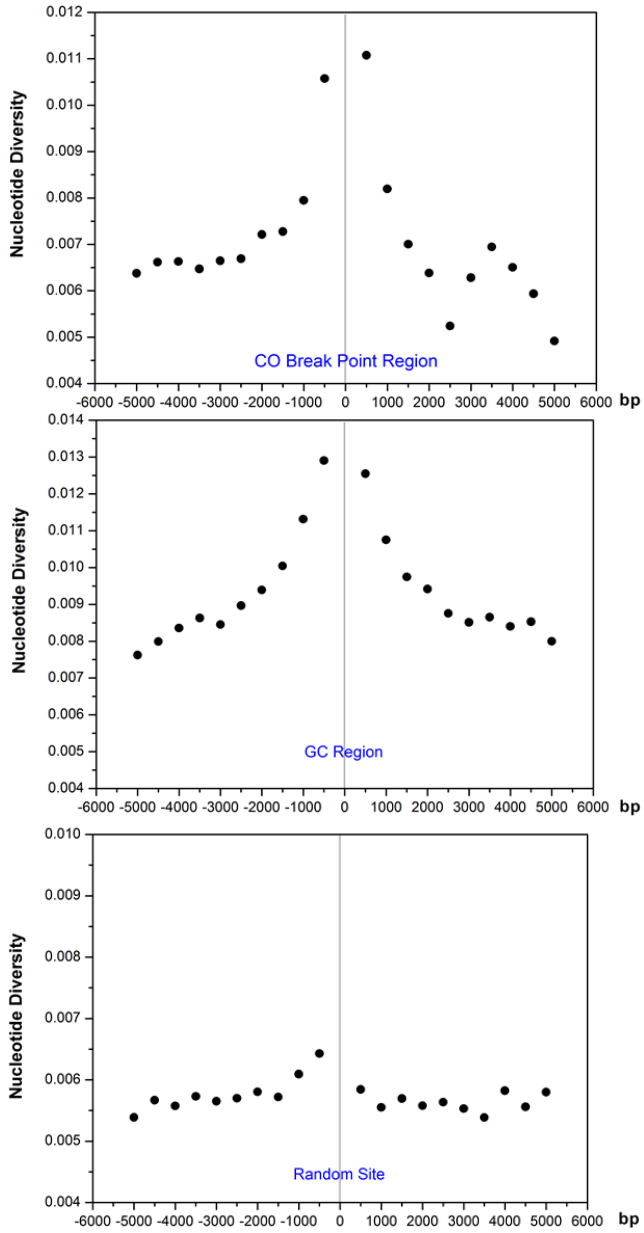


Fig. S6. The level of SNPs around crossover breakpoints (top), gene conversions (middle) and randomly sampled loci (bottom) for all events (a) and for intergenic events alone (b).



6a The level of SNPs around crossover breakpoints (top), gene conversions (middle) and randomly sampled loci (bottom). The Y-axis is the level of nucleotide diversity and the zero point of X-axis is the break positions of COs, two borders of GC tracts or randomly sampled positions. The absolutely number represents 100 bp distance from the zero position.



6b. The level of SNPs around intergenic crossover breakpoints (top), gene conversions (middle) and randomly sampled loci (bottom). Given that recombination events tend to avoid genes and genes tend to have higher constraint, it is worth asking whether the genome wide pattern (Fig S6a) is replicated if we just examine intergenic domains. This is presented here. As most The Y-axis is the level of nucleotide diversity and the zero point of X-axis is the break positions of COs, two borders of GC tracts or randomly sampled positions. The absolutely number represents 100 bp distance from the zero position.

Addendum: Diversity near GC events. A higher diversity in the vicinity of GC events is likely in part to be a definitional necessity. Imagine that DSB followed by SDSA events were randomly located around the genome. When this event occurs but the domain of influence (circa a few hundred bp on average) contains no SNPs we cannot detect it. This must lead to a correlation between rates of observed GC and local diversity, if only because in zones of zero diversity there must have zero GC events, rooting any regression through the origin. So one could say that the correlation is an ascertainment bias. However, if by gene conversion one means not just that DSB followed by SDSA (or whatever mechanism) occurs but also that this is accompanied by a shift in polymorphic alleles (the conversion part of GC), then by definition, GC can only occur when there is diversity. In this sense the GC/diversity correlation is a necessary consequence of definition, not an ascertainment bias: with random DSB and SDSA, GC has to be more common in zones of higher diversity.

A second question is whether the diversity in the vicinity of GC events is higher than would be expected given that higher diversity is expected in GC domains owing to the definition of GC requiring SNPs to convert. We address this via simulation. Here we sample from the observed size distribution of GC events. We randomly place these events on the genome and then reject any events that don't pass our filters: they must contain enough SNPs. We can then ask about the diversity in the vicinity of this pseudo random set. The diversity (0.012) around the breakpoints of our identified GCs is significantly higher than the random tracks (0.0073; from10,000 times' random repeats; P <0.0001) which in turn is higher than around unbiased randomly selected sites. This indicates that diversity is higher than expected around GC sites, even permitting that GC events require diversity to define them.

As regards crossing over we can be confident that there cannot be an ascertainment bias.

On average crossing over events are very long 100kb+ while SNPs are at a very fine resolution (one per every 250bp or so). It is next to impossible for us to miss any such crossing over events and so there can be no ascertainment bias whereby some biased set of events are missed. The probability a breakpoint would be in a zone of high diversity should then be approximately equal to the proportion of zones that are high diversity under the random location of breakpoints model.

Fig. S7. Distribution of crossover breakpoints (CBs) and gene conversion tracts (GCTs) in genes and their flanking sequences. Each intergenic sequence was partitioned into two fragments with equal length, both of which were assigned as the 3' or 5' flanking sequences to the two adjacent genes according to their transcription directions, respectively. Total CBs and GCTs were counted in CDS, UTR and intron regions, and 500-bp non-overlapping windows of flanking sequences, respectively. Then the numbers of CBs and GCTs were divided by the sequence length and by the total number of markers in different regions or windows, respectively. Therefore, the y-axis ($\times 10^{-2}$) reflects the occurrence rate of CBs and GCTs in corresponding regions of 31 F_2 plants, after excluding the uneven distribution of marks and the length difference.

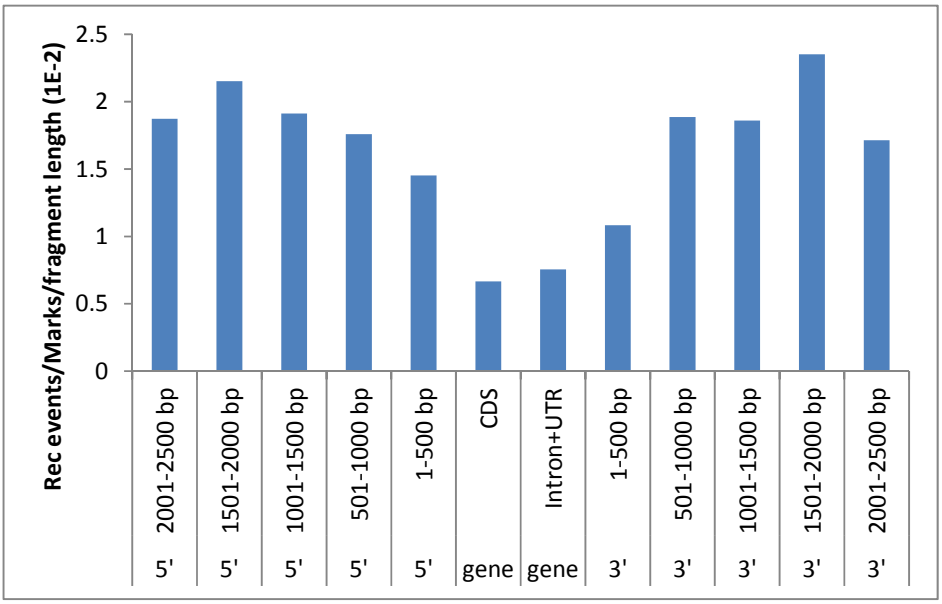


Fig. S8 Trees for shared GCs and COs. The F_2 plant trees grouped relative to their shared GCs (a) and COs (b), constructed by shared loci (using 1 for shared present or 0 for absent) and by Maximum Likelihood (ML) method using Phylip-3.69. The confidence for each branching node was assessed by bootstrap analysis with 1000 replicates.

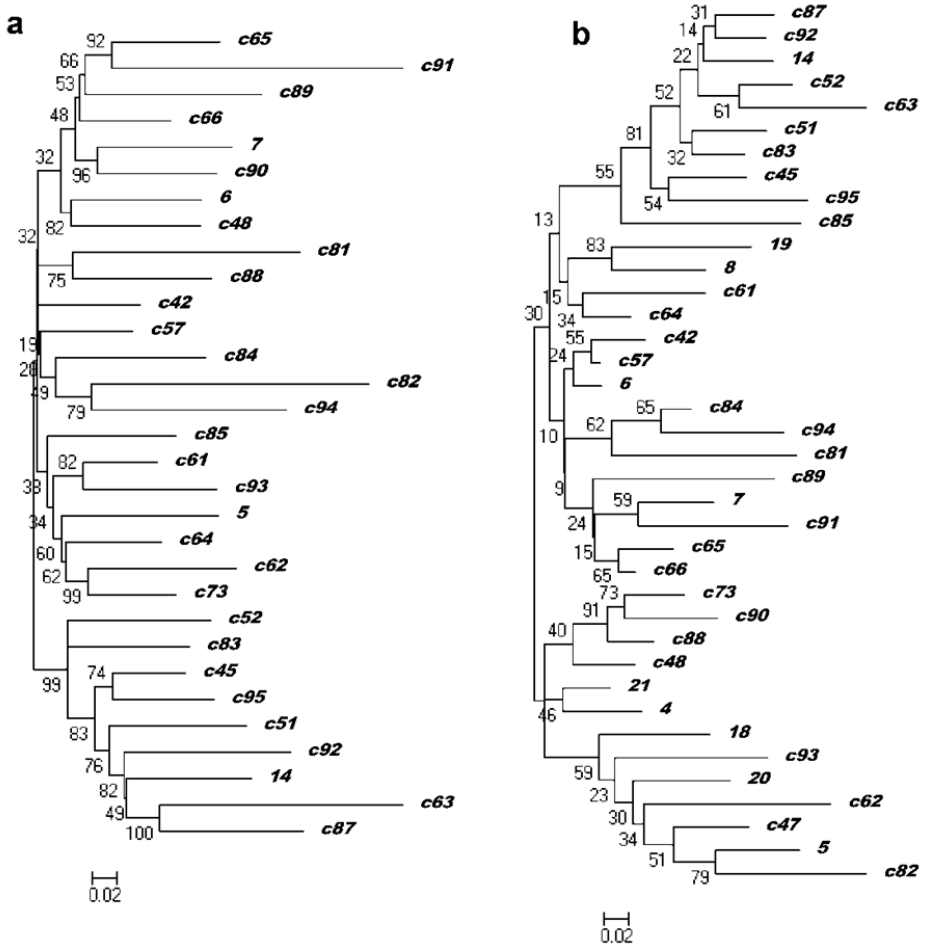


Fig. S9. Distribution of shared small COs on chromosomes

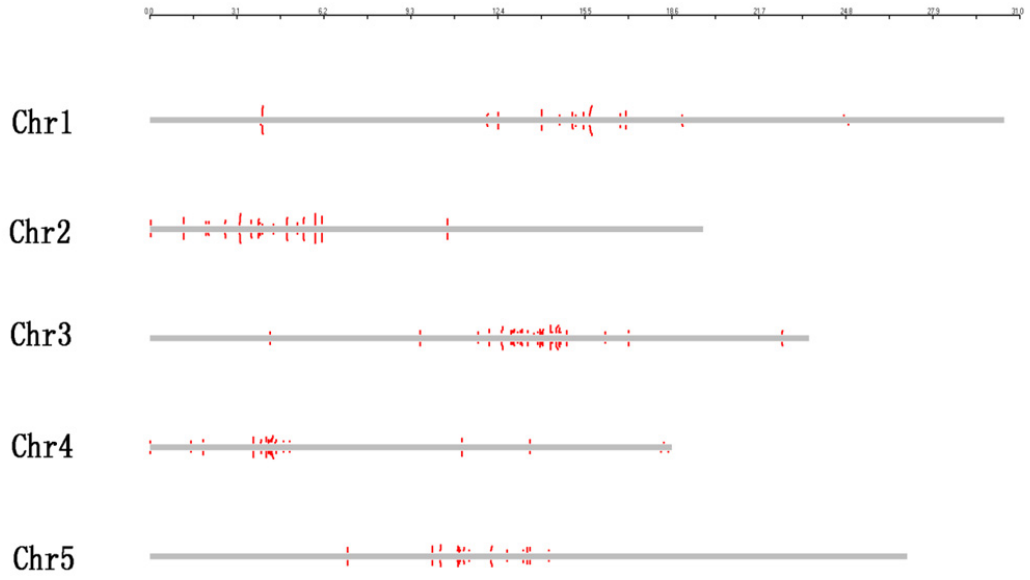


Fig. S10. Two examples of the consistent results between background genotypes and confirmed pattern of PCR and Sanger sequences in the shared small CO (a, 45623bp) and GC (b, 1570) among 3 and 7 different individual plants, respectively. The PCR positions and lengths are showed proportionally. The last (a) and the last four plants (b) are controls with different backgrounds, where no CO or GC events occur. There were two locations of PCRs in a and four different pairs of PCRs in b to show the consistent results, respectively. These results strongly suggest that the shared small CO and GC are not artifacts.

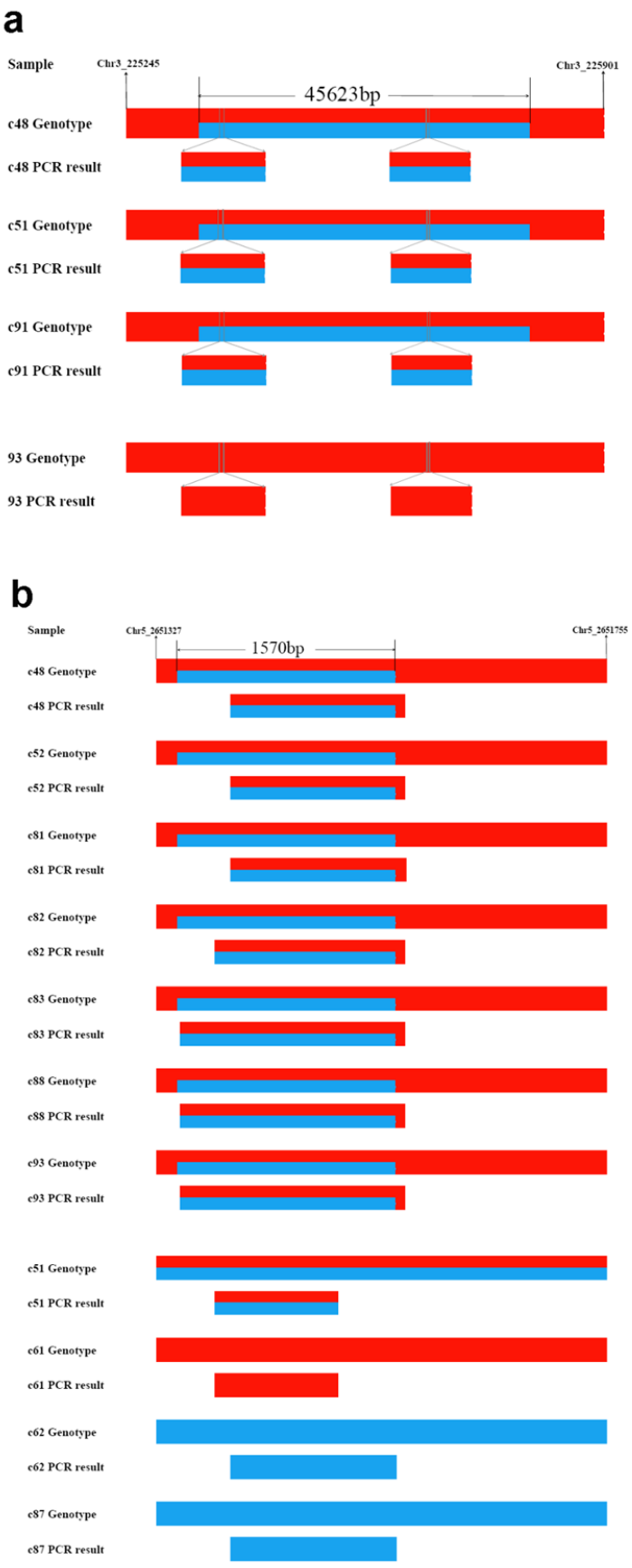


Fig. S11. Distribution of the Col- or Ler-homozygous sequences along five chromosomes in 40 F_2 plants. The average proportion of Col- or Ler-homozygous sequences was calculated for every Mb of 40 F_2 individuals along a pair of chromosomes.

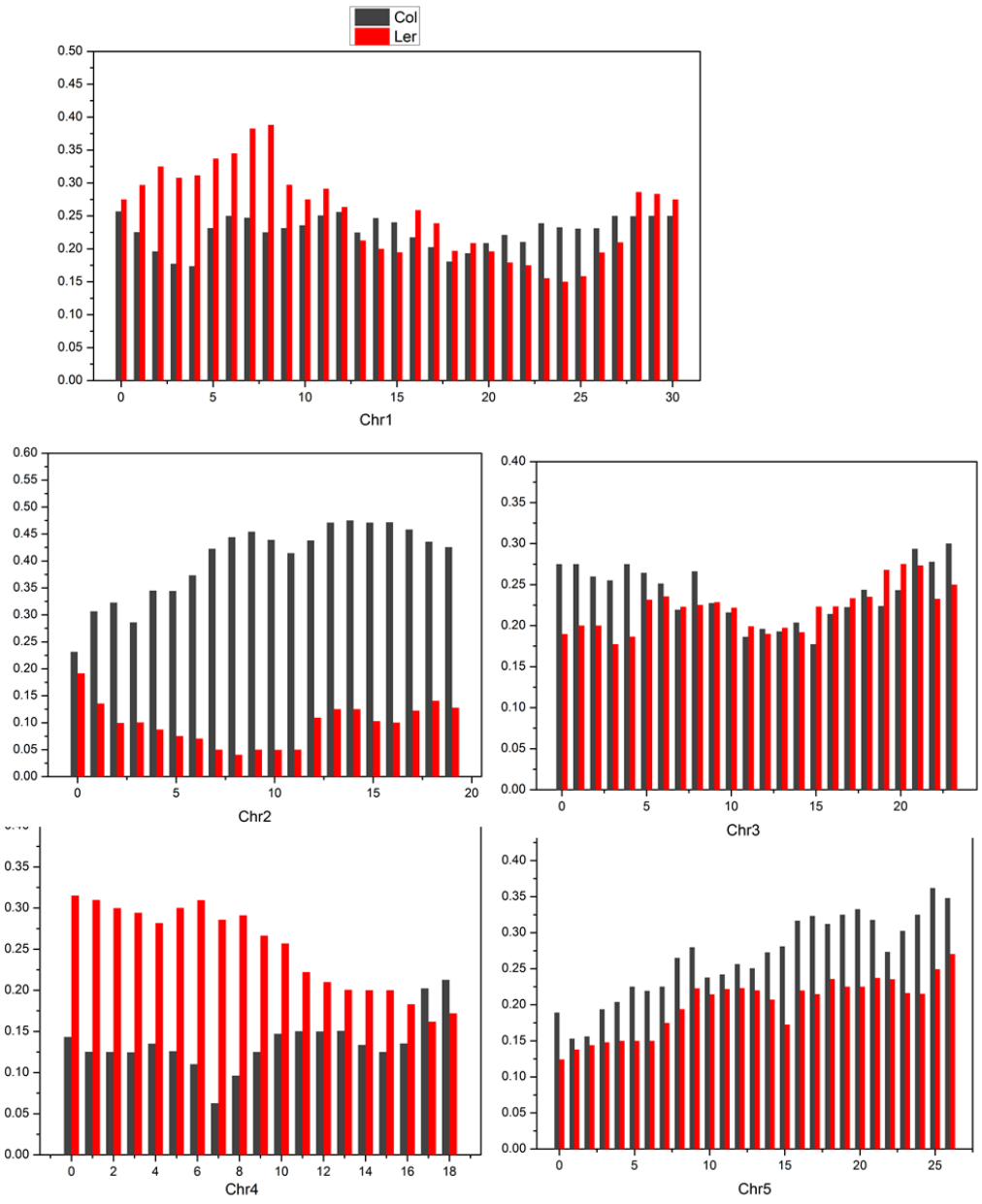
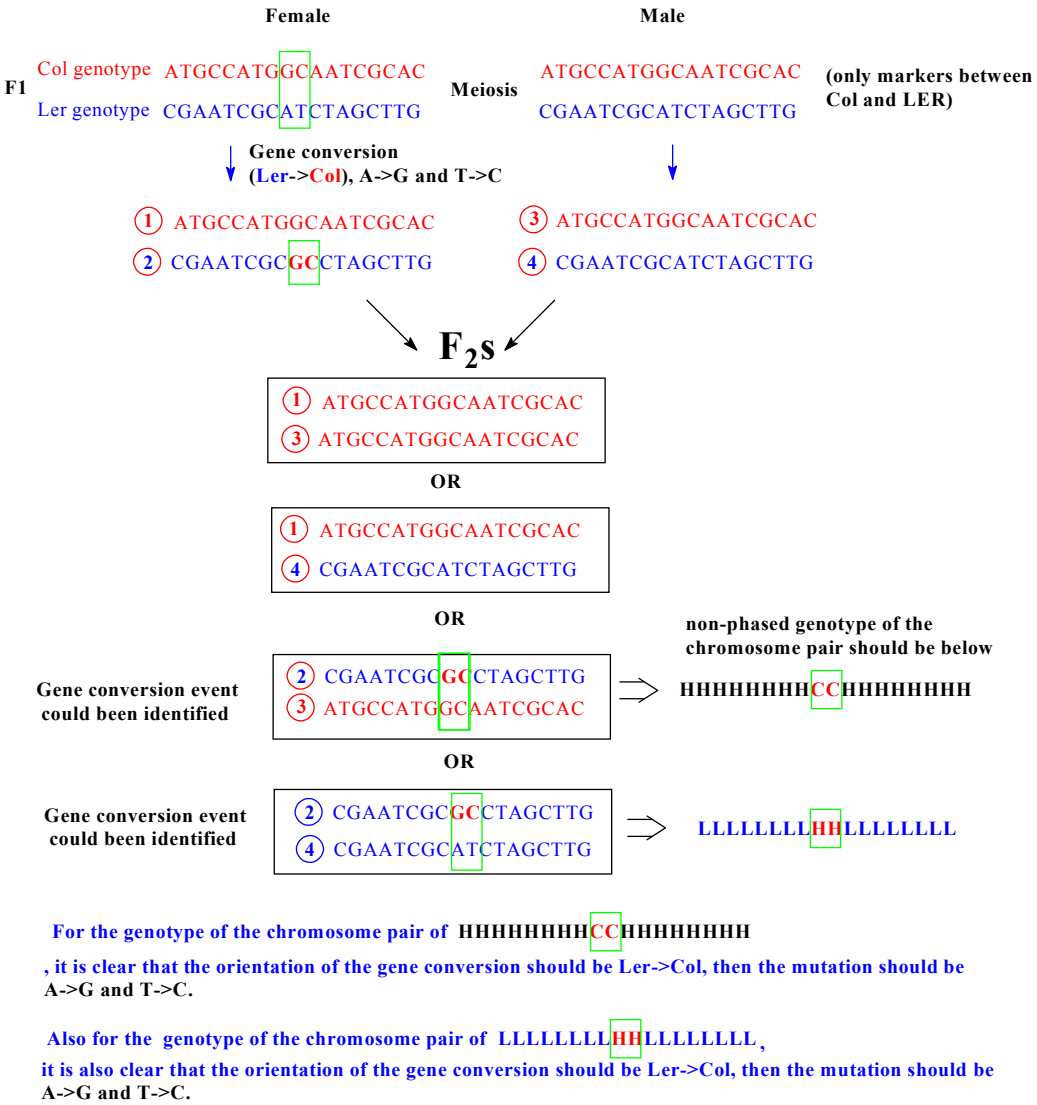


Fig. S12. Inference of the direction of SNPs

For any putative gene conversion event we can assess whether the conversion converted an AT to a GC or vice versa. Here a Ler->Col conversion events means the Ler genotype is replaced by the Col genotype, A->G means an A converted to a G etc.



For the genotype of the chromosome pair of HHHHHHHHHCCHHHHHHHHH , it is clear that the orientation of the gene conversion should be Ler->Col, then the mutation should be A->G and T->C.

Also for the genotype of the chromosome pair of LLLLLLLLHHLLLLLLLLL , it is also clear that the orientation of the gene conversion should be Ler->Col, then the mutation should be A->G and T->C.

Fig. S13. Evaluation of CO interference

The >500kb-track COs and >10kb-track COs are used to test the CO interference, respectively. In principle, only double COs on single chromosome could be used for measure of CO interferences (Salome, P. A. *et al.* 2011). However in our study, due to the lack of enough double COs, double and more COs were used. If one chromosome has 2 or more COs, the total physical distance of all CO-pairs (a) is used to calculate the mean (mean CO-pair distance) and s.d. (the associated standard deviation) of their distances. The function dgamma (chromosome length, shape, rate) in R is used to simulate the CO interference with gamma distribution (scale = $(s.d.)^2$/ mean; shape = $(mean / s.d.)^2$); b. Cross-over interference was calculated with >500 kb tracks of the COs, suggesting a positive CO interference affecting all chromosomes; c. When using tracks of COs with >10 kb, the figures do not meet the model of gamma distribution.

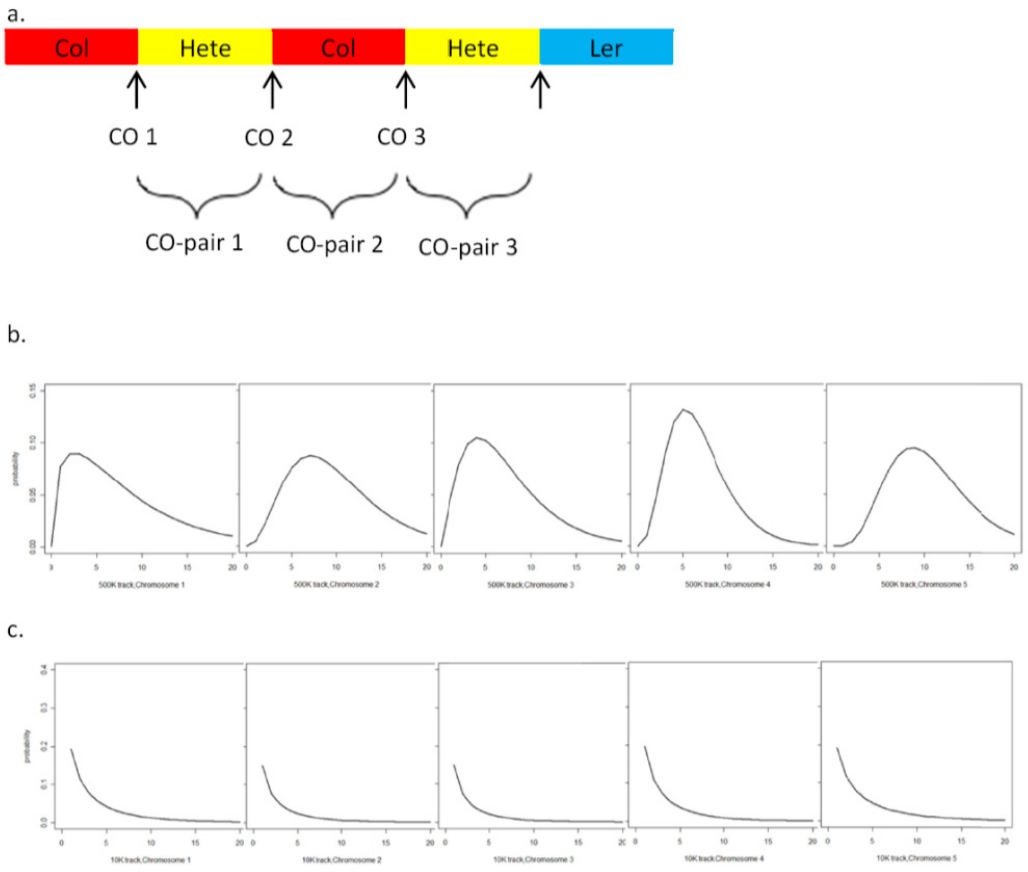
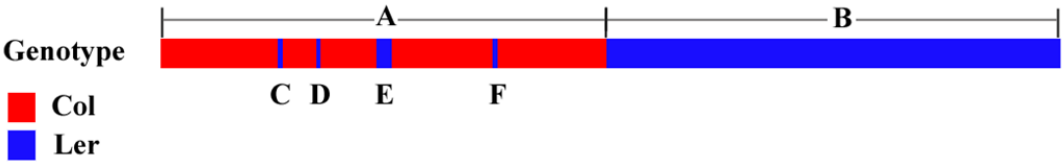


Figure S14. Inference of identification of crossover events. Each background is a tract of long COs when its net size is still ≥500 kb when excluding the total length of other genotype blocks, which in total must be <20% of a long CO. A small CO, identified from long COs and the other regions, can contain GCs but must still be from 10 to <500kb after excluding the GC blocks which in total must be <20%. All the spans of COs are used to calculate the proportion of H, C and L in a plant

Case 1: a background >500 kb



Assuming: A=520 kb, B=500 kb; C=1 kb, D=1 kb, **E=8 kb** (<10 kb), F=1 kb;

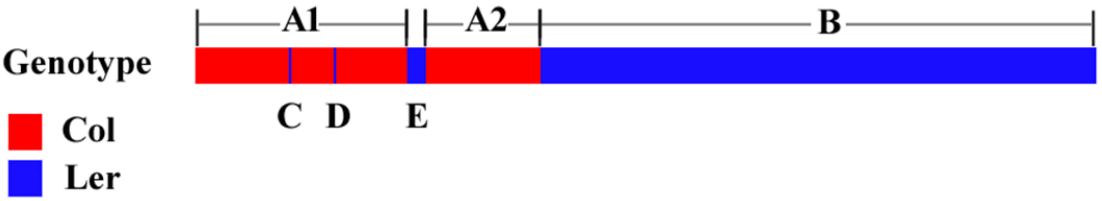
Region A: $\frac{C+D+E+F}{A} = \frac{1+1+8+1}{520} = 0.021 < 20\%$

and 520 kb -1-1-1-8 = 509 > 500 kb

Therefore, the genotype of region A in one chromosome should be Col.

Then, **one cross-over events** (between regions A and B) can be detected in this condition.

Case 2: a background <500 kb



Assuming: A1=150 kb, A2=100 kb, B=500 kb, E=50 kb (>10 kb); C=1 kb, D=1 kb;

(1) A1+E+A2=300 kb <500 bp;

(2) A1 = 150 kb >10 kb, A2 = 100 kb >10 kb, E=50 kb >10 kb;

Therefore, **one large** (between regions A2 and B; due to B ≥ 500kb) and **two small** (between regions A1 and E, E and A2; 10-500kb) **cross-over events** can be detected.

Other Supporting Information Files

Dataset S1 (XLSX)

Dataset S2 (XLSX)