BIOLOGICAL SCIENCES: Evolution

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, School of Life Sciences, Nanjing University, Nanjing 210093, 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

Corresponding author:

Dacheng Tian

School of Life Sciences, Nanjing University, Nanjing 210093, China

Telephone number: 86-25-83686406

E-mail: dtian@nju.edu.cn

Laurence D. Hurst

Department of Biology and Biochemistry, University of Bath, Bath, U.K, BA2 7AY

E-mail: bssldh@bath.ac.uk

Jian-Qun Chen

School of Life Sciences, Nanjing University, Nanjing 210093, China

E-mail: chenjq@nju.edu.cn

Abstract

The evolutionary importance of meiosis may not solely be associated with allelic shuffling owing to crossing-over but also to do with its more immediate effects such as gene conversion. While estimates of the crossing over rate are often well resolved, the gene conversion rate is much less clear. In *Arabidopsis*, for example, next generation sequencing approaches suggest that the two rates are about the same, which contrasts with indirect measures, these suggesting an excess of gene conversion. Here we provide analysis of this problem by sequencing 40 F_2 *Arabidopsis* plants and their parents. Small gene conversion tracts, with biased GC-content, represent over 90% (probably nearer 99%) of all recombination events. The rate of alteration of protein sequence owing to gene conversion is over 600 times that owing to mutation. Finally, our analysis reveals recombination hotspots and unexpectedly high recombination rates near centromeres. This may be responsible for the previously unexplained pattern of high genetic diversity near Arabidopsis centromeres.

\body

Introduction

When considering the population genetic impact of recombination, classical theories predominantly concentrate on the impact of allelic shuffling, mediated by crossing over, and the effect this has on linkage disequilibrium and, in turn, the effect the fate

of one allele has on its genomic neighbours (1). However, when programmed double strand breaks (DSBs) are introduced into chromosomes to initiate meiotic recombination, both crossover (CO) and non-crossover (non-CO) recombination events can occur. Non-crossover mechanisms, such as synthesis-dependent strand annealing (SDSA), typically result in gene conversion (GC) (2). Gene conversion skews segregation rates of alleles and thus has immediate effects on allele frequencies. While such direct consequences of recombination are generating more interest (3), relatively little is known about the rates of gene conversion, although its long term impact on sequence evolution is thought to be profound and phylogenetically widespread (4, 5). Despite this, in the construction of cross-over maps from linkage disequilibrium data, gene conversion events are typically ignored, being treated as though they were genotyping errors.

While many studies across diverse taxa have investigated the abundance and distribution of cross-overs during meiosis, few studies have resolved gene conversion rates, largely because such analysis is challenging. Based on tiling microarray data, an average of 90.5 COs and 46.2 non-COs were observed per meiosis in yeast, matching an estimate of 140-170 DSBs per meiosis (6). This contrasts with what is seen in mammals, where gene conversion events considerably outnumber crossover events (7).

Investigations in *Arabidopsis* have resulted in highly consistent estimates as regards cross-over events with under ten (3.74-8.3) per meiosis (8-11). Similarly, the most recent report, employing next-generation sequencing, revealed 9 COs per meiosis (9). According to the analyses in humans and yeast, meiotic gene conversion events typically have tract lengths less than 2kb (12, 13), commonly smaller (9). The small size of GCs makes them all but impossible to be detected in nearly all of the prior recombinational analyses for our species, as markers were on average usually hundreds of kb or a few Mb apart. Next generation sequencing (NGS) approaches can potentially be influential in this arena allowing markers every few hundred base pairs

to be employed. The one recent NGS analysis suggested there to be as many crossover events as gene conversions (9), making *Arabidopsis* more like yeast. This direct estimate, however, disagrees with indirect inferences. An immunolocalization study (14) suggests in excess of 200 recombination events per meiosis, while another (15) suggests a more modest 120-140. Assuming that these events mostly reflect non-crossover recombination events, this suggests a considerable excess of gene conversion compared to CO.

Here we employ NGS to provide a robust direct estimate of the rate of gene conversion in *Arabidopsis*. The above discrepancy between direct and indirect estimates of gene conversion rates may reflect little more than the difficulty of detecting gene conversion events through NGS if sequence quality is poor. Both density and accuracy of sequence are critical to detect a full spectrum of recombination events. With this concern foremost, we sequenced 40 Arabidopsis F_2 plants and their parents, Col and Ler, with unique sequencing strategies, incorporating high coverage, replicate independent sequencing, and long paired-end reads in long inserts. These strategies reveal abundant GCs in accord with, or possibly in excess of, the prior indirect estimates. We incidentally discover an unexpectedly rich world of recombination in and around centromeres. This may help resolve a prior paradox of *Arabidopsis* biology, namely why it is that its centromeres are unusual in having high levels of diversity (16, 17). Hotspots for recombination are also identified.

Results

We crossed two inbred strains (Col and Ler) to generate F_1 hybrids. These F_1s were self-crossed to generate an F_2 (Fig. 1a). We infer a recombination event (crossover or gene conversion) in the F_1 meiosis when in the F_2 progeny a run of markers from one strain switches to those from the other (see Methods). Were a recombination event to occur with matching breakpoints in both male and female meiosis our approach could be misleading. However, with abundant (>300,000) and well

scattered markers 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 $1000/300000^2 = 1.1 \times 10^{-8}$. We can thus identify almost every recombinational event (examples in Fig. 1b, compendium in Fig.S1).

Identification of accurate markers. To guarantee accuracy of markers, multiple stringent strategies were employed. First, 33 of 40 F_2s, from Col/Ler F_1 heterozygotes, were independently library-constructed and sequenced 2-3 times with high coverage (2×21.2× or 3×32.3×) and long paired-end reads (2×100 bp in 500bp inserts; Tables S1-S2). The other 7 F_2s were sequenced only once (Table S1). With high sequence quality and the addition of a second or third round, SNP calling and recombination block identification are expected to be almost 100% accurate (Tables S3-S4 and Dataset S1). Second, each of four Ler and three Col plants was sequenced 2-3 times with high coverage (up to 3×31.5× per plant). These sequences, combined with each parental genome being sequenced with 824× coverage in the 40 F_2 plants (Table S1), allow us to construct two accurate parental genomes, based on the well-sequenced Col. Third, three software packages (Novoalign, Shore and Stampy) were used to independently call SNPs against the reference. These filters resulted in a total of 586,231 SNPs identified by at least 2 of the software packages and 41,743 1-3 bp Ler deletions, these being identified by Shore alone. As a negative control for recombination events we sequenced a mixture of Col and Ler DNA (Table S4).

To be yet more stringent, we further refined a gold standard set. Only markers (a) identified by all three softwares (deletions only by Shore), (b) observed in >80% of 75(=7 + 31×2 + 2×3) sequenced F_2 genomes in corresponding heterozygous regions, and (c) concordant with the 461,070 identified previously (18), were considered to be adequate. This set comprises 373,614 SNPs and 41,743 1-3 bp Ler deletions, a total of 415,357 markers, representing on average one every 289 bp. These gold standard markers were used to detect COs and GCs. Over 3000 of these markers were sampled

for PCR amplification followed by Sanger sequencing confirming >99.7% of them (Dataset S1). From the initial 586,231 SNPs, the 212,617 eliminated in the second round of processing to generate the 373,614 gold standard SNPs, the less reliable SNPs, were used to corroborate the identified GCs.

Estimates of crossover rates accord with lower resolution studies. The blocks of runs of markers from a given genome are expected to come in a variety of lengths dependent on the manner of their creation. Spans >10kb we assume to be the outcome of crossing over. To enable comparison with prior studies (8-10, 12, 13), we group these events into long (>500 kb) and short spans (10kb-500 kb). The average number of long blocks are limited (8.4 or 3.6 cM/Mb per genome; Table 1) and consistent with prior reports for which a 500kb interval was about the limit of resolution (8-11). The position of every long CO is unique (Fig. S1a-e). The number (28.8 per genome) of small blocks identified is about four times greater than the long ones.

With 20 or more markers (77.4 on average) every CO can be clearly identified. However, a GC event may involve relatively few markers. False positive GC events are thus a possibility and the estimate of the number of GC events will be sensitive to the stringency of analysis. We start by attempting to define lower bounds for the total number of GC events (Table 2 and Tables S5-S6).

Estimating lower bounds for the number of gene conversion events. To define a stringent set we require that each GC tract must be between 20bp and 2kb and contain two or more of the gold standard markers, each of which must, in addition, be identified in all independently sequenced genomes for the same F_2 individual (the seven F_2 genomes sequenced only once were excluded from this analysis). In addition, we require that these GC tracts must be consistent with the slightly less reliable SNPs (see Methods). Even with these severe filters we identified 265.3 gene conversion tracts per meiosis (Table 2). The analysis of the two negative controls showed that the

error rates for GC identification to be 0 – 5% (Table S4), consistent with PCR results in Table 2. Our analysis hence largely confirms the prior higher estimates based on immunolocalization data (14) and suggests that at a minimum 90% of recombination events are resolved as gene conversions.

Confirmation of lower bounds. To confirm these estimates we sequenced 126 regions containing GC events from two randomly sampled F_2 plants (c52 and c66). This confirmed 100% of them in Col-homozygous regions, an average of 72.3 per genome in these regions. As the individual proportions of Col and Ler background (Table 2) are significantly correlated with their GC numbers (r=0.645 and 0.797; P<0.01 respectively), we can extrapolate the numbers seen for the Col homozygous background based on a random occurrence model. Given that the Col-homozygous regions account for 25.2% of the sequence, 286.9 GCs (=72.3/0.252) are predicted genome wide (Table 2). Notably, the two samples (c94 and c95 in Table 2 and S1), sequenced thrice with ~100×coverage per plant, produce similar GC numbers when the same extrapolation form Col homozygous regions is employed (288.5 =65.5/0.227).

Given difficulties in unambiguously assigning sequence to repeat rich areas of the genome, we checked that our estimates are not repeat-associated mapping artifacts. We determined for all GCs identified in Table 2 whether they are in repeat or non-repeat regions. The majority are fully or partially in non-repeat regions, suggesting that they are not products of repeat-associated mapping errors (Tables S6-S7). Even if we assume that only GC events in non-repeated regions can be identified unambiguously, we find that there are circa 161 GC events. As non-repeat regions account for 77.56% of the genome, this suggests there to be 207 GC events per genome (assuming a random occurrence model).

Inclusion of very short and long tracts modestly increases GC number estimates. The above analyses ignore possible very short (<20bp) and long GC (>2kb) tracts.

Applying the strict requirements above, but requiring the tracts to be 20bp to 10kb long (rather than 20bp to 2kb), we detect a further 30 tracts that likely reflect GC events (Table 1). If we add in very small (2-19bp) but well described GC tracts (Table S5) an additional 73 GC events are estimated. For this analysis we again require 2 reliable markers in the span and consistency on adding in intervening but less reliable SNPs. We detected 18.5 small GC events in Col homozygous background, 100% of which were confirmed by Sanger sequencing. Based on the equation in Table 2, 73 GCs per genome are expected. Including these longer and shorter events thus increases the estimate 35% to 390 GCs per meiosis (≈287+30+73).

Estimating upper bounds for the number of gene conversion events. In the above analyses we ignore the possible GC tracts supported by few markers. These are harder to confidently estimate. 2377 possible incidences of GC per meiosis were identified in a set where either one in the first round of sequencing or one or more markers in second round are required to define a GC event (Table S8). This provides one upper estimate.

An alternative estimate can be deduced via extrapolation. When focusing on the number of markers involved in each CO and GC, there is a smooth distribution relating the number of occurrences to the number of markers involved (Fig. S2). Assuming that the GC tracts with four or more markers are real, we can fit a frequency curve that, by extrapolation, can give a prediction for the GC tracts with 1-3 markers as approximately 2800 per plant. Together with the 207 tracts with more than 3 markers (used to define the frequency distribution), there may thus be >3000 gene conversion events, i.e. 80 times more gene conversion events than crossovers. When GCs in repeat regions are discarded from the set from which extrapolation is based there are still >2000 gene conversion events genome wide, meaning 50 times more gene conversion events than crossovers.

Assuming an upper bound of around 3000 gene conversion events, we conclude that

between 90% and 99% of recombination events are gene conversion events. The higher estimate is somewhat in excess of the most extreme prior indirect estimate. For estimates of the mean tract length and proportion of the genome that are part of such tracts (19) see Table S9. In terms of the total span of DNA recombined, the impact of COs is greater than GC, even using our upper estimate, as the span of each CO event is so long.

Evidence of abundant peri-centromeric recombination. The large (>500kb) and small (10-500kb) cross-over blocks have quite different patterns of distribution on chromosomes. The long COs distribute almost randomly along chromosomes (Fig. 2a, S2a-e and S3a) their density hence correlating with chromosomal length (P =0.038; Fig. S4a). This is as classically reported for coarsely resolved crossover maps. Unexpectedly, of the small crossover spans, 72.6% occur in pericentromeric regions (within 2 Mb; Fig. 1b, 2a & S3b), classically considered to be recombinational deserts (20, 21).

Can we be confident that this unexpected result is not a build or sequencing artifact? In negative controls (Table S4) no block with 10-500kb was identified. These pericentromeric blocks thus cannot be explained by sequencing errors. Moreover all or some of the markers in 71% of 10-500kb blocks are contained in non-repeat regions, which can be mapped without ambiguity. Furthermore, many of 10-500kb blocks, including those in pericentromeric regions, are located within a pure Col or Ler background as shown in Fig. 1b, indicating that they are unlikely to be from mapping errors. In fact, the high mapping quality can be clearly displayed by the long paired-ends reads in 500bp long inserts at the border of CO transitions (Fig.S5), which further indicates that they are neither artifacts nor rearrangements in Ler compared to Col. In addition, we examined by PCR followed by Sanger sequencing the markers for 9 small COs, 7 of which are peri-centromeric. We confirm all the markers for all of the 9 blocks.

Employing the most robustly defined GC events we observe a similar excess of gene conversion events near centromeres. To verify this we employed a single-stranded cloning strategy. This can resolve the sequence for each sister chromosome at the same region. 10 candidate GCs putatively near pericentromeric regions were analysed by this strategy with all 10 being confirmed as residing in proximity to centromeres (Table S10).

Recombination events are in domains of high diversity and low gene density. Breakpoints of both COs and gene conversion events are often located in regions with high diversity (Fig.S6). As regards GC events this is in part a definitional necessity (DSB followed by SDSA may occur in homozygous stretches but in the absence of variable markers involves no allelic gene conversion). However, through simulation we observe that there is more diversity than expected by chance, allowing for the definitional bias (Fig.S6). Given the long span of cross-over events allied with our very high marker density, we can be confident at having identified all cross-overs and thus have an unbiased assay of the location of breakpoints. The excess diversity in the vicinity of COs is thus neither easily accountable in terms of ascertainment bias nor definitional necessity.

GC and CO events both tend to be more prevalent in gene poor regions and, as commonly reported (8, 10), tend to be intergenic (Fig.S7). However, on average 16.1 GCs, containing 32.3 non-synonymous SNPs and 17.2 Ler deletions, are detected per meiosis. The rate of non-synonymous conversion is approximately 675 times higher than the non-synonymous mutation rate reported in laboratory conditions (a total of 11 detected in 5 individuals for 30 generations (22)). With 991 markers in intergenic DNA converted per meiosis the effects on sequence affecting gene expression may be more profound. These numbers, however, apply to our F_1 plants. For highly selfing wild populations of *A. thaliana* the number of heterozygous sites in any given meiosis is likely to be low and hence the actual conversion rate also much lower.

Shared recombination events. About 67% of GCs and 89.4% of the small COs are shared in two or more individuals and their track/spans lengths are on average 402 bp and 36 kb (98.5% of them <100kb), respectively. This rate of sharing is significantly greater than the expected value (1.1×10^{-8}) in a random occurrence model. A repeatable CO span (26 kb long) is shown at 3-4 Mb position with a frequency of 18/80 chromosomes in Fig. 1b. Each of the shared GC or small length CO loci is seen in 6.3% and 8.6% of individuals, respectively. In total, 59.3% of shared GCs/COs were located or partly located within non-repeat regions, suggesting that the majority of them could not be repeat-associated mapping artifacts. Based on the trees for shared GCs or COs (Fig. S7), every individual has a different set, suggesting independent occurrence. As expected given the location of small sized CO events, the shared events are common around centromeres (Fig. 2 and S9) and roughly coincident with the recombination hot-spots reported recently (15). Sanger sequencing of PCR products, from unique pairs of primers in the *Arabidopsis* genome, confirmed that 17 putatively shared loci sampled (8 GCs or 9 small COs) were indeed shared among different plants. Fig. S10 shows two confirmed examples where the PCR and sequencing results from multiple pairs of primers, including ones crossing the border of the breakpoint, were consistently positive among many plants.

MEME analysis reveals that conserved motifs with several copies per sequence are often located in 300-bp regions surrounding a GC (or CO) or within the converted sequences (Table S11), which could be associated with the frequent occurrence of GCs or small COs at specific positions among individuals (23).

Distorted segregation ratios and GC-content biased gene conversions. As with prior reports from interline crosses (8, 10) we find strong evidence for distorted segregation ratios (Fig. S11), with three of the five chromosomes significantly different from Mendelian expectations (Fig. S11). They are either Col-dominant (chromosome 2 and 5) or Ler- dominant (chromosome 4). The underlying cause is

unclear but may reflect meiotic drive, genes under selection for early viability (10) or genetic incompatibility (8).

Meiotic gene conversion is thought to be biased towards nucleotides G and C in the great majority of eukaryotes (4). If *u* is the number of AT→GC SNPs per A or T and *v* the number of GC→AT SNPs per G or C, then we can consider the ratio *u/v* (Fig.S12). For the gene conversion tracts (20bp-2kb) this is 1.22 which is significantly greater than the null (unity) (from randomization: $P < 0.0001$), consistent with biased gene conversion increasing the frequency of AT→GC SNPs as seen in other taxa (4). By contrast, the 10kb-500kb spans (u/v=0.96) are no different ($P = 0.18$) from unity, suggesting that these spans might be crossovers rather than gene conversion events.

Discussion

Our analysis supports early indirect approximations to the number of gene conversions events, strongly rejecting the one prior next generation sequencing based estimate which suggested equal numbers of CO and GC events (9). This rejection is rendered yet more robust by our conservative assumption that 10kb-500kb events are COs not GCs. The cause of these mid-sized blocks is, however, yet to be fully resolved. A few similarly sized recombination events were observed previously (9) and assumed to reflect an interference-free mode of crossing over. Consistent with this we find no evidence for interference for small COs (Fig. S13). The same is true for the shared COs which are almost only present in <100kb spans. Similarly, we find no evidence for distorted G and C content, consistent with an absence of gene conversion. In principle, however, tracts a little over 10kb may reflect gene conversions created by helicase-mediated resolution of double Holliday Junctions (14, 33) (rather than through the SDSA mechanism), tract lengths for which are unknown or could be mitotic conversion events (13, 32). However, mitotic conversion rates are typically 10^4-10^5 lower than meiotic conversion (32) making the latter unlikely. Unfortunately, with segregation distortion common across the chromosomes, we cannot perform a segregation analysis and so cannot definitively

conclude that these are crossover events.

The commonality of GC events has implications for population genetic inferences. Regular gene conversion events are likely to reduce the structure of linkage disequilibrium (24) and will have a strong effect on the distribution of nucleotide polymorphisms. Adding gene conversion to genetic models will make them more appropriate for the inference of population history from linkage disequilibrium (24). Crossover rate inference from linkage disequilibrium data is robust to moderate gene conversion rates (treating it as genotyping errors) and would have little or no problem were the recent (9) lower end estimate correct. With our new more extreme estimates, caution is advisable in application of such methods.

Why so much gene conversion? The abundant GCs in *Arabidopsis* suggest that plants are more like mammals (7) than yeast, the latter having relatively common crossing over compared to gene conversion (6). This difference between taxa we suggest may reflect differences in repeat content, as repetitive sequences are a source of genomic instability during meiosis (25), owing to non-allelic homologous recombination (26). Compared with COs, non-crossover (e.g., via SDSA (2)), which yields the most gene conversions (GCs), pose the least genomic threat among mechanisms that repair DSBs (27). Analysis of repeat poor genomes of multicellular species (e.g. *Oikopleura*) will be informative.

Mechanistically the above suggestion may require that organisms scan the local sequence environment to determine how to resolve a DSB. In a related vein, Dooner (31) has suggested that the local diversity levels may mediate the choice between gene conversion and crossing over following a DSB. However, we find no significant differences in the distribution of SNPs and indels around GCs and COs (t-test, $P = 0.42$). While this fails to support Dooner's conjecture, our evidence is not decisive as we consider only small indels ($\leq$3 bp) while Dooner's hypothesis concerns larger indels in addition.

Pericentromeric recombination may explain prior unusual observations. The apparently high frequency pericentromeric recombination events may explain some prior data. First our data could explain why the crossover frequencies were seen, in lower resolution maps, to increase adjacent to the centromeres (8). When examining the distribution of all COs (Fig. 1b), more frequent recombination can be identified between two arms of chromosome 1, due to the denser distribution of small COs around the centromere. Given that many of these COs are double crossovers, they are unable to cause a recombination between the two arms (or part of the arms). When excluding those COs, however, the potential frequency can still be as high as about 1/4 on this chromosome, suggesting a partly free exchange between two arms.

Second, the finding of abundant recombination, including crossing-over, near centromeres helps resolve a prior paradoxical result. In many taxa there is a positive correlation between intra-population diversity and genomically-local recombination rates (3). In *A. thaliana* (16, 17) and the outbred *A. lyrata* (28) there is, unusually, high sequence diversity near centromeres. This has been considered contra to classical theory as centromeres were assumed to have low crossover rates, thus prone to weak Hill-Robertson interference reducing diversity (28). Reduction of such interference under high crossover rates may not, however, be the full explanation. *A. thaliana* is a near obligate selfer and as such crossover should have relatively little effect on Hill-Robertson mediated diversity. Moreover, that we observe that the breakpoints of both COs and gene conversion events are often located in regions with high diversity (Fig.S6) suggests instead that either a) there is a preference for double strand breaks to occur in domains of high polymorphism or b) double strand breaks promote polymorphism. The latter may be mediated by a coupling between DSB repair and the mutation process (29) or reflect the activity of biased gene conversion which can increase load at gene conversion hotspots even if inbreeding levels are very high (30). Biased gene conversion is supported by SNP analysis (see above) and from the finding that in the 100-bp sequences around the tracts of gene conversion,

the GC-content (0.368) in shared loci, with two or more GCs among different individuals, is higher than that at unshared (0.345) or randomly sampled loci (0.348; $P = 14\times10^{-10}$). Recent evidence (22) supports a higher mutation rate in proximity to centromeres.

Materials and Methods

Plant material. The F_1 seeds were obtained from female Col individuals crossed with Ler male plants, both of which were either from a single Col or Ler seed. Thousands of F_2 plants were grown from seeds obtained from selfed F_1 plants. Finally 40 F_2, 4 Ler and 3 Col plants were used to extract DNA by the CTAB method for genome sequencing.

Re-sequencing. Paired-end sequencing libraries with insert size of 500 bp were constructed for each plant according to the manufacturer's instructions. Then 2×100 bp paired-end reads were generated on Illumina HiSeq 2000. Finally 47 plants (Table S1) were re-sequenced with >21.2× coverage and high quality for each by BGI-Shenzhen. To increase the accuracy, 2 parental plants, 33 F_2 plants (7 of the 40 were without enough DNA) were independently constructed for libraries and sequenced 2-3 times with the same coverage (3×29.8× for each of parent, 3×32.3× per plant for 2 F_2, and 2×21.2× for the remaining 31 F_2; Table S1). For the other 5 parental plants, an equal amount of DNA from any two of the five parental plants was mixed to perform re-sequencing, e.g., Sample_C_1C_2, C_2L_1, L_1L_2, L_2L_3 and L_3C_1 (C stands for Col and L for Ler). The two mixed Col-Ler samples (i.e. C_2L_1 and L_3C_1) were used as negative controls. The other parental samples were employed as references for SNP calling.

Marker identification. The Col genome (TAIR9) was downloaded from TAIR website (ftp://ftp.arabidopsis.org/home/tair/Sequences/whole_chromosomes). The assembly Ler scaffolds, SNPs and indels were downloaded from 1001 genomes

(http://1001genomes.org/projects/assemblies.html). To identify reliable markers, SHORE (22), Novoalign (www.novocraft.com) and Stampy (34) were used to independently call SNPs against the reference Col genome for our sequenced parental plants.

Identification of crossover events. Based on 415,357 markers, regions along chromosome pairs were converted into blocks of genotype H (heterozygosity), C (Col homozygosity) and L (Ler homozygosity) by searching for the genotype switching points, e.g., H→C, H→L, L→H or L→C. The <500-kb blocks were first ignored to construct the genotype background (for details see Fig. S14).

Detection of gene conversions. More strict criteria were applied to detect GCs. A quality evaluation for two or three rounds of independent sequencings was carried out for each chromosome pair. We split the genome into 300-kb non-overlapping windows and for each window calculated the proportion of markers that were in disagreement between the independent sequencings. Only those windows with less than 5% disagreement were used. 31 out of 33 F_2 plants have almost no regions with >5% difference. For the remaining two plants the first round of sequencing had variable quality while the second had much higher quality. We retained only domains with high quality in both (<5% discordance).

Only the 415,357 gold standard markers were used for GC detection. After identification of candidate GCs, however, the other SNPs were used to distinguish whether the form of a GC tract remains the same or changes to a different one when adding more SNPs between the gold standard markers. For example, a block was identified as pure Col (C-C-C) by three markers and two more non-gold standard SNPs were inserted between these markers. Imagine that the pattern was changed into C-H-C-H-C or C-H-H-C-C. In either case the original candidate GC was discarded. In the latter instance a new GC (C-C) was, however, accepted. A large number of GC candidates were randomly sampled for checking via PCR and Sanger sequencing.

Each pair of primers in all PCRs was unique in the well-sequenced Arabidopsis genome.

Estimation of sequence quality. To ensure the quality of our sequences and genome assembly, we employed numerous methods. First, two mixtures of Col and Ler DNAs were separately sequenced as negative controls (Table S4) to identify false positive recombination events (table S2-S4). Second, there are multiple rounds of independent sequencing for 33 F_2 plants, which can be used to estimate the numbers and types of gene conversions independently between the two sets of sequences for the same plant. The differences between multiple independent sequencings provide an estimate of the range of errors (Table S3-2).

We also compared our results with those from Lu et al. (9). The genome sequences of four progeny from one meiosis provided by them were used to (1) compare the sequencing strategy and quality with this study (Table S2) and (2) identify the possible GC events in Lu's samples by the criteria used in this study for different sets of gene conversions (Table S3-1). Note, while there were 8 progeny in total only four could be used.

Controlling for repeats. 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.

Checking borders. We analysed in detail the transition borders of 10-500k COs in sample c94 and c95 (each with 97× coverage). To directly see whether a CO might be incorrectly mapped or built (Fig. S5) we considered instances where there are two or more markers within 400bp of each other but with >100-200bp between them. These were analyzed in detail for all the paired reads (2×100bp) in long inserts (500bp). In total, there are only 12 COs with enough markers, enough space and an

unambiguous border to do such analysis (based on the software inGAP-sv (http://ingap.sourceforge.net/)). By this analysis, we confirmed all 12 COs. For a full list of recombination events see Dataset S2.

Acknowledgements This work was supported by National Natural Science Foundation of China (30930049, 30930008 & 31071062) and NSFC of Jiangsu province (BK2011015) to D.T. or J.Q.C. L.D.H is a Royal Society Wolfson Research Merit Award holder.

Footnotes

[1]S.Y., Y.Y., and L.W. contributed equally to this work.

[2]To whom correspondence may be addressed. E-mail: dtian@nju.edu.cn, bssldh@bath.ac.uk or chenjq@nju.edu.cn

Author contributions: D.T., L.D.H., S.Y., J.Q.C. and W.W. designed the experiments and analyses. S.Y. organized all aspects of the project. Y.Y., L.W. and S.Y. analyzed the sequence data. J.L. and H.L. prepared plant materials and performed experimental confirmations. D.T. and L.D.H. wrote the paper and L.D.H. finalized the paper.

The authors declare no conflict of interest.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database [accession no. XXXX].

Supporting Information

Figs. S1 to S14

Tables S1 to S11

Other Supporting Information Files

Dataset S1 (XLSX)

Dataset S2 (XLSX)

References

1. Hill WG & Robertson A (1966) The effect of linkage on limits to artificial selection. *Genet Res* 8(3):269-294.
2. Allers T & Lichten M (2001) Differential timing and control of noncrossover and crossover recombination during meiosis. *Cell* 106(1):47-57 .
3. Webster MT & Hurst LD (2012) Direct and indirect consequences of meiotic recombination: implications for genome evolution. *Trends Genet* 28(3):101-109 .
4. Pessia E*, et al.* (2012) Evidence for widespread GC-biased gene conversion in eukaryotes. *Genome Biol Evol* 4(7):675-682 .
5. Innan H (2002) A method for estimating the mutation, gene conversion and recombination parameters in small multigene families. *Genetics* 161(2):865-872 .
6. Mancera E, Bourgon R, Brozzi A, Huber W, & Steinmetz LM (2008) High-resolution mapping of meiotic crossovers and non-crossovers in yeast. *Nature* 454(7203):479-485 .
7. Paigen K & Petkov P (2010) Mammalian recombination hot spots: properties, control and evolution. *Nature reviews. Genetics* 11(3):221-233 .
8. Salome PA*, et al.* (2012) The recombination landscape in Arabidopsis thaliana F2 populations. *Heredity* 108(4):447-455 .
9. Lu P*, et al.* (2012) Analysis of Arabidopsis genome-wide variations before and after meiosis and meiotic recombination by resequencing Landsberg erecta and all four products of a single meiosis. *Genome research* 22(3):508-518.
10. Giraut L*, et al.* (2011) Genome-wide crossover distribution in Arabidopsis thaliana meiosis reveals sex-specific patterns along chromosomes. *PLoS genetics* 7(11):e1002354 .
11. Toyota M, Matsuda K, Kakutani T, Terao Morita M, & Tasaka M (2011) Developmental changes in crossover frequency in Arabidopsis. *The Plant journal : for cell and molecular biology* 65(4):589-599 .
12. Chen JM, Cooper DN, Chuzhanova N, Ferec C, & Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 8(10):762-775 .
13. Judd SR & Petes TD (1988) Physical lengths of meiotic and mitotic gene conversion tracts in Saccharomyces cerevisiae. *Genetics* 118(3):401-410 .
14. Chelysheva L*, et al.* (2007) Zip4/Spo22 is required for class I CO formation but not for synapsis completion in Arabidopsis thaliana. *PLoS Genet* 3(5):e83 .
15. Sanchez-Moran E, Santos JL, Jones GH, & Franklin FC (2007) ASY1 mediates AtDMC1-dependent interhomolog recombination during meiosis in Arabidopsis. *Genes Dev* 21(17):2220-2233 .
16. Borevitz JO*, et al.* (2007) Genome-wide patterns of single-feature polymorphism in

Arabidopsis thaliana. *Proceedings of the National Academy of Sciences of the United States of America* 104(29):12057-12062 .

17. Clark RM, *et al.* (2007) Common sequence polymorphisms shaping genetic diversity in Arabidopsis thaliana. *Science* 317(5836):338-342 .
18. Schneeberger K, *et al.* (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National Academy of Sciences of the United States of America* 108(25):10249-10254 .
19. Mansai SP KT, Innan H (2011) The rate and tract length of gene conversion between duplicated genes. *Genes* 2:313-331.
20. Talbert PB & Henikoff S (2010) Centromeres convert but don't cross. *PLoS biology* 8(3):e1000326 .
21. Gore MA, *et al.* (2009) A first-generation haplotype map of maize. *Science* 326(5956):1115-1117 .
22. Ossowski S, *et al.* (2008) Sequencing of natural strains of Arabidopsis thaliana with short reads. *Genome research* 18(12):2024-2033 .
23. Horton MW, *et al.* (2012) Genome-wide patterns of genetic variation in worldwide Arabidopsis thaliana accessions from the RegMap panel. *Nature genetics* 44(2):212-216 .
24. Wall JD (2004) Close look at gene conversion hot spots. *Nature genetics* 36(2):114-115 .
25. Sasaki M, Lange J, & Keeney S (2010) Genome destabilization by homologous recombination in the germ line. *Nature reviews. Molecular cell biology* 11(3):182-195 .
26. Vader G, *et al.* (2011) Protection of repetitive DNA borders from self-induced meiotic instability. *Nature* 477(7362):115-119 .
27. Hicks WM, Kim M, & Haber JE (2010) Increased mutagenesis and unique mutation signature associated with mitotic gene conversion. *Science* 329(5987):82-85 .
28. Kawabe A, Forrest A, Wright SI, & Charlesworth D (2008) High DNA sequence diversity in pericentromeric genes of the plant Arabidopsis lyrata. *Genetics* 179(2):985-995 .
29. Kulathinal RJ, Bennett SM, Fitzpatrick CL, & Noor MA (2008) Fine-scale mapping of recombination rate in Drosophila refines its correlation to diversity and divergence. *Proceedings of the National Academy of Sciences of the United States of America* 105(29):10051-10056 .
30. Glemin S (2010) Surprising fitness consequences of GC-biased gene conversion: I. Mutation load and inbreeding depression. *Genetics* 185(3):939-959 .
31. Dooner HK (2002) Extensive interallelic polymorphisms drive meiotic recombination into a crossover pathway. *The Plant cell* 14(5):1173-1183.
32. Lee PS, *et al.* (2009) A fine-structure map of spontaneous mitotic crossovers in the yeast Saccharomyces cerevisiae. *PLoS genetics* 5(3):e1000410 .
33. Hartung F, Suer S, Knoll A, Wurz-Wildersinn R, & Puchta H (2008) Topoisomerase 3α and RMI1 Suppress Somatic Crossovers and Are Essential for Resolution of Meiotic Recombination Intermediates in *Arabidopsis thaliana*. *PLoS genetics* 4(12):e1000285.
34. Lunter G & Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome research* 21(6):936-939 .

Figure Legends

Fig. 1. Schemed patterns (a) and examples (b) of crossovers (>10 kb) in F_2 plants. The red and blue bars represent the chromosomes of Col and Ler, respectively. One recombination of 40 F_2 chromosome pairs (chromosome 1) is showed as an example from male and female meiosis. The CO highlighted with an arrow is shown in further detail in Fig. S5.

Fig. 2. **Distribution of (a) COs and (b) GCs on chromosome 1 and 2.** The long (>500Kb) and small COs (10-500 Kb) are showed separately. Shared and non-shared GCs (20bp to 10 Kb) are demonstrated by different lines. The centromere regions were represented as grey bars. Only those GCs in Table 2 were used.