Detection of genomic variations and DNA polymorphisms and impact on analysis of meiotic recombination and genetic mapping

Ji Qi,a,b, Yamao Chen,a,b, Gregory P. Copenhaver,c,d, and Hong Ma,a,b,e,1

Edited* by Elliot M. Meyerowitz, California Institute of Technology, Pasadena, CA, and approved May 13, 2014 (received for review November 23, 2013)

DNA polymorphisms are important markers in genetic analyses and are increasingly detected using genome resequencing. However, the presence of repetitive sequences and structural variants can lead to false positives in the identification of polymorphic alleles. Here, we describe an analysis strategy that minimizes false positives in allelic detection and present analyses of recently published resequencing data from Arabidopsis meiotic products and individual humans. Our analysis enables the detection of sequencing errors, small insertions and deletions (indels), and structural variants, including large reciprocal indels and copy number variants, from comparisons between the resequenced and reference genomes. We offer an alternative interpretation of the sequencing data of meiotic products, including the number and type of recombination events, to illustrate the potential for mistakes in single-nucleotide polymorphism calling. Using these examples, we propose that the detection of DNA polymorphisms using resequencing data needs to account for nonallelic homologous sequences.

Results and Discussion

In many species, large-scale SVs often involve identical or highly similar sequences that differ in chromosomal contexts between structural variation | genotyping | insertions–deletions | high-throughput sequencing

DNA polymorphisms are ubiquitous genetic variations among individuals and include single nucleotide polymorphisms (SNPs), insertions and deletions (indels), and other larger rearrangements (1–3) (Fig. 1A and B). They can have phenotypic consequences and also serve as molecular markers for genetic analyses, facilitating linkage and association studies of genetic diseases, and other traits in humans (4–6), animals, plants, (7–10) and other organisms. Using DNA polymorphisms for modern genetic applications requires low-error, high-throughput analytical strategies. Here, we illustrate the use of short-read next-generation sequencing (NGS) data to detect DNA polymorphisms in the context of whole-genome analysis of meiotic products.

There are many methods for detecting SNPs (11–14) and structural variants (SVs) (15–25), including NGS, which can capture nearly all DNA polymorphisms (26–28). This approach has been widely used to analyze markers in crop species such as rice (29), genes associated with diseases (6, 26), and meiotic recombination in yeast and plants (30, 31). However, accurate identification of DNA polymorphisms can be challenging, in part because short-read sequencing data have limited information for inferring chromosomal context.

Genomes usually contain repetitive sequences that can differ in copy number between individuals (26–28, 31); therefore, resequencing analyses must account for chromosomal context to avoid mistaking highly similar paralogous sequences for polymorphisms. Here, we use recently published datasets to describe several DNA sequence features that can be mistaken as allelic (32, 33) and describe a strategy for differentiating between repetitive sequences and polymorphic alleles. We illustrate the effectiveness of these analyses by examining the reported polymorphisms from the published datasets.

Meiotic recombination is initiated by DNA double-strand breaks (DSBs) catalyzed by the topoisomerase-like SPORULATION 11 (SPO11). DSBs are repaired as either crossovers (COs) between chromosomes (Fig. 1C), or noncrossovers (NCOs). Both COs and NCOs can be accompanied by gene conversion (GC) events, which are the nonreciprocal transfer of sequence information due to the repair of heteroduplex DNA during meiotic recombination. Understanding the control of frequency and distribution of CO and NCO (including GC) events has important implications for human health (including cancer and aneuploidy), crop breeding, and the potential for use in genome engineering. COs can be detected relatively easily by using polymorphic markers in the flanking sequences, but NCO products can only be detected if they are accompanied by a GC event. Because GCs associated with NCO result in allelic changes at polymorphic sites without exchange of flanking sequences, they are more difficult to detect. Recent advances in DNA sequencing have made the analysis of meiotic NCOs more feasible (30–32, 34); however, SVs present a challenge in these analyses. We recommend a set of guidelines for detection of DNA polymorphisms by using genomic resequencing short-read datasets. These measures improve the accuracy of a wide range of analyses by using genomic resequencing, including estimation of COs, NCOs, and GCs.

Significance

Genetic analyses require allelic markers, which are often DNA polymorphisms and can be analyzed by using short reads from high-throughput sequencing. Therefore, accuracy in genetic studies depends on correct identification of DNA polymorphic markers, but genomic structural variants increase the complexity of allelic detection and must be carefully accounted for to avoid errors. Here, we examine potential mistakes in single-nucleotide polymorphism calling caused by structural variants and their impact on detecting meiotic recombination events. Our results demonstrate that it is crucial to examine structural variants in genetic analysis with DNA marker detection by using short reads, with implications for a wide range of genetic analyses.

Author contributions: H.M. designed research; J.Q. performed research; J.Q. and Y.C. analyzed data; and J.Q., Y.C., G.P.C., and H.M. wrote the paper.

The authors declare no conflict of interest.

1To whom correspondence should be addressed. E-mail: hongma@fudan.edu.cn.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1321897111/-/DCSupplemental.

www.pnas.org/cgi/doi/10.1073/pnas.1321897111

PNAS Early Edition | 1 of 6
such mistakes (31, 34). Studies have used some of the strategies described here to avoid version events (presumably accompanying NCOs), and recent sequences as SNPs can lead to the misidentification of gene conversion events on the same chromosome (Fig. 2B). When meiotic products from a cross between individuals with large SVs are analyzed by using unassembled short reads from resequencing, reads from the non-reference ecotype (Ler, TAIR10 assembly; ref. 35) and Landsberg erecta (Ler), including copy number variants (CNVs), large deletions, insertions, and inversions (31, 36–38). These SVs significantly influence genotyping, particularly SNP calling. Here, we focus on SV involving transposable elements (TEs) and CNVs, because their effects on false positive calling of SNPs are substantial.

**Mapping Nonallelic Sequence Reads Causes Artifactual SNP Calls.** SVs between Col and Ler that include TEs (Figs. 1B and 2) create regions of high sequence similarity that map to different (non-allelic) chromosomal positions. When meiotic products from a cross between individuals with large SVs are analyzed by using unassembled short reads from resequencing, reads from the non-reference ecotype (Ler) can be misaligned to nonallelic positions on the Col reference genome, resulting in the misidentification of similar sequences as polymorphisms, including SNPs. Because these sequences are not allelic, they can assort independently if they are on different chromosomes, or be redistributed in the genomes of meiotic progeny by COs or independent assortment, the same chromosome (Fig. 2B). Centromeres are denoted by yellow dots.

**Fig. 1.** (A) SNPs and small indels between two ecotype genomes. (B) Possible types of SVs. Col genotypes are marked in blue and Ler in red. Arrows indicate DNA segments involved in SVs between the two ecotypes. (C) Meiotic recombination events including a CO and a GC (NCO). Centromeres are denoted by yellow dots.

**Fig. 2.** Paralogous sequences between two ecotypes and their effects on allele ratio estimation. Redistribution of SV-related paralogue DNA segments in meiotic progeny and consequent genotyping using short read mapping for paralogs on the same chromosome (A), or paralogs on different chromosomes (B). Col and Ler are marked in the same colors as in Fig. 1, whereas dashed arrows indicate actual sites of short reads which were misplaced at nonallelic sites in the reference genome (Col). (C) Read depth, mapping distance, and orientation of PE reads from Col, Ler, and the F2 on the reference genome (Col) around the SV (large blue arrows in both Left and Right). On the Right, PE reads are shown above the Col (blue) sequences; Col read pairs and most Ler read pairs are mapped normally, except those Ler read pairs flanking a deletion, shown here as distantly mapped red reads with blue linkers. Only one chromatid is shown for both Col and Ler. An F2 plant has a nonallelic (Col/Ler) genotype around position 1 and homozygous (Col/Col) around position 2. (D) PE read mapping from Col, Ler, and an F2 plant by Yang et al. (32) in a 9-kb window (16,657,200 – 16,666,200) on chromosome 1 of Col reference genome. Col reads mapped normally; PE mapping patterns for Ler reads indicate a deletion: A TE is surrounded by a group of reads pairs (linked by blue lines) that mapped further apart than expected, and another two group of reads pairs (marked by pink lines) mapped to different chromosomes.
to 220 per meiosis (39–41). Even if every DSB resulted in a GC, the reported levels appear to be much too high. Interestingly, Yang et al. reported that many of the large GC tracts (crossover of consecutive polymorphisms) occurred at the same locations in multiple meioses (32). Our reanalysis of the sequencing data of two parental and two F2 plants (32) indicated that more than 67% of the reported large GC tracts (2 Kb ~10 Kb; Table 1) (32) were events from different meioses and were repeatedly detected with exactly the same boundaries. The discordance between the estimated number of DSBs and the reported number of GCs, and the striking repeated occurrence of large GC tracts at the same loci, leads us to seek alternative explanations.

Resequencing of *Arabidopsis* genomes by Yang et al. produces paired-end (PE) reads from both ends of short DNA fragments of similar lengths in the sequencing library. When a genome (e.g., L*) that has deletions compared with the reference genome (e.g., Col) is resequenced by using PE sequencing, the regions flanking the deleted DNA are also present and can be sequenced as PE reads from the same fragment. Such PE reads can be mapped to sites in the reference genome that span a greater distance than expected from the DNA fragment lengths of the sequencing library (Fig. 2C), providing strong support for a deletion in L*.

Furthermore, the deleted L* sequence might be found at a different (paralogous) genomic location, possibly resulting from historic transpositions. These nonallelic L* sequences are mapped back to the Col reference, contributing to false SNP calls in subsequent GC analysis. When an F2* plant lacks a paralogous L* copy because of CO or chromosome reassembly, the heterozygous region would be misrepresented as having the Col genotype (Fig. 2C). The situation is illustrated by an example shown in Fig. 2D: We found a transposable element (TE; AT1TE54925) on chromosome 1 of Col (at nucleotide 16,659,688–16,664,330 bp) that has a paralog on chromosome 2, but not on chromosome 1, in L* (at ~1.3 Mb; ref. 35). An F2* plant, designated C95 by Yang et al., was of the Col genotype for the entire length of chromosome 2, thus lacking the L* copy of AT1TE54925. As a result, no reads for the L* version of the TE were mapped to chromosome 1, producing a Col genotype at that locus. However, the regions flanking the TE on chromosome 1 were heterozygous (Col/L*) at polymorphic markers, leading to the interpretation that the AT1TE54925 locus had experienced a GC. Accounting for limited chromosomal context information for the structural differences between Col and L*, we allowed us to identify inappropriate GC calls at this site in 15 of 40 meiotic offspring, including C95 (32).

Accurate detection of meiotic GCs by polymorphisms requires knowledge of genomic SVs between the two parental genotypes. Because of the complex nature of SVs (large reciprocal indels, CNVs), it is necessary to examine all available sequence features, including read depth, mapping distance and orientation of PE reads, and mapping boundaries revealed by split reads, to determine the types and quality of SVs (15). As described above, unexpectedly long distances between a pair of reads indicate deletions in the resequenced genome (e.g., L*), relative to the assembled reference genome (Col) (Fig. 2C). As in the Yang et al. data, a cluster of 33 PE reads from L* were mapped to positions at a distance of 5,100 bp on average, which was significantly longer than the average length of 474 ± 13 bp of the sequenced DNA fragments in this study (P value <<10^-10 using the Kolmogorov–Smirnov test). The pattern of distinctly mapped reads (Fig. 2D, short bars linked with blue dashed lines) indicates a ~4.6-Kb deletion of a TE in the L* genome. However, this region is fully covered by mapped L* reads, indicating that the L* genome has this TE, which is on chromosome 2, as supported by the two clusters of reads (marked by pink lines in Fig. 2D) adjacent to the ends of this TE.

To investigate the extent of these SVs, we analyzed the published data (32) and identified 161 sequences (Dataset S1) that mapped to different genomic positions between Col and L*, affecting >500 Kb of the genome and leading to false positive SNP calls that relied on misplaced short reads from L* to nonallelic Col positions. More than 14% of large GCs (2 Kb ~10 Kb) and approximately 5% for shorter GCs (20 bp ~2 Kb) predicted by Yang et al. (32) were associated with this type of SVs (Table 1).

### Table 1. Reinterpretation of GCs by Yang et al. from sequencing data for two F2 plants (C94 and C95) and listed for various sizes

<table>
<thead>
<tr>
<th>Factors</th>
<th>2 Kb ~10 Kb</th>
<th>20 bp ~2 Kb</th>
<th>2 bp ~20 bp</th>
<th>1 bp</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transpositions, %</td>
<td>14.10</td>
<td>2.91</td>
<td>0.15</td>
<td>—</td>
</tr>
<tr>
<td>Copy number variants, %</td>
<td>30.77</td>
<td>12.56</td>
<td>0.30</td>
<td>—</td>
</tr>
<tr>
<td>Other type of SVs, %</td>
<td>7.69</td>
<td>1.79</td>
<td>0.45</td>
<td>—</td>
</tr>
<tr>
<td>Misplacement of reads, %</td>
<td>8.97</td>
<td>16.14</td>
<td>4.69</td>
<td>2.64</td>
</tr>
<tr>
<td>HDRs, †</td>
<td>19.23</td>
<td>6.95</td>
<td>1.06</td>
<td>—</td>
</tr>
<tr>
<td>Failure of gap-opening, %</td>
<td>2.56</td>
<td>32.96</td>
<td>87.44</td>
<td>32.90</td>
</tr>
<tr>
<td>Incorrect SNPs, %</td>
<td>12.82</td>
<td>19.73</td>
<td>3.78</td>
<td>62.46</td>
</tr>
<tr>
<td>Correct SNPs but no GCs, ‡</td>
<td>3.85</td>
<td>3.59</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Other factors, %</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Sum, %</td>
<td>100</td>
<td>100</td>
<td>100</td>
<td>99.98</td>
</tr>
<tr>
<td>Total reported GCs</td>
<td>78</td>
<td>446</td>
<td>661</td>
<td>9,924</td>
</tr>
</tbody>
</table>

*To distinguish from the type based on CNVs, this phrase refers to wrongly placing of a few reads, usually insufficient to contribute an extra coverage.

†HDRs refer to highly divergent regions with insufficient identities between two ecotypes resulting in low read coverage of L.*

‡SNPs predicted by Yang et al. are either consistent with those from 1001 Genomes (35, 36), or supported by resequencing reads of Col/L*. However, reads from the corresponding F2 plants do not support GCs in these SNP loci.

*Refers to false positive “GCs” due to incorrect prediction of CO borders, or the absence of reads mapped to the regions of GCs from the corresponding F2 plant.
Incomplete Coverage of Tandem Repeats Contributes to Genotyping Errors. Tandem repeats of DNA sequences, including mono-, di-, tri-, or longer oligonucleotides, are common and may form structures (such as loops and hairpins) (48, 49) that cause DNA polymerase slippage, resulting in frequent copy number changes (small indels). In short-read NGS data, many reads may not span an entire repetitive array and are thus unable to detect small indels. At nucleotides 1,553,129–1,553,140 on chromosome 1 of the Col genome, there is an array of six “TA” dinucleotides (Fig. 4A), but the corresponding array in Ler contains only five. There are two types of reads for this region: (1) The 3' end of the reads terminates within the tandem repeats, uninformative regarding the indel; the reads span the array and reveal the indel. The difference in the coverage of the repeat region was interpreted as evidence for two alleles, resulting in a heterozygosity call, even when only one allele was present. These heterozygosity calls were interpreted as GCs if their flanking regions were either homozygous Col or homozygous Ler.

Therefore, only reads that span a repeat array and include some flanking sequences can be reliably used to detect GCs in repetitive DNA. To estimate the number of reported short GCs (2 bp ~20 bp) (32) associated tandem repeats, we scanned the Col genome by using “Tandem Repeats Finder” (50) and found 948,905 of tandem repeat regions (8% of the genome). Strikingly, more than 87% of short GCs (2 bp ~20 bp) from two F2 plants (32) were associated with small indels in tandem repeats, although the vast majority (91%) of the genome is covered by neither tandem repeats nor indels. Furthermore, 96% of these GCs were reported as having converted from “Col- or Ler- genotypes” to “heterozygous genotypes.” As noted above, a conversion of homozygous to heterozygous genotypes is consistent with the misinterpretation of reads of differing lengths as different alleles. Reanalysis of reads mapping to the regions of 87% of the reported GCs (2 bp ~20 bp) indicated no support for heterozygosity or GC if reads that did not span the array were removed (Table 1).
Therefore, the vast majority of the short GCs are not supported by the data when using only reads that covered the entire repeat regions and provided unambiguous genotypic evidence.

Sequencing Errors Contribute to Artificial Alleles. NGS technologies have error rates of \(10^{-3}\) substitutions per nucleotide or higher (51). Sequence errors can be mistaken as evidence for polymorphism. At a given SNP site, 100x sequencing coverage will yield a 3% \((10^{-3} \times 1/3 \times 10^3)\) chance of observing a read with SNP-like changes due to error. When genome-wide SNPs are examined, many such false SNP calls are expected. To evaluate the effect of sequencing errors on GC prediction, we examined the distribution of the ratio of Col and Ler reads in regions designated as Col, Ler or heterozygous, using the SNP information for chromosome 1 and the read data from an F2 plant (C94) (32). As shown in Fig. 4B, the first 21.2 Mb of the chromosome was genotyped as Ler; as expected, the ratio of Col/(Col+Ler) reads was close to zero (Fig. 4C). Similarly, the average ratio was \(0.5\) for the next 5.5 Mb heterozygous region and \(\sim 1.0\) for the last 3.6 Mb region genotyped as Col (Fig. 4C). We then performed the same analysis on SNPs from sites on chromosome 1 that were reported to have GC of 1 bp from Ler to heterozygous (32). Strikingly, the ratio of Col/(Col+Ler) reads at these SNPs was close to \(0\%\) (Fig. 4D), indicating that the number of reads called as Col was extremely small and the true genotype for these sites was likely Ler, instead of heterozygous due to GC. Fig. 4E shows an examination of SNPs from the reported “Ler to Col” type of GCs and suggests that these SNPs do not provide support for conversions. Further analysis of all 1-bp GCs from the C94 and C95 F2 plants (32) revealed that \(62\%\) lacked sufficient read support for converted genotypes (Table 1). Because the sequencing error rate of NGS is relatively low, reads of a specific SNP allele due to error are rare compared with reads for the correct genotypes. Therefore, evaluation of observed “genotype ratio” followed by a statistical test can greatly reduce false GC calls due to sequencing errors.

Reanalysis of Reported Data for Potential GCs. We describe a sequence analysis pipeline for detection of GCs by integrating the filters described above (SI Materials and Methods and Fig. S1). Briefly, polymorphisms between Col and Ler ecotypes including SNPs, small indels, and large SVs were either collected from 1001 Genomes (36, 37) or predicted based on alignment of Col and Ler resequencing reads (32) on the reference genome (TAIR10) (35) by using BWA (52) and inGAP-sv (14). Short reads at C94 and C95 (32) were also mapped by using the same filtering strategy and uniquely mapped reads were genotyped according to polymorphic information. COs were identified by genotyping loci along each chromosome to provide “allelic background” for the identification of GCs. Sequencing depths, allelic ratios (SI Materials and Methods) and large-scale allelic information between adjacent COs were evaluated for all three genomes (Col, Ler, and the F2 plant) for each converted SNP/indel. Candidate GCs were regarded as having insufficient support if they overlapped with SVs or CNVs.

From the data of the two reported F2 plants (32), we identified 11 COs in each of C94 and C95 (diploids resulted from one male and one female meiotic events). Consistent with prior studies, each chromosome had an average of one CO per meiosis (31, 34). Because information on other meiotic products of the same meiosis was not available, potential GC associated with these COs could not be identified. Nevertheless, after apply filtering steps to data from Col, Ler, and F2 plants (SI Materials and Methods), six potential GCs (associated with NCOs) were predicted, five in C94, and one in C95 (Table S1 and Fig. S2). Among them, only one was corresponded to a small indel, consistent with the fact that there is an order of magnitude fewer small indels than SNPs in Arabidopsis (31, 33). Directions of GCs were either from “homozygous” to “heterozygous” (Fig. S3) or vice versa (Fig. S4), consistent with the allelic background of the chromosomal region. For example, all three GCs on chromosome 4 of plant C94 were from heterozygous to homozygous (one for Col/Col and two for Ler/Ler), in a background of 93% of the chromosome being Col/Ler. Two of the six GCs predicted in this study were also identified by Yang et al. (32). Optimally, predicted GCs would be validated by using PCR and conventional sequencing, but in this case, the relevant plant material was not available. The small number of GCs detected here is consistent with previous findings (31, 34, 53) and suggests relatively small sizes of the gaps repaired by NCOs, although an underestimation of GCs due to the stringent criteria here cannot be ruled out.

Variations in the Human Genome and Potential Effects on SNP Calling. To examine the effects of SVs on SNP calling in a nonplant genome by using short reads, we examined the human genome using human chromosome 1 (hg19/GRCh37) (54) as an example. It has 432,854 repeat regions (45.7% of the chromosome), including SINE (37.4%), LINE (28.2%), and other repeats. We compared the HG00656 dataset from the 1000 Genomes Project (33) (\(\sim 5x\) coverage) with the human reference genome (hg19/GRCh37) (54).

As illustrated in Fig. 4F, the human genome also has small indels associated with tandem repeats, with potential problems using short reads at low coverage. Among 58,735 tandem repeats and low-complexity regions, 3,120 have small indels and were covered by reads without gap opening (see an example in Fig. 4F), making it possible for these indels to be interpreted as SNPs when coverage is not high or without proper statistical analysis. In addition, a study (33) of large deletions included 54 deletions in HG00656 (see one example in Fig. S5 A–C), 21 of which contain SNPs compared with the reference and would be considered as homozygous when they are, in fact, hemizygous. The use of these SNPs without consideration of the deletions would affect the outcome of genetic mapping, because the breakpoints of the deletions would be considered recombination points (Fig. S5C). Analysis of these deletions suggested that 10 of 54 deletions corresponded to sequences on chromosome 1 of the HG00656 dataset, whereas the remaining 26 were less widely distributed, potentially affecting genetic studies of the relevant populations. To investigate the influence of nonallelic similar sequences, such as those related to TE s (as illustrated in Fig. 2), we examined the HG00656 dataset by using inGAP-sv to identify complex SVs. On chromosome 1 of the HG00656 dataset, we identified 38 SVs after filtering out low quality ones: 24 of these SVs corresponded to sequences on chromosome 1 of HG00656 but in the “decoy genome” (named “hs37d5”) of the reference (54). The decoy genome contains 4,715 contigs totaling 35 Mb, including viral sequences, unassembled genomic segments, or de novo assembled sequences from other human genome projects. Thus, SVs uncovered here could be genome variations or reflect incomplete assembly of the reference. These 24 segments ranged in size from 1 to 7.4 Kb, covering 61 Kb of the decoy genome and included 133 nucleotide differences, which would be misidentified as “SNPs” when SVs are not considered. We also identified 82 duplications in HG00656, mostly tandem repeats within introns or intergenic regions. The mapping of two or more such nonallelic similar sequences to the same site would result in false “heterozygosity.” Our analyses indicate that human genomes contain a large number of variations that can potentially affect erroneous SNP calling if not accounted for properly.

Conclusions Whole genome resequencing is now feasible for a variety of studies, many of which involve the analysis of sequence variants as genetic markers. It is important to correctly identify nonallelic sequence variants to avoid mistaking them as alleles. When the genomes being analyzed have indels, CNVs, and other types of SVs in comparison with the reference genome, short reads of
nonallelic sequence can originate from a different location in the resequenced genome and be misinterpreted as polymorphisms. If such false SNPs are included, frequency measurements will be unreliable. False SNPs can be minimized by using PE reads to reveal SVs between the newly sequenced and reference genomes. In addition, reanalysis of the parental genomes of genetic crosses can uncover slightly divergent duplicates and avoid calling the variant duplicates within an individual as alleles between individuals. In outcrossing species in which individuals are heterozygous for most alleles, this analysis should reveal more than two kinds of reads for sequences with two or more similar copies, thus highlighting the need to distinguish nonallelic variants from the allelic ones.

Our reanalysis of the recently reported resequencing data for Arabidopsis meiotic recombination provides strong evidence that most of the reported GCs can be explained by the presence of highly similar but nonallelic DNA segments in the Ler genome (nonreference and unassembled) and the redistribution of such nonallelic sequence by meiotic COs or independent assortment. In addition, restricting the analysis of short GCs to strictly unambiguous genotypes drastically reduced the GC number. Therefore, there is compelling evidence that GCs are a less frequent outcome of meiotic recombination in Arabidopsis, consistent with the findings of relatively few GCs per meiosis using independent methods (31, 34, 35).

Materials and Methods

Published short read sequences from Arabidopsis (32) and human (33) were analyzed by using reported methods to detect SVs associated with TEs, CNVs, short indels relating to tandem repeats, and likely sequencing errors. The SVs and other polymorphisms in Arabidopsis were then matched with the position of reported GCs in the study (32) for evaluation. In addition, two F2 genomes were generated on each polymorphic locus to identify meiotic recombination events including COS and GCs. See details in SI Materials and Methods.

ACKNOWLEDGMENTS. This research was supported by Ministry of Sciences and Technology of China 973 Program Grant 2012CB910503 (to J.Q.), National Natural Science Foundation of China Grant 91131007 (to H.M.), US National Science Foundation Grant MCB-1121563 (to G.P.), and the biological supercomputing server of Computing Center of Beijing Institute of Life Science.
Supporting Information

Qi et al. 10.1073/pnas.1321897111

SI Materials and Methods

Analysis of Resequencing Datasets. The Arabidopsis thaliana (Col ecotype) genome sequences and corresponding annotations were downloaded from The Arabidopsis Information Resource (TAIR) website (Release TAIR10) (1). The TAIR10 release differs from TAIR9 only in updated gene annotations. Resequencing datasets of Columbia (Col), Landsberg erecta (Ler), and two F2 plants (C94 and C95) were produced by Yang et al. (2) by using 2 × 100-bp paired-end sequencing technology (insert size of 500 bp).

Identification of Meiotic Recombination Events. Because of the complex nature of Arabidopsis genomes and structural variation among ecotypes, genomic polymorphisms between Col and Ler genomes must be carefully examined to exclude artificial callings before identification of meiotic recombination events in progeny genomes. Here, we used a three-step strategy (Fig. S1) to describe the prediction processes in detail as below.

Collection of polymorphisms including SNPs, small indels, and large SVs. Single-nucleotide polymorphisms (SNPs) between Col and Ler ecotypes were downloaded from 1001 Genomes (3, 4) (available at http://1001genomes.org/projects/assemblies.html) and primary validated by using resequencing reads of the two ecotypes. Short reads were aligned against the Col reference genome by using short read aligner BWA (5), and those with mapping quality scores ≥ 20 were considered uniquely mapped and were used in subsequent analyses. A qualified SNP must be supported by sufficient coverage of Col or Ler specific reads (90% of total mapped reads or higher, minimum 10 reads) in the homozygous genotypes, otherwise it will be considered as a false SNP and will be screened out. In highly divergent regions between the ecotypes, when few reads could be mapped on Ler genome, SNPs are densely crowded and sometimes adjacent to or within indels or other types of SVs. These SNPs were undoubtedly filtered out in subsequent analyses. Besides collecting SNPs from the 1001 Genomes Project (3, 4), we further applied inGAP (6) on the mapping results of paired-end Ler reads against TAIR10 reference genome to predict small indels (1–20 bp) and other SNPs not listed by 1001 Genomes (3, 4). These SNPs and indels were also examined by the procedure described above. Furthermore, Tandem Repeats Finder (7) was used with default parameters to scan the reference genome for tandem repetition of nucleotides with a minimum alignment score of 10 and maximum period size of 20. Indels overlapping tandemly repeated regions were further examined for gain/loss of tandem units between ecotypes. In such loci, reads that failed to span the entire tandem repeats were ignored for indel evaluation.

The inGAP-sv program (8), which identifies structural variants based on information of paired-end read mapping, split read mapping, and depth of coverage, was applied to the filtered mapping results of Ler reads to identify large-scale insertions, deletions, inversions, transpositions, and copy number variants. Although the Col assembly was based on long-read sequencing of BAC clones, it is possible that two or more copies of a segment in the Col genome might have been reported only once in the assembly and cause reads (either from Col or Ler) to “pile up” in one region. To avoid false prediction of SVs, Col reads were mapped to the Col reference genome and those regions (bin size of 200 bp) were excluded if they had both abnormally mapped reads and excessive sequencing coverage (with at least 50% greater read depth than both average values and that of flanking regions, additional details were described in ref. 8).

Primary genotyping for progeny genomes and prediction of COs. Detailed analyses of crossovers (COs) were described (9). Briefly, resequencing reads of two F2 plants, C94 and C95, from Yang et al. (2) were mapped to Col reference genome (TAIR10) by using the same filtering strategy as that for parental sequences. Uniquely mapped reads were genotyped when they overlapped with one or more SNP/indel loci. For reads containing indels of tandem units, only those that fully span tandem arrays were eligible for SNP calling. The polymorphic loci were recognized as Col, Ler, or heterozygous after summing up the genotypic information of the corresponding reads. Eventually COs were identified as the allelic information of polymorphic loci for a whole chromosome was gathered. The CO boundaries (adjacent to double Holliday junctions if have polymorphisms) were defined by the closest detected markers to maximize flanking regions with continuous and consistent genotypes.

Primary prediction of GCs and further examinations. Comparing with the prediction of COs that used allelic information from chromosome-scale polymorphic markers, identification of GCs were much more challenging because they changed genotypes on limited loci. Because not all artificial SNPs/indels were excluded from collections, many false positive GCs could be predicted when hundreds of thousand of markers were analyzed simultaneously. Therefore, mapping details of both parental and progeny reads must be examined carefully on polymorphic loci related to GC candidates. Here, we present a brief description on the basic procedures of the prediction and inspection of GC events.

First, sequencing depth and read distribution were inspected for Col, Ler, and F2 plants. Converted SNPs/indels were ignored if they had less read coverage than the lower quartiles (bottom 5% of all SNPs, possibly due to insufficient amplification for sequencing in high or low guanine-cytosine content regions, or unable to map reads lacking sequence similarity with reference in highly divergent regions), or more than the higher quartiles (top 5% of all SNPs, possible due to copy number variance of DNA segments).

Second, we calculated allelic ratio, defined as the proportion of Col-allellic reads to the total of Col- and Ler-allellic reads, for each polymorphic locus by using resequencing reads of parental genomes, because Col or Ler loci are not necessarily covered by Col- or Ler-allellic reads only (due to sequencing errors or wrong mapping of short reads). Distributions of allelic ratios were obtained for both Col and Ler genomes, and polymorphic loci were ruled out if not confirmed as homozygous confidently (threshold with 95%). Evaluation of allelic ratios were more complicated when considering reads from F2 plants: Allelic ratios were calculated for Col, Ler, or heterozygous regions respectively inferred from CO predictions to investigate consistency of genotypes among loci. In the analysis of reads of C94, ~99% of loci with Col/Col alleles were covered by 100% Col-genotypic reads, 96% of loci with Ler/Ler alleles by 100% Ler-genotypic reads, and 88% of loci with Col/Ler alleles had ratios ranging from 30 to 70%. Allelic ratios of SNPs/indels within GC candidates were examined with the same confidence threshold as that for Col or Ler genomes.

Finally, all predicted GCs candidates, were examined manually to exclude artifacts due to misplacement of short reads caused by SVs, especially by historic transpositions and CNVs. Those candidates passed these filters need further experimental verifications.
Analysis of SVs in Human Genomes. Resequencing data (≃5× depth) of a human genome, HG00656, were reported by Mills et al. (10) with predicted SNPs and large indels. This dataset consists of 91-bp PE reads sequenced on an Illumina platform with a mean insert size of 470 bp. The human reference genome (hg19/GRCh37) was used for comparison. The distribution of 54 large deletions found in HG00656 was also examined in each of 14 human population groups (totaling 1,092 individuals) (10) by using R, and further clustered by using “gplots” with default parameters. Complex SVs in the human genome were investigated according to the mapping results of PE reads by using inGAP-sv (8), with the sample procedure as in the analyses of *Arabidopsis* genomes of meiotic progeny as described above.

• Quality control of read sequencing
• Quality control of read mapping
• Checking of tandem repeats

Comparison with reference genome

Calling of genomic polymorphisms

SNPs  Small indels  Structural Variants

Primary genotyping for progenies on each SNP/indel

Prediction of cross-over events during meiosis

Manual examination of boundaries of crossovers

Primary prediction of gene conversions

Examination of genotyping on SNPs/indels

Paternal genome  Maternal genome  Progeny genomes

Filtering "GCs" by expected sequencing depths and allele ratios

Filtering "GCs" if overlapped with genomic structural variants

Manual examination of predicted GCs and further experimental verifications

Fig. S1. The pipeline of investigating potential GCs, with details described in SI Materials and Methods. (A) The workflow for calling of genomic polymorphisms including SNPs, indels, and SVs. (B) Prediction of COs for each meiotic progeny by primary genotyping. (C) Prediction of potential GCs and illustration of further examinations.
Fig. S2. Display of six potential GCs discovered in this analysis on C94 (A) and C95 (B). Col, Ler, and heterozygous genotypes are marked in blue, red, and stripes, respectively. The six GCs are pointed out at the corresponding positions with converted directions.

Fig. S3. A potential GC discovered in this analysis on the F2 plant C94. (A) PE read mapping from Col, Ler, and F2 plants adjacent to 7,133,180 bp on chromosome 3. Reads allelic to Col and Ler were colored in blue and red, respectively. (B) Detailed alignments of reads from Col and Ler genomes confirmed the SNP at 7133180 bp, on which reads from the F2 genome were consistent with either Col or Ler.
Fig. S4. A potential GC discovered in this analysis on the F$_2$ plant C95. (A) PE read mapping from Col, Ler, and F$_2$ plants adjacent to 7434533 bp on chromosome 1. (B) Detailed alignments of reads from Col and Ler genomes confirmed two SNPs at 7434521 bp and 7434533 bp. Some F$_2$ reads were consistent with Col at both two SNPs, whereas the other reads (colored in green in A) were identified as Ler allelic at the first SNP and as Col allelic at the second SNP.
Fig. S5. Effects of deletions in human genome on allele ratio estimation. (A) Compared with the reference genome, the DNA sequence is retained for one chromosome (the first allele) but lost in the homolog with a deletion (the second allele). (B) PE reads from the second allele of resequenced genome mapped to regions flanking the deletion, appearing to be abnormally distant, whereas SNPs from the first allele could be detected and would be considered as “homozygous” if the deletion is not recognized. (C) PE reads mapping and genotyping results within and adjacent to a 2.8-kb deletion in chromosome 1 of the human genome HG00656 (1). (D) Detection of heterozygous SNPs based on reads from two alleles flanking the deletion and of homozygous SNPs on reads from only one allele within the deletion. (E) Detection frequency of the 54 deletions in each of the 14 population groups (totaling 1,092 individuals). Label names are consistent with those in ref. 1.


Table S1. A list of potential gene conversions discovered in this analysis on two F\textsubscript{2} plants, C94 and C95

<table>
<thead>
<tr>
<th>Sample</th>
<th>Chr</th>
<th>Site</th>
<th>Col</th>
<th>Lerp</th>
<th>Type</th>
<th>to Col</th>
<th>to Lerp</th>
<th>Ratio, %</th>
<th>Potential GC direction</th>
<th>Appearance in the GC list by Yang et al. (1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C94</td>
<td>3</td>
<td>3545989</td>
<td></td>
<td></td>
<td>A</td>
<td>73</td>
<td>0</td>
<td>100</td>
<td>Heterozygous to Col</td>
<td>No</td>
</tr>
<tr>
<td>C94</td>
<td>3</td>
<td>7133180</td>
<td>T</td>
<td>C</td>
<td>SNP</td>
<td>31</td>
<td>27</td>
<td>53</td>
<td>Ler to heterozygous</td>
<td>No</td>
</tr>
<tr>
<td>C94</td>
<td>4</td>
<td>8986595</td>
<td>C</td>
<td>T</td>
<td>SNP</td>
<td>0</td>
<td>56</td>
<td>0</td>
<td>Heterozygous to Ler</td>
<td>No</td>
</tr>
<tr>
<td>C94</td>
<td>4</td>
<td>12358751</td>
<td>T</td>
<td>C</td>
<td>SNP</td>
<td>85</td>
<td>0</td>
<td>100</td>
<td>Heterozygous to Col</td>
<td>Yes</td>
</tr>
<tr>
<td>C94</td>
<td>4</td>
<td>13651179</td>
<td>A</td>
<td>T</td>
<td>SNP</td>
<td>0</td>
<td>60</td>
<td>0</td>
<td>Heterozygous to Ler</td>
<td>Yes</td>
</tr>
<tr>
<td>C95</td>
<td>1</td>
<td>7434533</td>
<td>C</td>
<td>T</td>
<td>SNP</td>
<td>60</td>
<td>0</td>
<td>100</td>
<td>Heterozygous to Col</td>
<td>No</td>
</tr>
</tbody>
</table>


Other Supporting Information Files

Dataset S1 (XLSX)
Dataset S2 (XLSX)
Dataset S3 (XLSX)