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]

[a]State Key Laboratory of Genetic Engineering and Collaborative Innovation Center for Genetics and Development, Institute of Plant Biology, Center for Evolutionary Biology, School of Life Sciences, and [b]Ministry of Education Key Laboratory for Biodiversity Science and Ecological Engineering, Institute of Biodiversity Sciences, Fudan University, Shanghai 200433, China; [c]Department of Biology and the Carolina Center for Genome Sciences, University of North Carolina, Chapel Hill, NC 27599-3280; [d]Lineberger Comprehensive Cancer Center, University of North Carolina School of Medicine, Chapel Hill, NC 27599-3280; and [e]Institutes of Biomedical Sciences, Fudan University, Shanghai 200032, China

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 by 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 accurate 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.

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. 1 *A* 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. 1*C*), 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.

Results and Discussion

In many species, large-scale SVs often involve identical or highly similar sequences that differ in chromosomal contexts between

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.

*This Direct Submission article had a prearranged editor.

[1]To 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.

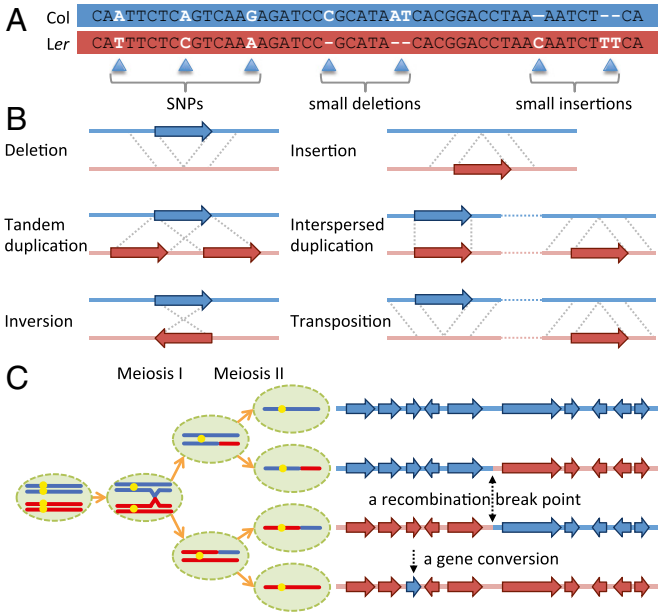

Fig. 1. (*A*) SNPs and small indels between two ecotype genomes. (*B*) Possible types of SVs. Col genotypes are marked in blue and L*er* in red. Arrows indicate DNA segments involved in SVs between the two ecotypes. (*C*) Meiotic recombination events including a CO and a GC (NCO). Centromeres are denoted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

individuals. Numerous genomic polymorphisms have been reported between two *Arabidopsis* ecotypes (geographic variants), Columbia (Col, TAIR10 assembly; ref. 35) and Landsberg *erecta* (L*er*), 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 L*er* that include TEs (Figs. 1*B* and 2) create regions of high sequence similarity that map to different (nonallelic) 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 nonreference ecotype (L*er*) 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 if they are on the same chromosome (Fig. 2 *A* and *B*). Mistaking these nonallelic sequences as SNPs can lead to the misidentification of gene conversion events (presumably accompanying NCOs), and recent studies have used some of the strategies described here to avoid such mistakes (31, 34).

Specifically, when nonallelic homologous Col and L*er* sequences are redistributed by COs or independent assortment, the failure to recognize such nonallelic sequences can result in false GC events (from heterozygous to Col) in the progeny that lack L*er* sequences. For example, if a DNA segment in Col (Fig. 2*C*, blue arrow) is nearly identical to a nonallelic sequence in L*er* (Fig. 2*C*, red arrow), a CO between the two loci (or an independent assortment between two chromosomes if the segments are unlinked) results in an F_2 plant with the second locus having the Col genotype, without the L*er* allele. Consequently, short-read NGS data of the F_2 plant will yield only the Col genotype at the first locus, which is different from its heterozygous flanking genotype (Col/L*er*) (Fig. 2*C*). In this example, if the nonallelic homologous sequences are not recognized, it is not possible to distinguish between a true GC event and a CO between SVs, resulting in an overestimation of GCs.

Recently, Yang et al. reported that more than 1,000 GC events occurred in each progeny of *Arabidopsis* meiosis (93,696 GCs for 40 progeny of products of both male and female meiosis) (32). This result is surprising because each GC should originate from a SPO11-mediated DSB, regardless of whether the GC is associated with a CO or a NCO. Although DSBs have not been directly measured in *Arabidopsis*, several groups have used immunolocalization of recombination proteins such as RAD51 and DMC1 to provide indirect estimates of DSBs ranging from 120

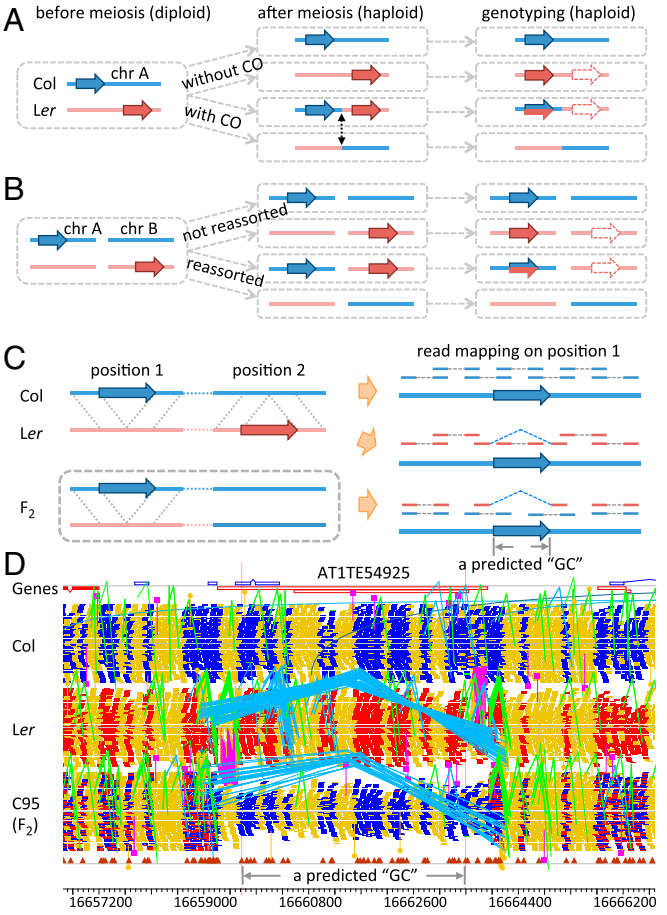

Fig. 2. Paralogous sequences between two ecotypes and their effects on allele ratio estimation. Redistribution of SV-related paralogous DNA segments in meiotic progenies and consequent genotyping using short read mapping for paralogs on the same chromosome (*A*), or paralogs on different chromsomes (*B*). Col and L*er* 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, L*er*, and the F_2 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 L*er* read pairs are mapped normally, except those L*er* read pairs flanking a deletion, shown here as distantly mapped red reads with blue linkers. Only one chromatid is shown for both Col and L*er*. An F_2 plant has a heterozygous (Col/L*er*) genotype around position 1 and homozygous (Col/Col) around position 2. (*D*) PE read mapping from Col, L*er*, and an F_2 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 L*er* reads indicate a deletion: A TE is surrounded by a group of reads pairs (linked by blue lines) that mapped farther 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 (coconversion of consecutive polymorphisms) occurred at the same locations in multiple meioses (32). Our reanalysis of the sequencing data of two parental and two F_2 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*er*) that has deletions compared with the reference genome (e.g., Col) is resequenced by using PE sequencing, the regions flanking the deleted DNA are adjacent 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. 2*C*), providing strong support for a deletion in L*er*. Furthermore, the deleted L*er* sequence might be found at a different (paralogous) genomic location, possibly resulting from historic transpositions. These nonallelic L*er* sequences are mapped back to the Col reference, contributing to false SNP calls in subsequent GC analysis. When an F_2 plant lacks a paralogous L*er* copy because of CO or chromosome reassortment, the heterozygous region would be misrepresented as having the Col genotype (Fig. 2*C*). The situation is illustrated by an example shown in Fig. 2*D*: We found a transposable element (TE; AT1TE54925) on chromosome 1 of Col (at nucleotide position 16,659,688–16,664,330 bp) that has a paralog on chromosome 2, but not on chromosome 1, in L*er* (at ~1.3 Mb; ref. 35). An F_2 plant, designated C95 by Yang et al., was of the Col genotype for the entire length of chromosome 2, thus lacking the L*er* copy of AT1TE54925. As a result, no reads for the L*er* 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*er*) 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*er* allowed us to identify inappropriate GC calls at this site in 13 of 40 meiotic offspring, including C95 (32).

Accurate detection of meiotic GCs by using 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*er*) relative to the assembled reference genome (Col) (Fig. 2*C*). As in the Yang et al. data, a cluster of 33 PE reads from L*er* 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 $<<10^{-3}$ using the Kolmogorov–Smirnov test). The pattern of distantly mapped reads (Fig. 2*D*, short bars linked with blue dashed lines) indicates a ~4.6-Kb deletion of a TE in the L*er* genome. However, this region is fully covered by mapped L*er* reads, indicating that the L*er* genome has this TE, which is on chromosome 2, as supported by the two clusters of reads (marked by pink lines in Fig. 2*D*) 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*er*, affecting >500 Kb of the genome and leading to false positive SNP calls that relied on misplaced short reads from L*er* to nonallelic Col positions. More than 14% of large GCs (2 Kb ~10 Kb) and approximately 3% for shorter GCs (20 bp ~2 Kb) predicted by Yang et al. (32) were associated with this type of SVs (Table 1).

Artifactual SNPs from CNVs. Approximately 20% of the *Arabidopsis* genome (35) is comprised of TE-related sequences, which can vary in copy number and position among ecotypes (42–44). In addition, movement of TEs can cause CNV of nearby sequences (45–47). CNVs of both TE and non-TE sequences can also result in the misidentification of genotypes due to the mismapping of short reads (Fig. 3 *A* and *B*). We identified 1,429 CNVs between Col and L*er* (Dataset S2), covering 2.7 Mb of the Col genome, and found 30% of large GCs (2 Kb ~10 Kb) and 12% of shorter GCs (20 bp ~2 Kb) identified by Yang et al. are located in these regions (Table 1). For example, a sequence has two copies in L*er* but only one copy in Col (nucleotides 2,246,169–2,256,074 on chromosome 2; Fig. 3*C*). The number of reads mapped to the reference Col sequence was significantly higher (P value $<<10^{-3}$) than those in its flanking regions (Fig. 3*C*) because reads from the additional L*er* copy were mapped to the single Col copy.

When the duplicated copies have diverged slightly, their sequence differences can be mistaken for SNPs, but their true paralogous nature can be definitively recognized because they

Table 1. Reinterpretation of GCs by Yang et al. from sequencing data for two F_2 plants (C94 and C95) and listed for various sizes

Factors	2 Kb ~10 Kb	20 bp ~2 Kb	2 bp ~20 bp	1 bp
Transpositions, %	14.10	2.91	0.15	—
Copy number variants, %	30.77	12.56	0.30	—
Other type of SVs, %	7.69	1.79	0.45	—
Misplacement of reads,* %	8.97	16.14	4.69	2.64
HDRs,† %	19.23	6.95	1.06	—
Failure of gap-opening, %	2.56	32.96	87.44	32.90
Incorrect SNPs,‡ %	_	3.36	2.12	1.98
Correct SNPs but no GCs,§ %	12.82	19.73	3.78	62.46
Other factors,¶ %	3.85	3.59	—	—
Sum, %	100	100	100	99.98
Total reported GCs	78	446	661	9,924

*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*er*. SNPs predicted in these regions lack enough support from sequencing data.
‡SNPs predicted by Yang et al. (32) are not supported by sufficient Col/L*er* genotypic reads.
§SNPs predicted by Yang et al. are either consistent with those from 1001 Genomes (35, 36), or supported by resequencing reads of Col/L*er*. However, reads from the corresponding F_2 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 F_2 plant.

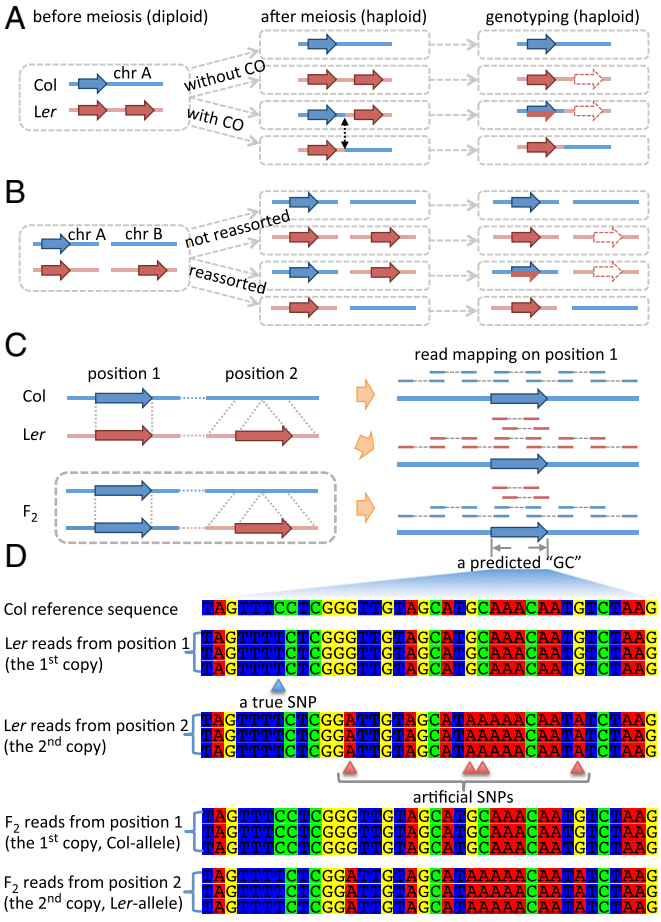

Fig. 3. PE reads mapping patterns for CNVs and corresponding genotype predictions. Genotyping of SV-related genomic regions in meiotic progenies due to misplacement of short read in the case of CNVs on the same chromosome (*A*), or CNVs on different chromosomes (*B*). (*C*) Mapping pattern of reads from Col, L*er*, and the F_2 plant on the reference genome. The F_2 plant (an example) has two Col-homologous copies and one L*er*-homologous copy. Both L*er* and the F_2 show abnormally higher read depth within the duplicated segment. (*D*) Both L*er* and the F_2 could have two or three kinds of "genotypes," depending on the sequence divergence of the three segments in *C*, where Col has one copy and L*er* has two copies. Comparison of allelic sequences between Col and L*er* give one true SNP (in blue), whereas misplacing the nonallelic copy of L*er* creates three artifactual SNPs (in red).

occur in the homozygous L*er* parent, even when there are more than two copies. As shown in Fig. 3*D*, the sequence of the true L*er* allele has only one substitution: from the Col "C" to "T." However, when reads from a nonallelic duplicate copy were also mapped to the Col site, the four SNP-like differences were used inappropriately to call for GC (32) because these sequence differences were thought to be genotypes different from their flanking regions. An examination of the read depth and mapping distance/orientation of the reported large GCs (2 Kb ~10 Kb) (32) showed that 30% are associated with loci that show evidence of copy number variation between the two parental ecotypes (Table 1). These results suggested that meiotic COs and independent assortment of chromosomes involving duplicated copies can contribute to the inflation of GC estimates from short reads.

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. 4*A*), but the corresponding array in L*er* 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 (2); 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 L*er*.

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 F_2 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 L*er*- 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).

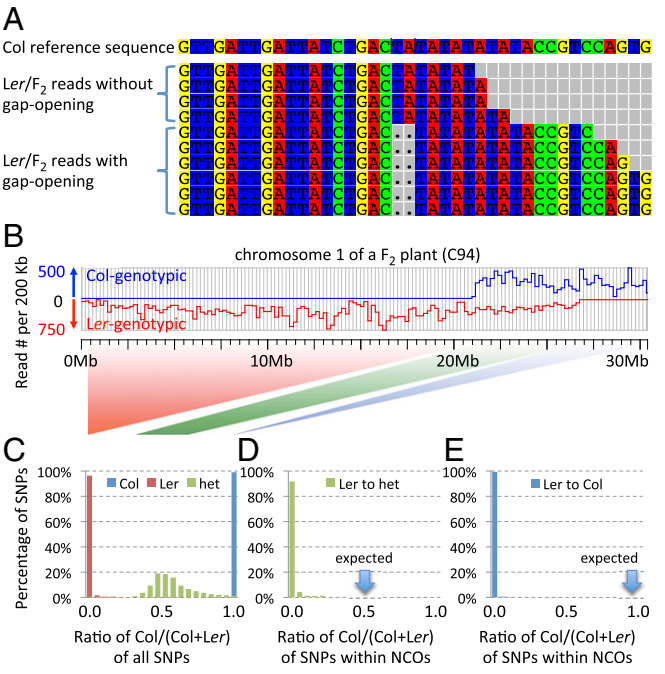

Fig. 4. (*A*) Alignment of short reads with small indels in tandem repeat arrays. Gaps are opened only for L*er*/F_2 reads that span the entire tandem array. (*B*) Genotypes of chromosome 1 of an F_2 plant (C94): Upper blue line indicates the number of reads genotyped as Col per 200-kb window, and the lower red line is the depth of reads genotyped as L*er*. (*C*) Distribution of allele ratio for SNPs from left L*er* homozygous region (background in red), middle heterozygous region (background in green), and right Col homozygous region (background in blue). (*D*). Distribution of allele ratio of SNPs within predicted GCs by Yang et al. as "heterozygous converted from L*er*." Allelic ratio of Col at these SNPs is expected to be approximately 50% (blue arrow). (*E*) Distribution of allele ratio of SNPs within predicted GCs predicted by Yang et al. as "Col converted from L*er*". The expected ratio (100% of Col-genotypes) is also represented by a blue arrow.

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 Artifactual Alleles. NGS technologies have error rates of $\sim 10^{-3}$ substitutions per nucleotide or higher (51). Sequence errors can be mistaken as evidence for polymorphism. At a given SNP site, 100× sequencing coverage will yield a 3% ($10^{-3} \times 1/3 \times 10^2$) 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 L*er* reads in regions designated as Col, L*er* or heterozygous, using the SNP information for chromosome 1 and the read data from an F_2 plant (C94) (32). As shown in Fig. 4*B*, the first 21.2 Mb of the chromosome was genotyped as L*er*; as expected, the ratio of Col/(Col+L*er*) reads was close to zero (Fig. 4*C*). Similarly, the average ratio was ∼0.5 for the next 5.5 Mb heterozygous region and ∼1.0 for the last 3.6 Mb region genotyped as Col (Fig. 4*C*). We then performed the same analysis on SNPs from sites on chromosome 1 that were reported to have GC of 1 bp from L*er* to heterozygous (32). Strikingly, the ratio of Col/(Col+L*er*) reads at these SNPs was close to 0% (Fig. 4*D*), indicating that the number of reads called as Col was extremely small and the true genotype at these sites was likely L*er*, instead of heterozygous due to GC. Fig. 4*E* shows an examination of SNPs from the reported "L*er* 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 F_2 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 GCs 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 L*er* ecotypes including SNPs, small indels, and large SVs were either collected from 1001 Genomes (36, 37) or predicted based on alignment of Col and L*er* resequencing reads (32) on the reference genome (TAIR10) (35) by using BWA (52) and inGAP-sv (14). Short reads of 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, L*er*, and the F_2 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 F_2 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, L*er*, and F_2 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 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 L*er*/L*er*), in a background of 93% of the chromosome being Col/L*er*. 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) (∼5× coverage) with the human reference genome (hg19/GRCh37) (54).

As illustrated in Fig. 4*A*, 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. 4*A*), 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. S5*D*). An analysis of the distribution of the 54 deletions among 1,092 individuals from 14 populations (Dataset S3 and Fig. S5*E*) revealed that 16 of these deletions were detected in each population, indicative of ancient variations. Twelve other deletions were found frequently in American/East Asian/European populations, 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 TEs (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 sequences 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 stringently 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, 53).

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 F_2 genomes were genotyped 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.C.), and the biological supercomputing server of Computing Center of Beijing Institutes of Life Science.

1. Feuk L, Carson AR, Scherer SW (2006) Structural variation in the human genome. *Nat Rev Genet* 7(2):85–97.
2. Sharp AJ, Cheng Z, Eichler EE (2006) Structural variation of the human genome. *Annu Rev Genomics Hum Genet* 7:407–442.
3. Mitchell-Olds T, Schmitt J (2006) Genetic mechanisms and evolutionary significance of natural variation in Arabidopsis. *Nature* 441(7096):947–952.
4. Stankiewicz P, Lupski JR (2002) Genome architecture, rearrangements and genomic disorders. *Trends Genet* 18(2):74–82.
5. Hurles ME, Dermitzakis ET, Tyler-Smith C (2008) The functional impact of structural variation in humans. *Trends Genet* 24(5):238–245.
6. Stankiewicz P, Lupski JR (2010) Structural variation in the human genome and its role in disease. *Annu Rev Med* 61:437–455.
7. Johanson U, et al. (2000) Molecular analysis of FRIGIDA, a major determinant of natural variation in Arabidopsis flowering time. *Science* 290(5490):344–347.
8. Michaels SD, He Y, Scortecci KC, Amasino RM (2003) Attenuation of FLOWERING LOCUS C activity as a mechanism for the evolution of summer-annual flowering behavior in Arabidopsis. *Proc Natl Acad Sci USA* 100(17):10102–10107.
9. Koornneef M, Alonso-Blanco C, Vreugdenhil D (2004) Naturally occurring genetic variation in Arabidopsis thaliana. *Annu Rev Plant Biol* 55:141–172.
10. Krieger U, Lippman ZB, Zamir D (2010) The flowering gene SINGLE FLOWER TRUSS drives heterosis for yield in tomato. *Nat Genet* 42(5):459–463.
11. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 18(11):1851–1858.
12. Brockman W, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res* 18(5):763–770.
13. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: Short oligonucleotide alignment program. *Bioinformatics* 24(5):713–714.
14. Qi J, Zhao F, Buboltz A, Schuster SC (2010) inGAP: An integrated next-generation genome analysis pipeline. *Bioinformatics* 26(1):127–129.
15. Qi J, Zhao F (2011) inGAP-sv: A novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res* 39(Web Server issue):W567–575.
16. Chen K, et al. (2009) BreakDancer: An algorithm for high-resolution mapping of genomic structural variation. *Nat Methods* 6(9):677–681.
17. Korbel JO, et al. (2009) PEMer: A computational framework with simulation-based error models for inferring genomic structural variants from massive paired-end sequencing data. *Genome Biol* 10(2):R23.
18. Lee S, Hormozdiari F, Alkan C, Brudno M (2009) MoDIL: Detecting small indels from clone-end sequencing with mixtures of distributions. *Nat Methods* 6(7):473–474.
19. Hormozdiari F, Alkan C, Eichler EE, Sahinalp SC (2009) Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res* 19(7):1270–1278.
20. Hajirasouliha I, et al. (2010) Detection and characterization of novel sequence insertions using paired-end next-generation sequencing. *Bioinformatics* 26(10):1277–1283.
21. Zeitouni B, et al. (2010) SVDetect: A tool to identify genomic structural variations from paired-end and mate-pair sequencing data. *Bioinformatics* 26(15):1895–1896.
22. Sindi S, Helman E, Bashir A, Raphael BJ (2009) A geometric approach for classification and comparison of structural variants. *Bioinformatics* 25(12):i222–i230.
23. Ye K, Schulz MH, Long Q, Apweiler R, Ning Z (2009) Pindel: A pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. *Bioinformatics* 25(21):2865–2871.
24. Abel HJ, et al. (2010) SLOPE: A quick and accurate method for locating non-SNP structural variation from targeted next-generation sequence data. *Bioinformatics* 26(21):2684–2688.
25. Wong K, Keane TM, Stalker J, Adams DJ (2010) Enhanced structural variant and breakpoint detection using SVMerge by integration of multiple detection methods and local assembly. *Genome Biol* 11(12):R128.
26. Kidd JM, et al. (2008) Mapping and sequencing of structural variation from eight human genomes. *Nature* 453(7191):56–64.
27. Graubert TA, et al. (2007) A high-resolution map of segmental DNA copy number variation in the mouse genome. *PLoS Genet* 3(1):e3.
28. Emerson JJ, Cardoso-Moreira M, Borevitz JO, Long M (2008) Natural selection shapes genome-wide patterns of copy-number polymorphism in Drosophila melanogaster. *Science* 320(5883):1629–1631.
29. Huang X, et al. (2010) Genome-wide association studies of 14 agronomic traits in rice landraces. *Nat Genet* 42(11):961–967.
30. Qi J, et al. (2009) Characterization of meiotic crossovers and gene conversion by whole-genome sequencing in Saccharomyces cerevisiae. *BMC Genomics* 10:475.
31. 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 Res* 22(3):508–518.
32. Yang S, et al. (2012) Great majority of recombination events in Arabidopsis are gene conversion events. *Proc Natl Acad Sci USA* 109(51):20992–20997.
33. Mills RE, et al.; 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59–65.
34. Wijnker E, et al. (2013) The genomic landscape of meiotic crossovers and gene conversions in Arabidopsis thaliana. *eLife* 2:e01426.
35. Lamesch P, et al. (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210.
36. Schneeberger K, et al. (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc Natl Acad Sci USA* 108(25):10249–10254.
37. Cao J, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* 43(10):956–963.
38. Long Q, et al. (2013) Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nat Genet* 45(8):884–890.
39. Mercier R, et al. (2005) Two meiotic crossover classes cohabit in Arabidopsis: One is dependent on MER3,whereas the other one is not. *Curr Biol* 15(8):692–701.
40. Chelysheva L, et al. (2005) AtREC8 and AtSCC3 are essential to the monopolar orientation of the kinetochores during meiosis. *J Cell Sci* 118(Pt 20):4621–4632.
41. 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.
42. Tsukahara S, et al. (2009) Bursts of retrotransposition reproduced in Arabidopsis. *Nature* 461(7262):423–426.
43. Initiative TAG; Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature* 408(6814):796–815.
44. Le QH, Wright S, Yu Z, Bureau T (2000) Transposon diversity in Arabidopsis thaliana. *Proc Natl Acad Sci USA* 97(13):7376–7381.
45. Jiang N, Bao Z, Zhang X, Eddy SR, Wessler SR (2004) Pack-MULE transposable elements mediate gene evolution in plants. *Nature* 431(7008):569–573.
46. Juretic N, Hoen DR, Huynh ML, Harrison PM, Bureau TE (2005) The evolutionary fate of MULE-mediated duplications of host gene fragments in rice. *Genome Res* 15(9):1292–1297.
47. Morgante M, et al. (2005) Gene duplication and exon shuffling by helitron-like transposons generate intraspecies diversity in maize. *Nat Genet* 37(9):997–1002.
48. Djian P (1998) Evolution of simple repeats in DNA and their relation to human disease. *Cell* 94(2):155–160.
49. Petruska J, Hartenstine MJ, Goodman MF (1998) Analysis of strand slippage in DNA polymerase expansions of CAG/CTG triplet repeats associated with neurodegenerative disease. *J Biol Chem* 273(9):5204–5210.
50. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573–580.
51. Loman NJ, et al. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol* 30(5):434–439.
52. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
53. Sun Y, et al. (2012) Deep genome-wide measurement of meiotic gene conversion using tetrad analysis in Arabidopsis thaliana. *PLoS Genet* 8(10):e1002968.
54. Meyer LR, et al. (2013) The UCSC Genome Browser database: Extensions and updates 2013. *Nucleic Acids Res* 41(Database issue):D64–D69.

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* (L*er*), and two F_2 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 L*er* genomes must be carefully examined to exclude artifactual 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 L*er* 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 L*er* 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 L*er* 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 L*er* 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 L*er* 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 L*er*) 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 F_2 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, L*er*, 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, L*er*, and F_2 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-allelic reads to the total of Col- and L*er*-allelic reads, for each polymorphic locus by using resequencing reads of parental genomes, because Col or L*er* loci are not necessarily covered by Col- or L*er*-allelic reads only (due to sequencing errors or wrong mapping of short reads). Distributions of allelic ratios were obtained for both Col and L*er* 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 F_2 plants: Allelic ratios were calculated for Col, L*er*, 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 L*er*/L*er* alleles by 100% L*er*-genotypic reads, and 88% of loci with Col/L*er* 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 L*er* 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.

1. Lamesch P, et al. (2012) The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210.
2. Yang S, et al. (2012) Great majority of recombination events in Arabidopsis are gene conversion events. *Proc Natl Acad Sci USA* 109(51):20992–20997.
3. Schneeberger K, et al. (2011) Reference-guided assembly of four diverse Arabidopsis thaliana genomes. *Proc Natl Acad Sci USA* 108(25):10249–10254.
4. Cao J, et al. (2011) Whole-genome sequencing of multiple Arabidopsis thaliana populations. *Nat Genet* 43(10):956–963.
5. Li H, Durbin R (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 25(14):1754–1760.
6. Qi J, Zhao F, Buboltz A, Schuster SC (2010) inGAP: An integrated next-generation genome analysis pipeline. *Bioinformatics* 26(1):127–129.
7. Benson G (1999) Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res* 27(2):573–580.
8. Qi J, Zhao F (2011) inGAP-sv: A novel scheme to identify and visualize structural variation from paired end mapping data. *Nucleic Acids Res* 39(Web Server issue): W567–575.
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 Res* 22(3):508–518.
10. Mills RE, et al.; 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59–65.

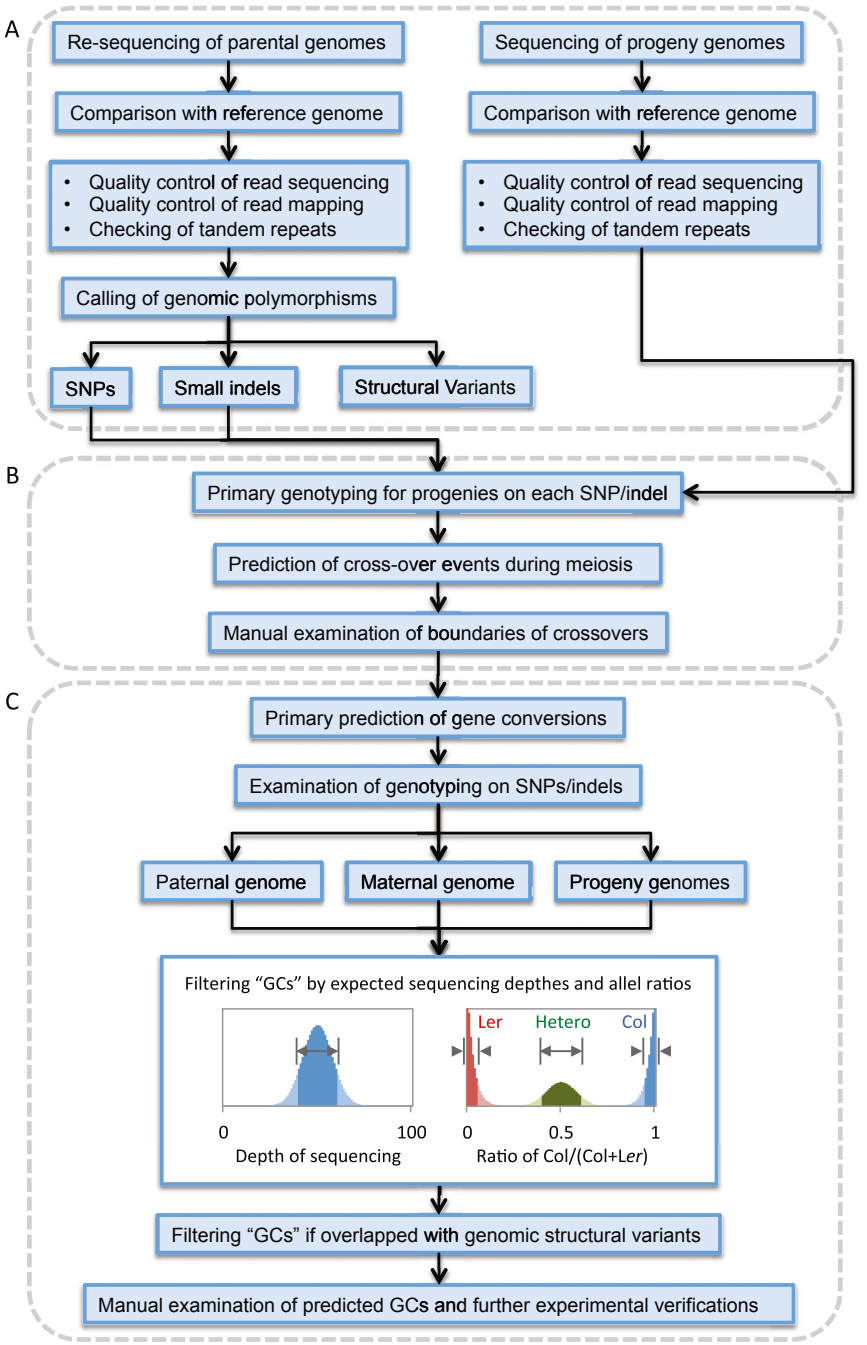

Fig. S1. The pipeline of investigating potential GCs, with details described in *SI Materials and Methods*. (*A*) The workflow for calling of genomic polymorphisms including SNPs, indels, and SVs. (*B*) Prediction of COs for each meiotic progenies by primary genotyping. (*C*) Prediction of potential GCs and illustration of further examinations.

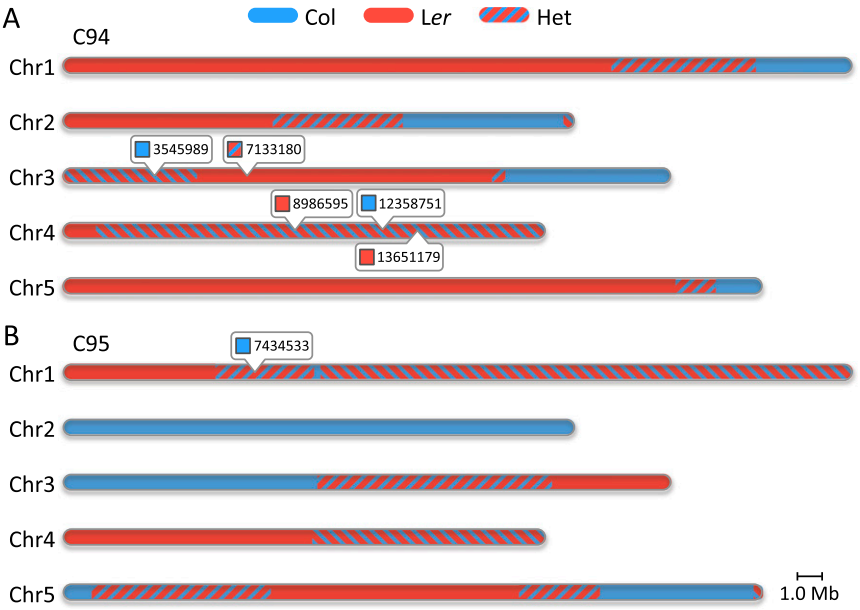

Fig. S2. Display of six potential GCs discovered in this analysis on C94 (*A*) and C95 (*B*). Col, L*er*, and heterozygous genotypes are marked in blue, red, and stripes, respectively. The six GCs are pointed out at the corresponding positions with converted directions.

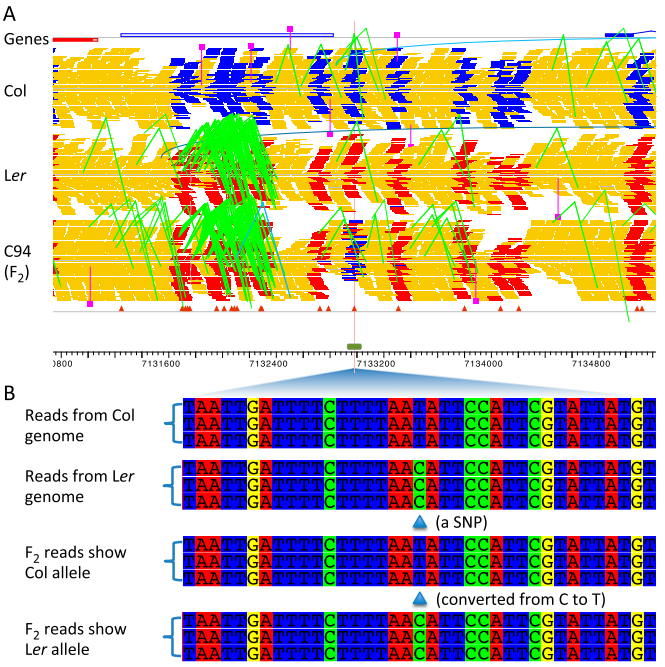

Fig. S3. A potential GC discovered in this analysis on the F_2 plant C94. (*A*) PE read mapping from Col, L*er*, and F_2 plants adjacent to 7,133,180 bp on chromosome 3. Reads allelic to Col and L*er* were colored in blue and red, respectively. (*B*) Detailed alignments of reads from Col and L*er* genomes confirmed the SNP at 7133180 bp, on which reads from the F_2 genome were consistent with either Col or L*er*.

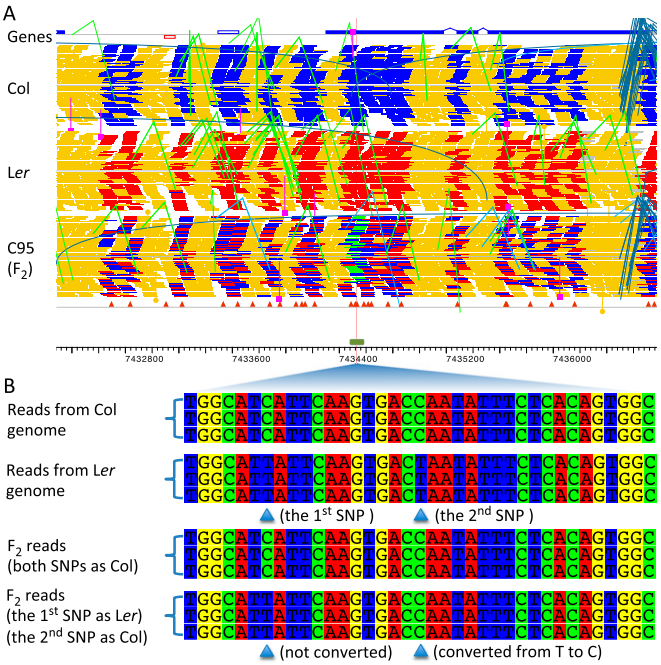

Fig. S4. A potential GC discovered in this analysis on the F_2 plant C95. (*A*) PE read mapping from Col, L*er*, and F_2 plants adjacent to 7434533 bp on chromosome 1. (*B*) Detailed alignments of reads from Col and L*er* genomes confirmed two SNPs at 7434521 bp and 7434533 bp. Some F_2 reads were consistent with Col at both two SNPs, whereas the other reads (colored in green in *A*) were identified as L*er* allelic at the first SNP and as Col allelic at the second SNP.

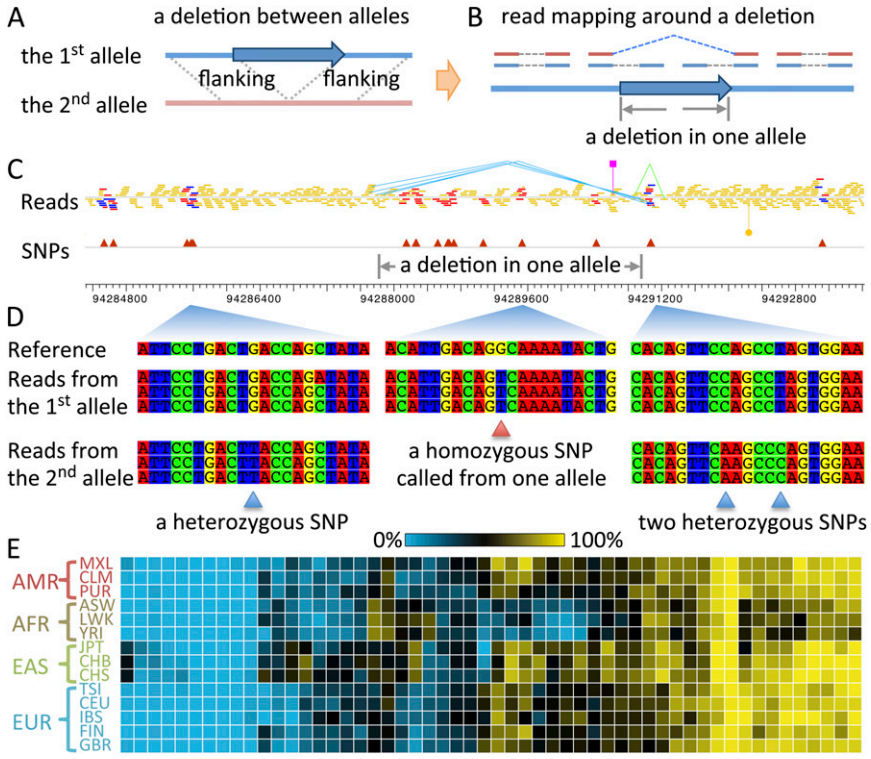

Fig. S5. Effects of deletions in human genome on allele ratio estimation. (*A*) Compared with the reference genome, the DNA sequence is retained for one chromosome (the first allele) but lost in the homolog with a deletion (the second allele). (*B*) PE reads from the second allele of resequenced genome 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.

1. Mills RE, et al.; 1000 Genomes Project (2011) Mapping copy number variation by population-scale genome sequencing. *Nature* 470(7332):59–65.

Table S1. A list of potential gene conversions discovered in this analysis on two F_2 plants, C94 and C95

						No. of read in F_2 allelic to parental genomes				
Sample	Chr	Site	Col	L*er*	Type	to Col	to L*er*	Ratio, %	Potential GC direction	Appearance in the GC list by Yang et al. (1)
C94	3	3545989	—	A	INDEL	73	0	100	Heterozygous to Col	No
C94	3	7133180	T	C	SNP	31	27	53	L*er* to heterozygous	No
C94	4	8986595	C	T	SNP	0	56	0	Heterozygous to L*er*	No
C94	4	12358751	T	C	SNP	85	0	100	Heterozygous to Col	Yes
C94	4	13651179	A	T	SNP	0	60	0	Heterozygous to L*er*	Yes
C95	1	7434533	C	T	SNP	60	0	100	Heterozygous to Col	No

1. Yang S, et al. (2012) Great majority of recombination events in Arabidopsis are gene conversion events. *Proc Natl Acad Sci USA* 109(51):20992–20997.

Other Supporting Information Files

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