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.

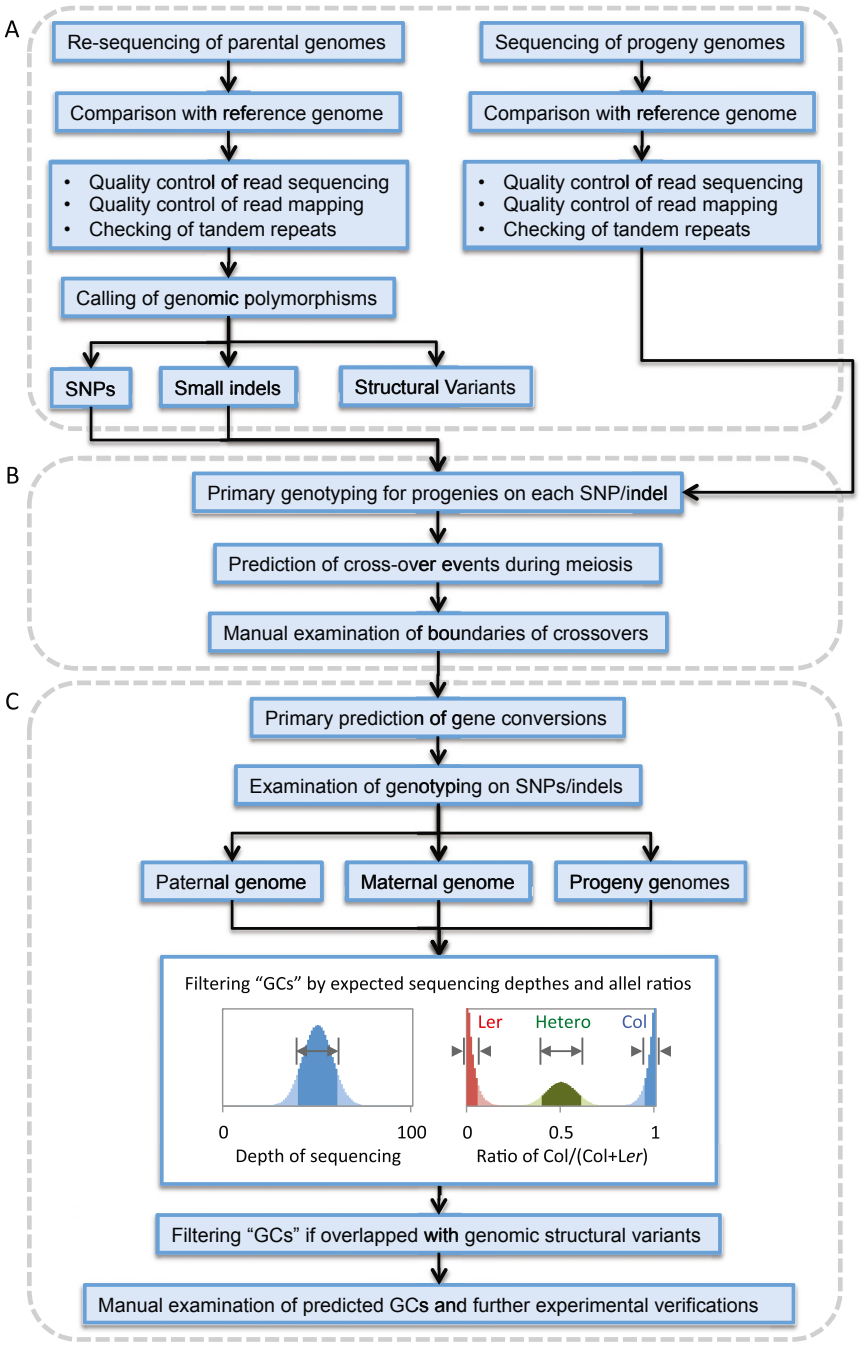

Fig. S1. The pipeline of investigating potential GCs, with details described in *SI Materials and Methods*. (*A*) The workflow for calling of genomic polymorphisms including SNPs, indels, and SVs. (*B*) Prediction of COs for each meiotic progenies by primary genotyping. (*C*) Prediction of potential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

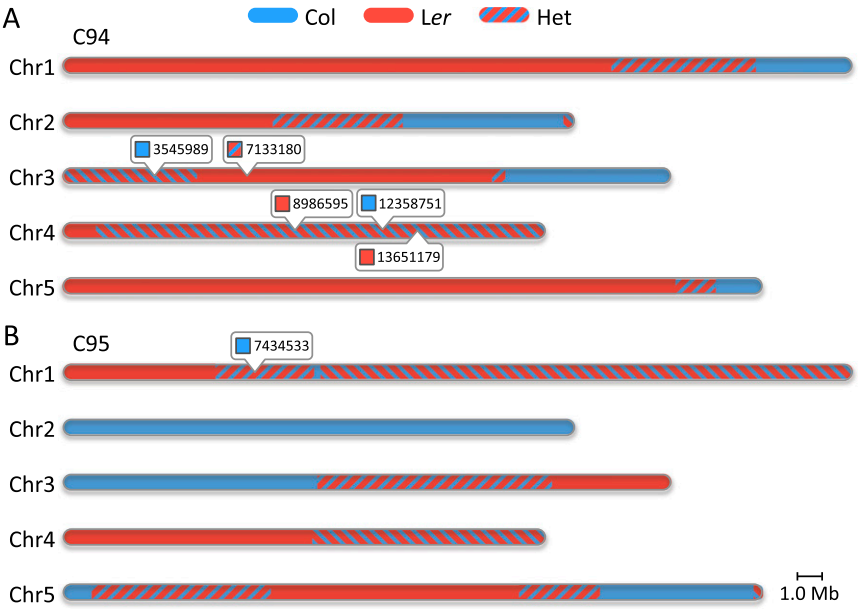

Fig. S2. Display of six potential GCs discovered in this analysis on C94 (*A*) and C95 (*B*). Col, L*er*, and heterozygous genotypes are marked in blue, red, and stripes, respectively. The six GCs are pointed out at the corresponding positions with converted directions.

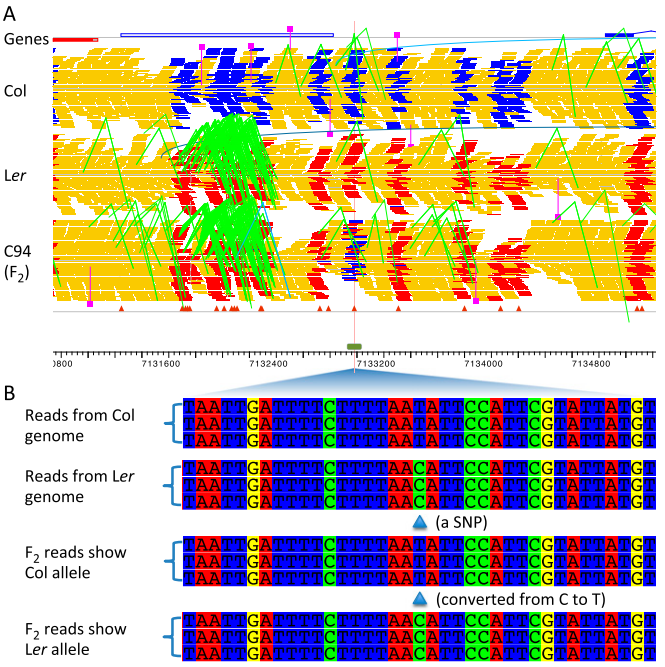

Fig. S3. A potential GC discovered in this analysis on the F_2 plant C94. (*A*) PE read mapping from Col, L*er*, and F_2 plants adjacent to 7,133,180 bp on chromosome 3. Reads allelic to Col and L*er* were colored in blue and red, respectively. (*B*) Detailed alignments of reads from Col and L*er* genomes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

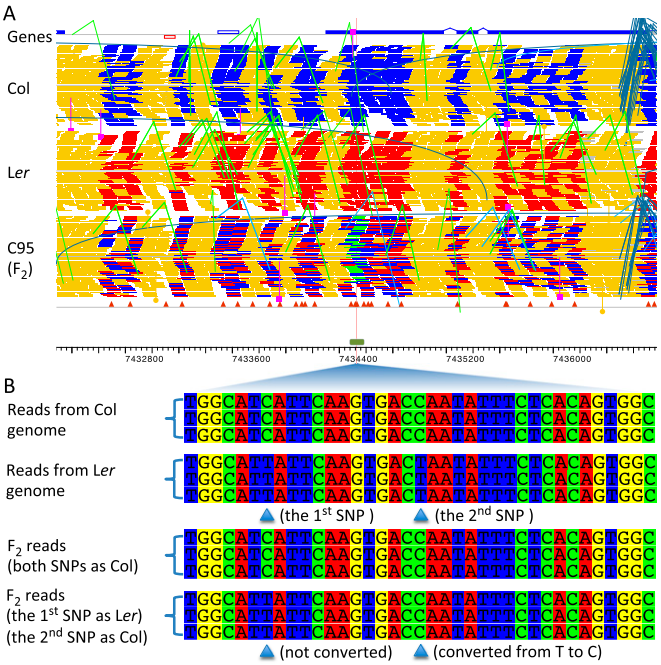

Fig. S4. A potential GC discovered in this analysis on the F_2 plant C95. (*A*) PE read mapping from Col, L*er*, and F_2 plants adjacent to 7434533 bp on chromosome 1. (*B*) Detailed alignments of reads from Col and L*er* genomes confirmed two SNPs at 7434521 bp and 7434533 bp. Some F_2 reads were consistent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

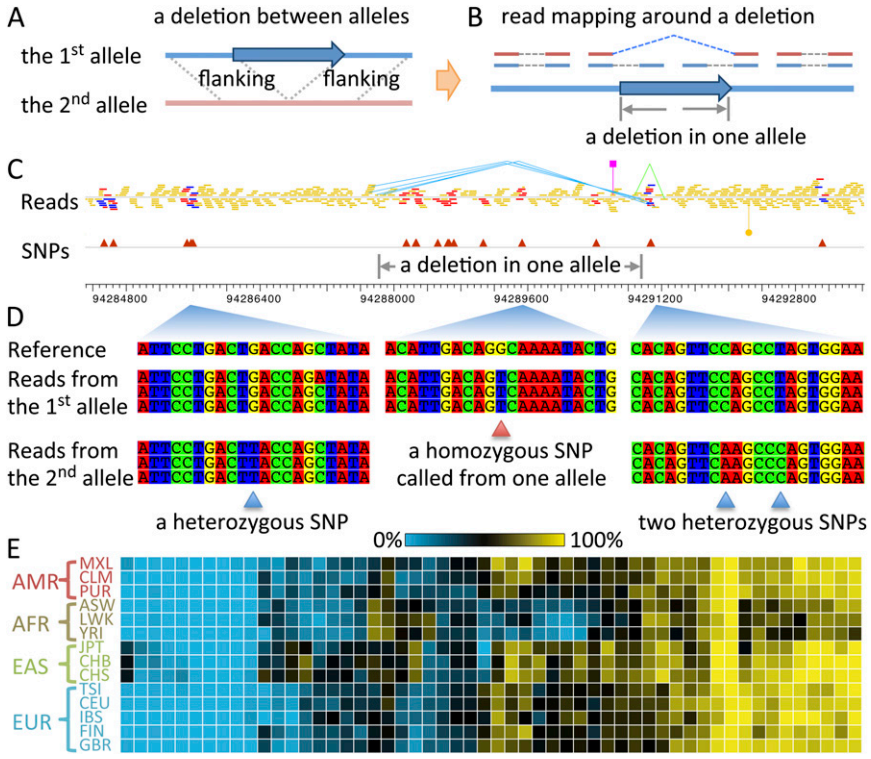

Fig. S5. Effects of deletions in human genome on allele ratio estimation. (*A*) Compared with the reference genome, the DNA sequence is retained for one chromosome (the first allele) but lost in the homolog with a deletion (the second allele). (*B*) PE reads from the second allele of resequenced genome 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	*Ler*	Type	to Col	to *Ler*	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	*Ler* to heterozygous	No
C94	4	8986595	C	T	SNP	0	56	0	Heterozygous to *Ler*	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 *Ler*	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)