Supporting Information

Xie et al. 10.1073/pnas.1005931107

SI Results

Several factors may affect the accuracy of maximum parsimony of recombination (MPR) inference: the number of single-nucleotide polymorphisms (SNPs) processed each time (window size), the density of putative SNPs (the distance between SNP sites), the number of recombinant inbred lines (RILs), and the maximum step size of the heuristic perturbation. We conducted Monte Carlo simulations to assess the effects of these factors using the 11,948 high-quality SNPs that were identified in both multiplexed sequencing of RILs and parental deep sequencing. The results of 10,000 random samplings are shown in Figs. S3 and S4. As expected, when the window size was fixed at 50 SNPs, the number of recombination events in a window increases with the length of the genomic region. This is the case for using data from both deep sequencing and MPR-inferred genotypes of the parents (Fig. S3 *A* and *B*). When SNP density is greater than 105 kb/SNP and the maximum step size is greater than 1, more than 99.9% of the deduced parental genotype calls are identical to the sequencing results (Fig. S3*C*), and more than 95% of the total number of recombination events calculated using the sequencing results are equal to the predictions (Fig. S3*D*). The results also showed that consistency would drop quickly when SNP density is lower than 232 kb/SNP, providing a reference for the applicability and the limitation of the MPR method. However, even when the SNP density is very low, >93% of the parental genotype calls could be deduced correctly, which indicates that the MPR method is robust even for species with lower SNP densities or even with lower sequence coverage.

The results showed that influence of window size on inference at high SNP densities is small, and the accuracy of deduced parental genotypes stabilizes when there are 30 SNPs or more per window (Fig. S4*A*). When SNP density is low, a larger window size leads to lower accuracy. Moreover, a population size of 110 RILs suffices to achieve an accuracy of 99% with SNP density of 49 kb/SNP or higher (Fig. S4*B*). More lines would be needed with lower SNP density to achieve the same accuracy.

We also evaluated the influence of heterozygosity on parental genotyping by simulations. A subset (denoted as N) of the 238 RILs was randomly selected and divided into two groups. Individuals in the two groups were mated pairwise in silico, resulting in $N/2$ F_2-like lines, similar to the immortalized F_2 design (1, 2). Such a mating process was repeated twice, resulting in N F_2-like lines. Combined with the unmated lines ($238 - N$), we obtained a mixed population of 238 lines. A series of different Ns was selected, resulting in populations with expected heterozygosity ranging from 0 to 50%. The simulated results of MPR analyses of these populations showed that over 98% of the parental genotypes can be inferred correctly even when heterozygosity is 50% (Fig. S5), indicating the robustness of the method. The results also showed that the inference of the MPR method is more sensitive to heterozygosity when SNP density is low.

SI Materials and Methods

Identifying High-Quality SNPs Using Parental Sequences. The "pileup" text format files of the two parents were obtained separately using the same process as for RIL sequences. For each parent, a high-quality SNP was identified when satisfying the following criteria: (*i*) only one nucleotide on this position with a sum of base quality for this nucleotide ≥ 60; (*ii*) the nucleotide is supported by at least three reads; (*iii*) at least one base call of the nucleotide with base quality ≥ 20; (*iv*) ≥80% of the base calls on this position agree with the nucleotide; and (*v*) the consensus nucleotides between parents are different.

Identifying Inferior SNPs Using Parental Sequences. To eliminate inferior SNPs due to copy-number variation or incorrect alignment which produces two or more different nucleotides at a site within a parent, we processed the "pileup" text format files of the two parents separately and identified such SNPs using the criteria for identification of potential SNPs using RIL sequences (*Materials and Methods*). SNPs that had two or more different nucleotides at a site within a parent were regarded as inferior SNPs.

1. Hua JP, et al. (2002) Genetic dissection of an elite rice hybrid revealed that heterozygotes are not always advantageous for performance. *Genetics* 162:1885–1895.

2. Hua JP, et al. (2003) Single-locus heterotic effects and dominance by dominance interactions can adequately explain the genetic basis of heterosis in an elite rice hybrid. *Proc Natl Acad Sci USA* 100:2574–2579.

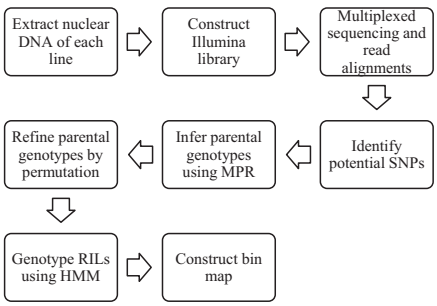

Fig. S1. The workflow of parent-independent genotyping.

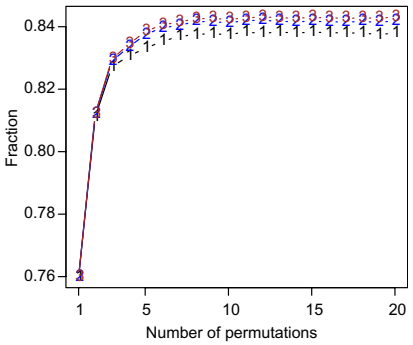

Fig. S2. The effect of the number of permutations on the fraction of SNPs remaining after removing inferior SNPs. Three lines in the figure are results produced by different assumed error rates of the MPR method in genotype inference (0.05%, 0.5%, and 10% denoted by 1, 2, and 3, respectively).

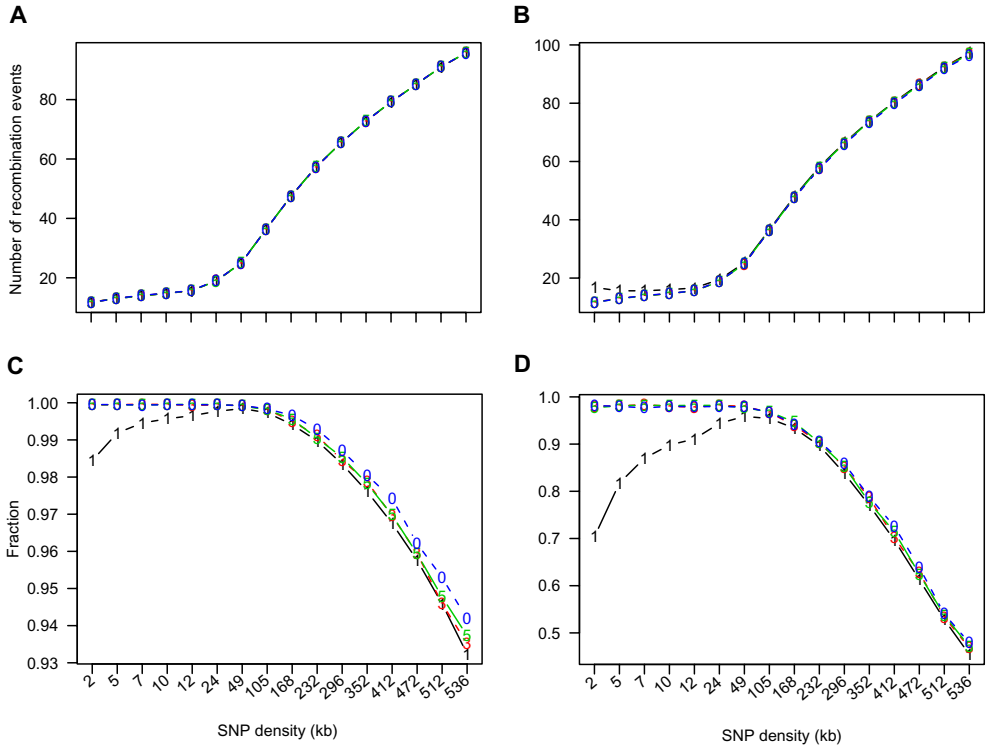

Fig. S3. Performance of the MPR method at different SNP densities (kb/SNP) and maximum step sizes of the perturbation in windows of 50 SNPs. Monte Carlo simulations were executed 10,000 times for each evaluation and the results were averaged. High-quality SNPs detected in sequencing of both 238 RILs and parents were randomly sampled to obtain different SNP densities from one SNP per 2 kb to 536 kb. The *x* axis shows the SNP densities. The *y* axis represents the number of recombination events that would occur in the RILs based on parental genotypes from sequencing results (*A*) or the MPR-inferred genotypes (*B*), or the fraction of parental genotype calls inferred correctly (*C*) or the fraction of the numbers of recombination events calculated using the sequencing results equaling the predictive results (*D*). Four lines in each figure are results of evaluations with four different maximum step sizes (1, 3, 5, and 10 denoted by 1, 3, 5, and 0, respectively).

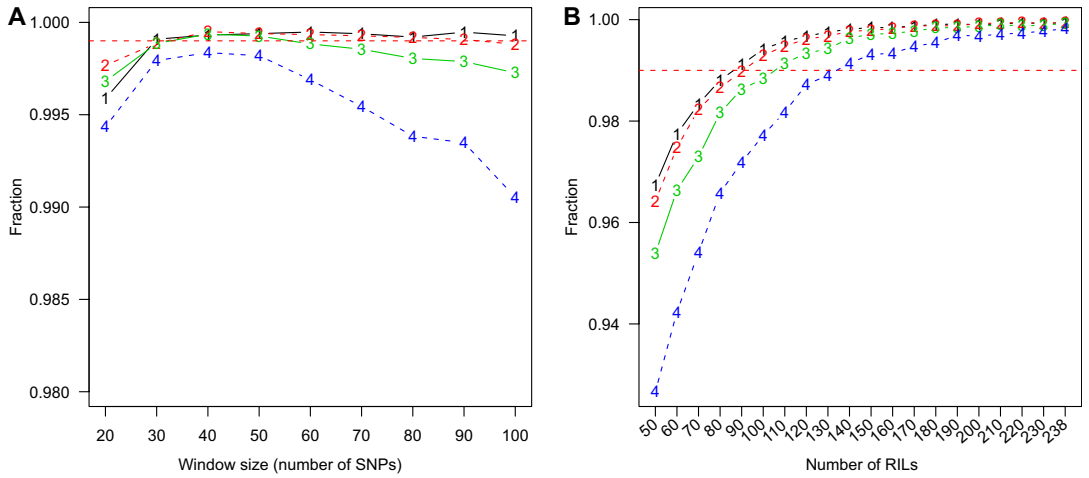

Fig. S4. Performance of the MPR method with different window sizes and numbers of RILs. Monte Carlo simulations were executed 10,000 times for each evaluation and the results were averaged. High-quality SNPs were randomly sampled to obtain four different SNP densities (2, 24, 49, and 105 kb/SNP denoted by 1, 2, 3, and 4, respectively). The *y* axis represents the fraction of parental genotype calls inferred correctly. (*A*) All RILs were used in evaluation, and the maximum step size of heuristic perturbation was set to 3. The horizontal red dashed line denotes the fraction of 0.999. (*B*) The maximum step size was set to 3 and the window size was fixed to 50 SNPs. The horizontal red dashed line denotes the fraction of 0.99.

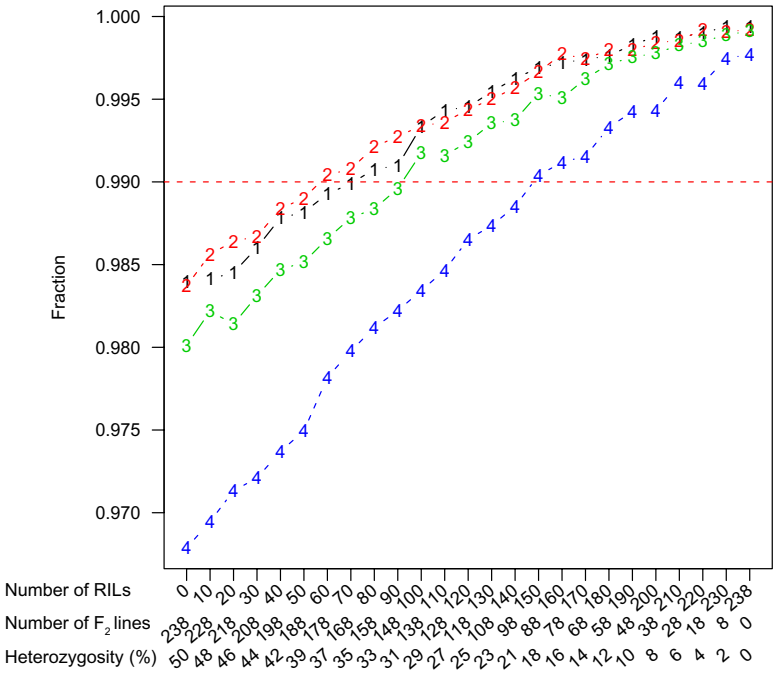

Fig. S5. Performance evaluation for the MPR method at different heterozygosities. The *y* axis shows the fraction of parental genotype calls inferred correctly. Four lines are results of evaluations with four different SNP densities (2, 24, 49, and 105 kb/SNP denoted by 1, 2, 3, and 4, respectively). The horizontal red dashed line denotes the fraction of 0.99. The maximum step size was set to 3 and the window size was fixed at 50. The results show that the accuracy of the MPR method in low SNP density is more sensitive to heterozygosity.

Other Supporting Information Files

Dataset S1 (XLS)
Dataset S2 (XLS)