# High-throughput genotyping by whole-genome resequencing

Xuehui Huang,[1,6] Qi Feng,[1,2,6] Qian Qian,[3,6] Qiang Zhao,[1,2,6] Lu Wang,[1,6] Ahong Wang,[1,6] Jianping Guan,[1] Danlin Fan,[1] Qijun Weng,[1] Tao Huang,[1] Guojun Dong,[3] Tao Sang,[1,4] and Bin Han[1,5,7]

*[1]National Center for Gene Research and Institute of Plant Physiology and Ecology, Shanghai Institutes of Biological Sciences, Chinese Academy of Sciences, Shanghai 200233, China; [2]College of Life Science and Biotechnology, Shanghai Jiaotong University, Shanghai 200240, China; [3]State Key Lab of Rice Biology, China National Rice Research Institute, Chinese Academy of Agricultural Sciences, Hangzhou 310006, China; [4]Department of Plant Biology, Michigan State University, East Lansing, Michigan 48824, USA; [5]Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100029, China*

**The next-generation sequencing technology coupled with the growing number of genome sequences opens the opportunity to redesign genotyping strategies for more effective genetic mapping and genome analysis. We have developed a high-throughput method for genotyping recombinant populations utilizing whole-genome resequencing data generated by the Illumina Genome Analyzer. A sliding window approach is designed to collectively examine genome-wide single nucleotide polymorphisms for genotype calling and recombination breakpoint determination. Using this method, we constructed a genetic map for 150 rice recombinant inbred lines with an expected genotype calling accuracy of 99.94% and a resolution of recombination breakpoints within an average of 40 kb. In comparison to the genetic map constructed with 287 PCR-based markers for the rice population, the sequencing-based method was ~20× faster in data collection and 35× more precise in recombination breakpoint determination. Using the sequencing-based genetic map, we located a quantitative trait locus of large effect on plant height in a 100-kb region containing the rice "green revolution" gene. Through computer simulation, we demonstrate that the method is robust for different types of mapping populations derived from organisms with variable quality of genome sequences and is feasible for organisms with large genome sizes and low polymorphisms. With continuous advances in sequencing technologies, this genome-based method may replace the conventional marker-based genotyping approach to provide a powerful tool for large-scale gene discovery and for addressing a wide range of biological questions.**

**[Supplemental material is available online at www.genome.org. Pseudomolecules harboring 1,226,791 SNPs identified between *Oryza sativa* ssp. *indica* cv. 9311 and ssp. *japonica* cv. Nipponbare are available at http://www.ncgr.ac.cn/english/edatabase.htm. The raw Illumina sequencing data are available in the EBI European Nucleotide Archive (ftp://ftp.era.ebi.ac.uk/) with accession number ERA000078.]**

The first use of DNA-based markers decades ago laid the groundwork for gene discovery through forward and reverse genetics. The types of markers and methods for constructing genetic maps have evolved rapidly with advances in molecular biology techniques. The development of PCR triggered the burst of a generation of markers that considerably simplified experimental procedures for marker designing and scoring. However, these markers, although still widely used, have shown growing limitations in chromosomal coverage, time, and cost effectiveness. The development of genomics concepts and tools has set the stage for replacing the marker-based mapping approach with genome-based high-throughput strategies.

The availability of genome sequences opened the door to high-throughput genotyping. This was initially accomplished by adopting microarray technology, which detects single nucleotide polymorphisms (SNPs) through hybridizing genomic DNA to oligonucleotides spotted on gene chips. This genotyping method substantially improved the efficiency of marker collection by allowing the detection of hundreds to thousands of markers in a single hybridization (Winzeler et al. 1998). It has been applied to model systems such as human, *Arabidopsis*, and rice (Meaburn et al. 2006; Singer et al. 2006; Jeremy et al. 2008). Although the goal of high-throughput was achieved, serious limitations remain for the array-based method. It is laborious, time-consuming, and expensive to design, produce, and process microarrays suited for specific mapping populations.

The advent of the next-generation sequencing technology holds the promise for a methodological leap forward in genotyping and genetic mapping. The new sequencing techniques not only increase sequencing throughput by several orders of magnitude but also allow simultaneously sequencing a large number of samples using a multiplexed sequencing strategy (Craig et al. 2008; Cronn et al. 2008). These recent technical advances have paved the way for the development of a sequencing-based high-throughput genotyping method that combines advantages of time and cost effectiveness, dense marker coverage, high mapping accuracy and resolution, and more comparable genome and genetic maps among mapping populations and organisms.

Here we describe the first high-throughput genotyping method that uses SNPs detected by whole-genome resequencing.

---

**[6]These authors contributed equally to this work.**
**[7]Corresponding author.**
**E-mail bhan@ncgr.ac.cn; fax 86-21-64825775.**

Article published online before print. Article and publication date are at http://www.genome.org/cgi/doi/10.1101/gr.089516.108. Freely available online through the *Genome Research* Open Access option.

19:1068–1076 © 2009 by Cold Spring Harbor Laboratory Press; ISSN 1088-9051/09; www.genome.org

This type of SNP data differs from traditional genetic markers primarily in two aspects. First, it is often not the case that all members of a recombinant population can be scored at a given SNP site. Second, an individual SNP site is no longer a reliable marker or locus for genotyping due to several potential sources of sequence errors. To deal with these unique features of the SNP data generated by the next-generation sequencing, we developed a new analytical framework, that is, a sliding window approach for evaluating SNPs collectively rather than individually. The method was applied to analyzing 150 rice recombinant inbred lines (RILs) derived from a cross between *indica* and *japonica* rice cultivars using sequences generated on the Illumina Genome Analyzer (GA).

## Results

### Experimental design

As shown in Figure 1, RILs were developed from a cross between rice cultivars, *Oryza sativa* ssp. *indica* cv. 93-11 and ssp. *japonica* cv. Nipponbare, whose genome sequences were previously reported (International Rice Genome Sequencing Project 2005; Yu et al. 2005). SNPs between the two genome sequences were identified as potential markers for genotyping. A total of 150 RILs were sequenced using the bar-coded multiplexed sequencing strategy. Indexed DNA samples of 16 RILs were combined and sequenced in a lane of the Illumina GA. The 33-mer sequences of each RIL were sorted according to the indexes and aligned with the parental genome sequences for SNP detection. Detected SNPs were arranged according to their physical positions, with genotypes specified. Consecutive SNPs were examined in a sliding window of 15 SNPs in size. The ratio between the numbers of SNPs from the two parents was calculated in each window and used for genotype calling. As the window slid along the chromosome, recombination breakpoints were determined. Recombination maps of the RILs with a high density of SNPs and precisely defined recombination breakpoints were constructed.

### Sequencing and SNP identification

The genomes of the RILs were resequenced on the Illumina GA. The utility of three-base indexes for multiplexed sequencing allowed us to combine 16 RILs into one lane of the sequencer, sequence 112 samples in seven lanes for a sequencing run (the eighth lane was used for the control), and complete sequencing of 150 RILs in two runs. For each RIL, the reads of 33-bp sequences (33-mers excluding the index) were sorted according to the 5′ indexes. Given the throughput of 1 Gb per sequencing run, ~7.2 Mb (1000 Mb/128) × 33/36 sequences were generated for each RIL, equivalent to ~0.02× coverage of the rice genome.

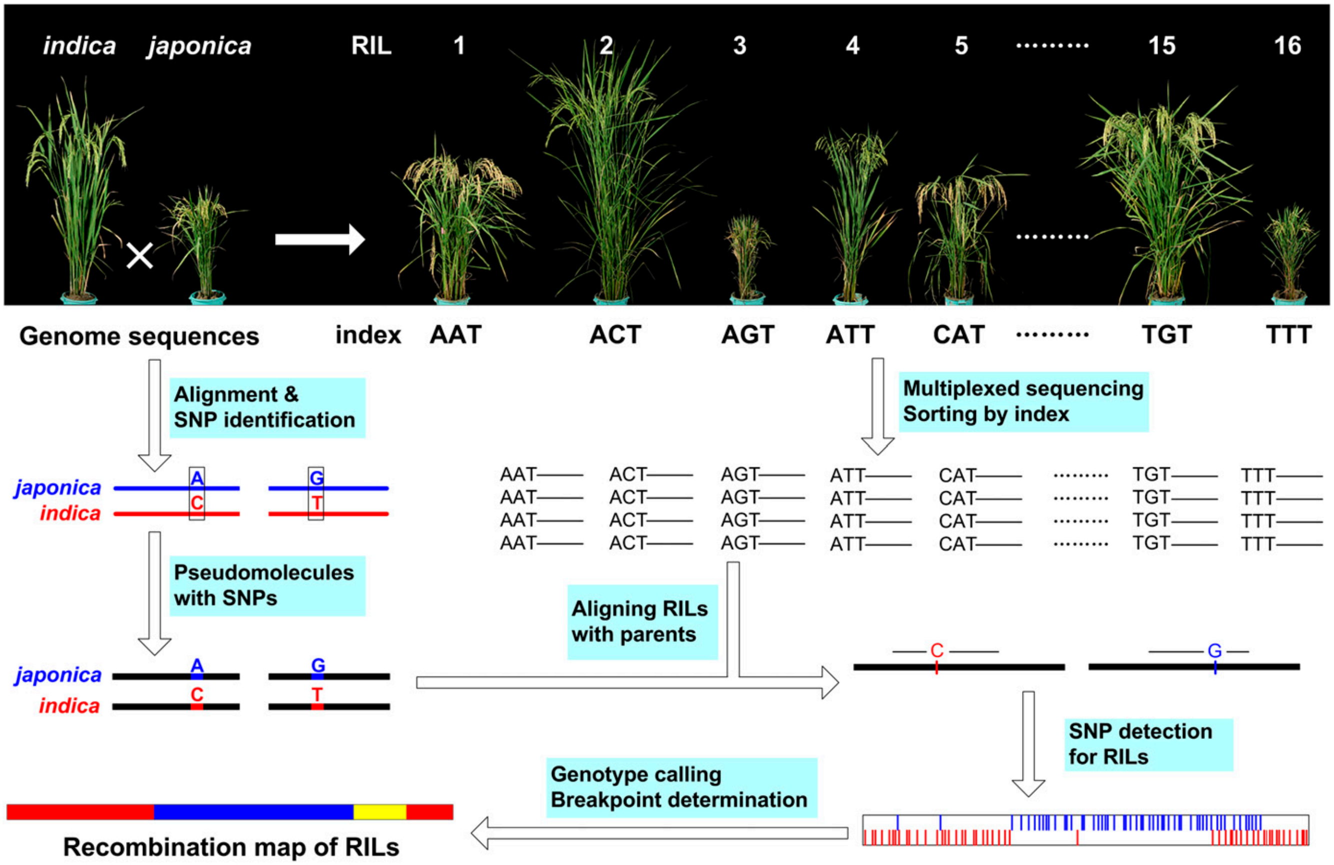

indica
japonica
RIL
1
2
3
4
5
15
16
Genome sequences
index
AAT
ACT
AGT
ATT
CAT
TGT
TTT
Alignment & SNP identification
Multiplexed sequencing Sorting by index
japonica
indica
A
C
G
T
Pseudomolecules with SNPs
Aligning RILs with parents
SNP detection for RILs
Genotype calling Breakpoint determination
Recombination map of RILs

**Figure 1.** Sequence-based high-throughput genotyping. Rice RILs were developed from a cross between *indica* and *japonica* cultivars. Genome sequences of the parents were aligned and SNPs were identified. Genomes of the RILs were resequenced on the Illumina Genome Analyzer using the multiplexed sequencing strategy. Three-base indexed DNAs of 16 RILs were combined and sequenced in one lane. Sequences were sorted and aligned with the pseudomolecules of parental genome sequences for SNP detection. Detected SNPs were arranged along chromosomes according to their physical locations with genotypes indicated. A sliding window approach was used for genotype calling, recombination breakpoint determination, and map construction.

The 33-mer short reads of RILs were aligned with the genome sequences of the parents. Through analysis of the most updated genome sequences of *indica* cv. 93-11 and *japonica* cv. Nipponbare, we identified 1,226,791 SNPs or 3.2 SNPs/kb between them (http://www.ncgr.ac.cn/english/edatabase.htm). When a 33-mer of a RIL was aligned to a region where an SNP was detected between the parents, the genotype of the RIL was assigned at this nucleotide position. From high-quality sequences obtained for the 150 RILs, a total of 1,493,461 SNPs were detected, which gave an average density of 25 SNPs/Mb or 1 SNP every 40 kb for the RILs.

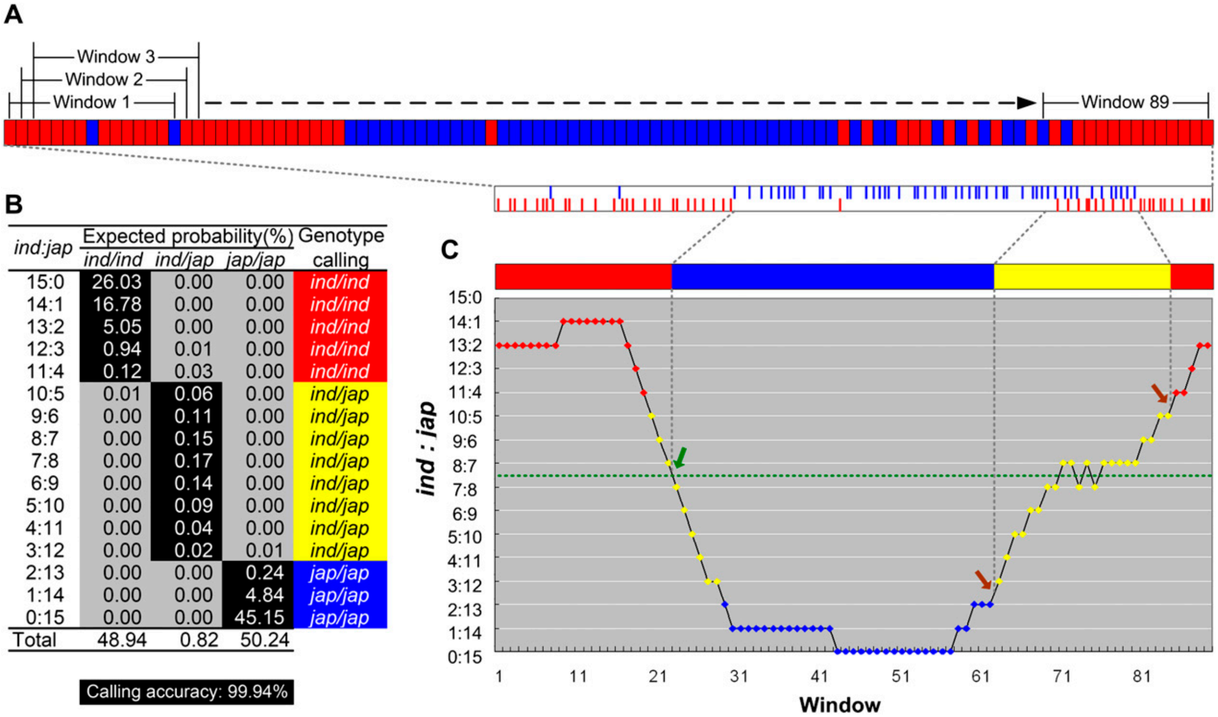
## Genotype calling

When SNPs detected from the RILs were placed along the chromosomes, we found that typically in a chromosomal region, SNPs representing one parent were predominant and those representing the other parent were scattering among them. The presence of minority SNPs was a result of sequence errors. Because of these noise SNPs, the genotype of the RILs could not be simply determined based on individual SNPs. That is, a SNP can no longer be treated as a mapping locus, as in the traditional way of using molecular markers. A sliding window approach was developed to evaluate a group of consecutive SNPs for genotyping (Fig. 2A).

We also resequenced both parents with the three-base indexes on the Illumina GA under the same experimental condition as the RILs, i.e., mixed with RILs in 16 samples per lane. After the 33-mers were aligned with the pseudomolecules of the parental genome sequences, we found that the resequencing data had SNP error rates of 4.12% and 0.71% for *indica* and *japonica* parents, respectively. This means that at a given SNP site, the chance of getting a wrong nucleotide from the short reads is 4.12% and 0.71% for *indica* and *japonica* sequences, respectively. Based on these error rates, we expected to detect 3.41% (4.12% minus 0.71%) more *japonica* SNPs than *indica* SNPs in the heterozygous regions of RILs.

Taking the SNP error rates into consideration, we could calculate the probability of occurrence of each genotype for a given SNP ratio in the sliding window (Equations 1–3, Methods). In addition to SNP errors, we need to consider the proportion of each genotype in the RIL population. We began with the theoretical expectation that the ratios of the three genotypes, *ind/ind:ind/jap:jap/jap*, were 49.98:0.05:49.98 in the $F_{11}$ generation of RILs. Using these ratios and the estimated SNP errors, we calculated the expected probabilities of the three genotypes based on SNP ratios in the window of 15 SNPs (Equations 4–6, Methods).

We made genotype calling based on the highest probability of a genotype (Equation 7, Methods) and continued this process as the window slid base-by-base along the chromosome. After this process was completed, we obtained a new estimate of genotype ratios, which were then used as the new start point to calculate the



| *ind:jap* | Expected probability(%) *ind/ind* | *ind/jap* | *jap/jap* | Genotype calling |
|---|---|---|---|---|
| 15:0 | 26.03 | 0.00 | 0.00 | *ind/ind* |
| 14:1 | 16.78 | 0.00 | 0.00 | *ind/ind* |
| 13:2 | 5.05 | 0.00 | 0.00 | *ind/ind* |
| 12:3 | 0.94 | 0.01 | 0.00 | *ind/ind* |
| 11:4 | 0.12 | 0.03 | 0.00 | *ind/ind* |
| 10:5 | 0.01 | 0.06 | 0.00 | *ind/jap* |
| 9:6 | 0.00 | 0.11 | 0.00 | *ind/jap* |
| 8:7 | 0.00 | 0.15 | 0.00 | *ind/jap* |
| 7:8 | 0.00 | 0.17 | 0.00 | *ind/jap* |
| 6:9 | 0.00 | 0.14 | 0.00 | *ind/jap* |
| 5:10 | 0.00 | 0.09 | 0.00 | *ind/jap* |
| 4:11 | 0.00 | 0.04 | 0.00 | *ind/jap* |
| 3:12 | 0.00 | 0.02 | 0.01 | *ind/jap* |
| 2:13 | 0.00 | 0.00 | 0.24 | *jap/jap* |
| 1:14 | 0.00 | 0.00 | 4.84 | *jap/jap* |
| 0:15 | 0.00 | 0.00 | 45.15 | *jap/jap* |
| Total | 48.94 | 0.82 | 50.24 | |

Calling accuracy: 99.94%

**Figure 2.** Sliding window approach for genotype calling and recombination breakpoint determination. (*A*) The *top* stripe of blocks represents SNPs along the hypothetical chromosomal region. This was redrawn from the two stripes of short vertical lines *below* illustrating SNPs detected by aligning 33-mers with the parental genome sequences. (Red) *Indica* genotype; (blue) *japonica* genotype. A sliding window covering 15 SNPs moves from *left* to *right* one base at a time. For each window, the ratio of the number of *indica* to *japonica* SNPs (*ind:jap*) is calculated. (*B*) Genotype calling based on the highest expected probabilities: Call homozygous *indica* genotype (*ind/ind*) when *ind:jap* ≥ 11:4; call heterozygous genotype (*ind/jap*) when 10:5 ≥ *ind:jap* ≥ 3:12; call homozygous *japonica* genotype (*jap/jap*) when *ind:jap* ≤ 2:13. Adding together the probabilities of these callings (shaded in black) gives the calling accuracy of 99.94%. (*C*) As the window slides, genotypes are called and recombination breakpoints are determined. Green and brown arrows point to breakpoints *between* two homozygous genotypes and *between* the heterozygous and homozygous genotypes, respectively. The resulting recombination map for this chromosomal region is illustrated in a solid bar, in which red, blue, and yellow represent genotypes *ind/ind*, *jap/jap*, and *ind/jap*, respectively. Identified breakpoints are indicated *between* SNPs.

probabilities and make genotype calling again. This process was repeated until the ratios stabilized at 48.94:0.82:50.24 for *ind/ind*:*ind/jap*:*jap/jap*. The higher proportion of the heterozygous genotype than expected could be due to selection for heterozygosity and/or occasional cross-pollen contamination during the process of population development.

With this presumably closest estimate of genotype ratios, we calculated the genotype probabilities and used them for the final genotype calling (Fig. 2B). For a given *indica*:*japonica* SNP ratio, the genotype with the highest expected probability was called. A window with an *indica*:*japonica* SNP ratio of 11:4 or higher was called *ind/ind*, 2:13 or lower was called *jap/jap*, and any ratio in between was called *ind/jap*. The slightly different thresholds for the two homozygous genotypes were due to unequal SNP error rates between the parental genotypes (also see below). This genotype calling strategy had an expected calling accuracy of 99.94% (Equation 8, Methods).

## Recombination breakpoint determination

As the window slid along the chromosome, genotypes were called based on SNP ratios. A genotype remained unchanged until it hit a recombination breakpoint (Fig. 2C). There are two kinds of breakpoints, with one separating two different homozygous genotypes and the other separating a homozygous genotype and the heterozygous genotype; the former is predominant in RILs and the latter occurs most frequently in $F_2$ populations. When the sliding window hit a homozygous/homozygous breakpoint, genotype changed from homozygous into transiently heterozygous, followed by a change to the other homozygous genotype. During the process, the SNP ratios passed through the 8:7/7:8 boundary only once, where the breakpoint was determined. When the sliding window hit a homozygous/heterozygous breakpoint, genotype changed from homozygous into heterozygous, followed by the fluctuation of SNP ratios at the 8:7/7:8 boundary before any change into a homozygous genotype again. The breakpoint was determined at the boundary of the homozygous and heterozygous genotypes. After all genotypes were called and recombination breakpoints were determined, we identified a total of 5074 breakpoints for the 150 RILs, or 33.8 per RIL.

## Error analyses

To set up suitable experimental parameters for a genotyping study using this method, the SNP error rate is a key factor to consider. While the error rates were estimated experimentally above from resequencing of the parental genomes, an independent theoretical calculation would help dissect the source of errors and their relative contribution. Three sources of sequence errors could contribute to total SNP errors, including (1) RIL sequence errors occurring in the three-base indexes, defined as $E_i$; (2) RIL sequence errors occurring in 33-mers, defined as $E_m$; and (3) errors existing in the genome sequences of the mapping parents, 1 and 2, defined as $E_{p1}$ and $E_{p2}$ (Supplemental Fig. 1).

Because any error at the third base of the index automatically disqualifies the sequences from further analysis, errors at the first two sites of the index together determine $E_i$. Given the estimated average per-base error rate of 0.3% at the 5′ end of Illumina GA reads (Dohm et al. 2008), the sequence error rate for the first two bases of the index combined is 0.6%. Because there are two alleles (one from each parent) of a roughly equal frequency in our mapping population, an index error causing incorrect sorting of a 33-mer has a 50% chance to yield a wrong genotype. Thus, $E_i$ is estimated at 0.3%.

For sequence errors in the 33-mers, an error leads to wrong genotype assignment only when the error occurs at the SNP site (1/33 chance) and happens to match the base of the other parental allele (1/3 chance). Given the reported average error rate of 2.8% for the 33-mers (Craig et al. 2008), $E_m$ is estimated at 0.03% (2.8% × 1/33 × 1/3).

Errors in the genome sequences of the parents can cause artificial SNP detection in RILs. We resequenced the genome of *indica* 93-11 with 0.6× coverage on the Illumina GA and estimated that the errors in the original genome sequences would give an SNP error rate, $E_{pi}$, of 3.9%. Because the map-based sequences of *japonica* cv. Nipponbare were about one order of magnitude more accurate than the shotgun sequences of *indica* cv. 93-11 (99.99% versus 99.9% [International Rice Genome Sequencing Project 2005; Yu et al. 2005]), the SNP error rate for *japonica* cv. Nipponbare, $E_{pj}$, is estimated at 0.39%. Taken together, the SNP error rates of homozygous *indica* and *japonica* genotypes are calculated as: $E_{ind/ind} = E_i + E_m + E_{pi} = 0.3\% + 0.03\% + 3.9\% = 4.23\%$, $E_{jap/jap} = E_i + E_m + E_{pj} = 0.3\% + 0.03\% + 0.39\% = 0.72\%$. These estimates are remarkably close to the SNP errors observed from our multiplexed resequencing of both parents (4.12% for *indica* and 0.71% for *japonica*).

Of these sources of errors, the quality of parental genome sequences varies among organisms whose genomes are sequenced with different strategies. To analyze this variable, we conducted simulation for the two most plausible situations (Fig. 3). First, one parent often has high-quality genome sequences in the case where one strain, such as *japonica* cv. Nipponbare, was sequenced to serve as a model system. Then, the genome of the other parent can be resequenced using the next-generation sequencers. In this scenario, one parent has an invariably low error rate (e.g., $E_{p1}$ = 1%), while the error rate for the other parent can vary depending on resequencing coverage (e.g., $E_{p2}$ = 2, 4, . . ., 20%). Second, we consider that the strain with high-quality genome sequences does not serve as a mapping parent but provides the reference for resequencing the genomes of the mapping parents. In this scenario, sequence errors of both parents can vary (e.g., $E_{p1} = E_{p2}$ = 2, 4, . . ., 20%). We also considered two types of mapping populations for the simulation, including a RIL population with genotype ratios set at 49.5:1:49.5, and an $F_2$ population with genotype ratios set at 1:2:1.

In the first scenario, the accuracy of genotype calling stays above 99% for the RIL population even when the error rate for one parent goes up to 20% (Fig. 3A, left). For the $F_2$ population, the accuracy drops below 99% when the error rate of the parent goes up to 6%, but it stays above 95% all the way to the 20% error rate. In the second scenario, the accuracy for RILs drops below 99% at an error rate of 16% for both parents but stays above 95% all the way up to the 20% error rate. For the $F_2$ population, the accuracy drops below 99% and 95% at the error rates of 4% and 12%, respectively (Fig. 3A, right).

We then asked whether increase in the window size, i.e., windows containing a larger number of SNPs for a given physical distance, would improve the accuracy of genotype calling. For the first scenario analyzed above, we took the critical error rate of 6% of the one parent that dropped the accuracy of the $F_2$ population below 99% for simulation. We found that a slight increase in the window size brought the calling accuracy back to the 99% level (Fig. 3B, left). For the second scenario, we conducted simulation for three critical errors rates, including the error rate of 16% for

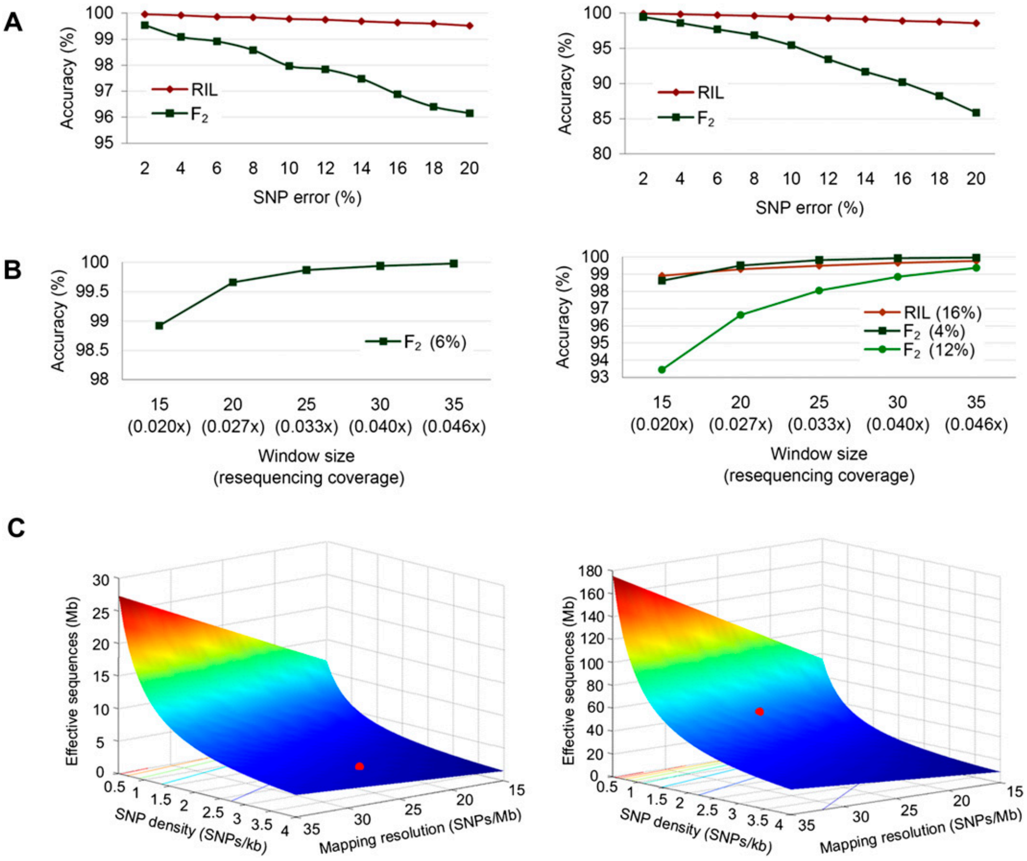


**Figure 3.** Simulation of genotype calling accuracy. (*A*) Effect of parental genome sequence quality on calling accuracy. (*Left*) One parent has high-quality genome sequences that give an SNP error rate of 1%, while the genome sequence quality of the other parent is allowed to vary and gives SNP error rates from 2% to 20%. (*Right*) Genome sequence qualities of both parents are allowed to vary and give the same SNP error rates from 2% to 20%. Two types of populations, RIL and $F_2$, are considered, with ratios of three genotypes set at 49.5:1:49.5 and 1:2:1, respectively. Window size is set at 15. Genotype calling accuracy is calculated according to Equation 8 in Methods. (*B*) The effect of window size on calling accuracy. (*Left*) The critical error rate of 6% that drops the calling accuracy of $F_2$ below 99% in the above figure is used. (*Right*) Three critical error rates are used, including 16% for both parents that drops the calling accuracy of RIL below 99%, 4% for both parents that drops the accuracy of $F_2$ below 99%, and 12% for both parents that drops the accuracy of $F_2$ below 95%, in the *above* figure. When window sizes are measured by the number of SNPs covering the same physical distance, increase in window sizes is equivalent to the increase in resequencing coverage. Rice is taken as an example to show resequencing coverage for the corresponding window size. (*C*) The amount of effective sequences ($S_e$) required for a RIL to reach a range of mapping resolutions (R) as SNP densities (D) vary. (*Left*) Simulation for the rice genome size, 389 Mb. Red dot indicates the location of the rice RIL of this study (D = 3.2 SNPs/kb, R = 25 SNPs/Mb). (*Right*) Simulation for the mouse genome size, 2500 Mb. Red dot indicates $S_e$ required for a mouse RIL with D = 1.3 and R = 25.

both parents that dropped the accuracy below 99% for the RIL population, and the error rates of 4% and 12% that dropped the accuracy of the $F_2$ population below 99% and 95%, respectively. As the window size increased from 15 to 20, which was equivalent to the increase in sequencing coverage of mapping populations from 0.020× to 0.027× for rice, the accuracy rose from 98.9% to 99.3% for RILs with the parental error rate of 16%, and from 98.6% to 99.5% for $F_2$ with the error rate of 4% (Fig. 3B, right). For $F_2$, with the parental error rate of 12%, the accuracy rose from 93.5% to 96.6% as the window size increased from 15 to 20, and to 99.4% as the window size further increased to 35, equivalent to the increase in resequencing coverage from 0.020× to 0.046× for rice.

Finally, we consider the feasibility of this method for mapping larger genomes with variable SNP densities between the parents. We conducted a simulation to estimate the amount of sequences required for a RIL to reach certain mapping resolutions as the SNP density varied. The simulation was run for two genome sizes, 389 Mb and 2500 Mb, equivalent to those of rice and mice, respectively (Equation 9, Methods) (Mouse Genome Sequencing Consortium 2002; International Rice Genome Sequencing Project 2005). The results are illustrated in Figure 3C.

For our rice mapping population, where SNP density between the parents was D = 3.2 SNPs/kb, an average of 3.0 Mb effective sequences ($S_e$) were required for each RIL to reach the resolution R = 25 SNPs/Mb (or 1 SNP per 40 kb). The total amount of sequences obtained for each line on average was 7.2 Mb, of which 3.0 Mb or 42% were defined as effective sequences while the remaining low-quality sequences could not be used for genotyping. Considering rice RILs derived between varieties with a lower SNP density, for example, 1 SNP/kb between two varieties of the *indica* cultivar (data not shown), an average $S_e$ of ~9 Mb, or the total amount of ~21 Mb sequences (assuming the same effective sequence ratio of 42%), is required to reach the same mapping resolution of 25 SNPs/Mb (Fig. 3C, left). Considering next the example of mouse populations that have larger genomes and lower SNP densities than rice (g = 2500 Mb [Mouse Genome Sequencing Consortium 2002] D ≈ 1.3 on average [Frazer et al. 2007]), our simulation indicates that ~48 Mb effective sequences or ~115 Mb total sequences (assuming the same effective sequence ratio of 42%) are required for a mouse RIL to reach the same resolution of R = 25 SNPs/Mb (Fig. 3C, right).

### Bin map construction and quantitative trait loci (QTL) analysis

To conduct genetic analyses, we converted the recombination maps into a skeleton bin map (van Os et al. 2006). We aligned all chromosomes of the 150 RILs and compared them for the minimal of 100-kb intervals (Fig. 4A). Adjacent 100-kb intervals with the same genotype across the entire RIL population were recognized as a single recombination bin (e.g., Fig. 4B,C). In this way, we obtained a total of 2334 recombination bins for the 150 RILs, which captured the vast majority of recombination events detected in the population. The average physical length of the recombination bins was 164 kb, ranging from 100 kb to 5.8 Mb. The genotypes and physical locations of the bins are given in Supplemental Table 1.

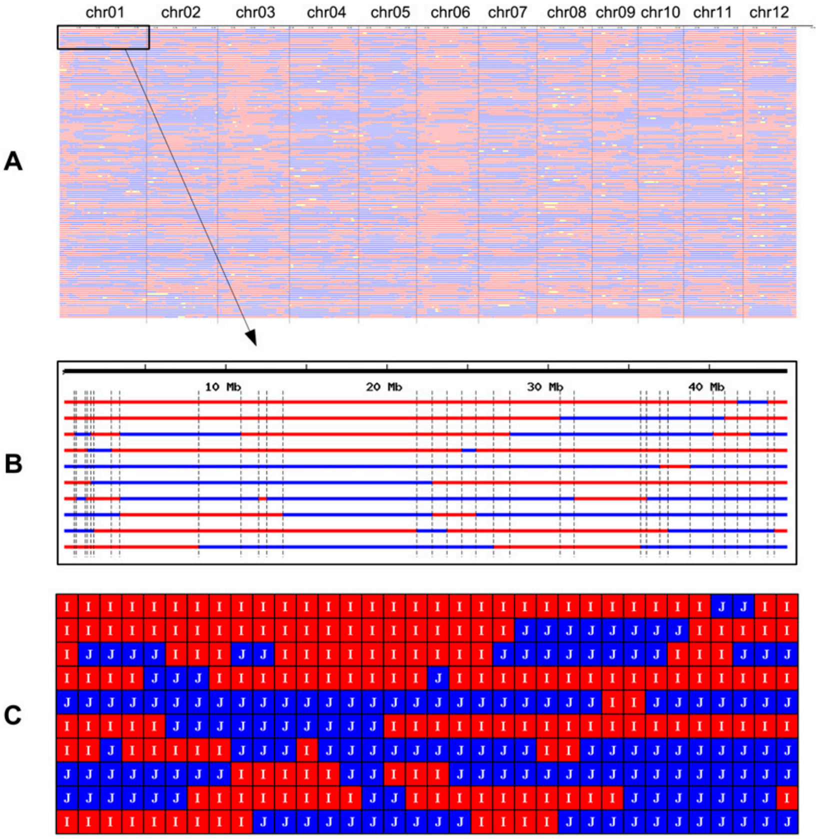
A linkage map was constructed with these bins serving as markers. It had a total genetic distance of 1539.5 cM with an average interval of 0.66 cM between the bins (Supplemental Table 1). This map was used for identifying QTL controlling plant height. When grown in Hangzhou, China, *indica* cv. 93-11 and *japonica* cv. Nipponbare were 124.4 and 84.3 cm tall, respectively; the height of the RILs ranged from 72.0 to 181.0 cm. Our analysis of the 150 RILs detected four QTL, with likelihood of odds (LOD) peaks overlapping with Bin 248 on chromosome 1, Bin 501 on chromosome 2, Bin 731 on chromosome 3, and Bin 2300 on chromosome 12, which explained 31.3%, 11.9%, 7.6%, and 6.6% of phenotypic variance, respectively. The QTL of the largest effect was mapped on Bin 248 occupying physical position of 40.1–40.2 Mb on chromosome 1, which contains the semi-dwarf gene, *sd1*, located at 40.14 Mb, responsible for rice "green revolution" (Sasaki et al. 2002). The result demonstrates that this new genotyping method provides a powerful tool for accurate QTL mapping and subsequent gene cloning.

## Discussion

Recombinant populations were the basis for Mendel's genetic experiments and continued to serve as a key to the study of genes, genomes, and genetic variations. Genotyping with various types of molecular genetic markers has long been a laborious and time-consuming step that limited the power and efficiency of genetic analyses and gene discovery. Here we show that the next-generation sequencing technology allows the development of an ever-fast, cost-effective, informative, and reliable genotyping method. Genotyping a typical mapping population of several hundred individuals with ultraresolution can be completed at a genomics service center in weeks rather than months to years required for conventional types of markers.

Our studies of the rice RILs demonstrated the advantages of this new genotyping method over the commonly used PCR-based approach. Before the sequencing-based method was developed for the $F_{11}$ RIL population, we genotyped the RIL population at its $F_8$ stage using 287 insertion–deletion markers, including microsatellites, which were amplified by PCR and scored on agarose gels (data not shown). The linkage map constructed from the PCR markers had an average coverage of a genetic distance of ~5 cM or a physical distance of ~1.4 Mb per marker, which is higher than the majority of previously reported rice genetic maps. Designing, screening, and collection of these PCR markers took three people more than one year of intensive work to complete. In this study, in contrast, we obtained an average coverage of 40 kb per SNP from data that could be generated within two weeks on the Illumina GA. Thus, the sequencing-based high-throughput method is much more time and cost effective than the conventional PCR-based genotyping approach.

More important is the markedly improved mapping accuracy and resolution of the sequencing-based method. A recombination breakpoint is determined between two SNPs that are 40 kb apart on average. This provides a much finer resolution than the PCR-based mapping, which only allows a resolution equivalent to the marker density, in this case, a 1.4-Mb interval on average (Supplemental Fig. 2A). The new method thus improves the resolution of recombination breakpoints by 35×. Additionally, an average coverage of one SNP per 40 kb almost eliminates the chance of



| | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | J | J | I | I |
| I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | J | J | J | J | J | J | J | J | I | I | I | I | I |
| I | J | J | J | J | I | I | I | J | J | I | I | I | I | I | I | I | I | I | I | J | J | J | J | J | J | J | J | I | I | I | J | J | J |
| I | I | I | I | J | J | J | I | I | I | I | I | I | I | I | I | I | J | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I |
| J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | I | I | J | J | J | J | J | J |
| I | I | I | I | I | J | J | J | J | J | J | J | J | J | J | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I | I |
| I | I | J | I | I | I | I | I | J | J | J | I | J | J | J | J | J | J | J | J | J | J | I | I | J | J | J | J | J | J | J | J | J | J |
| J | J | J | J | J | J | J | J | I | I | I | I | I | J | J | I | I | I | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J | J |
| J | J | J | J | J | J | I | I | I | I | I | I | I | I | J | J | I | I | I | I | I | I | I | I | I | I | J | J | J | J | J | J | J | I |
| I | I | I | I | I | I | I | I | I | J | J | J | J | J | J | J | J | J | J | I | I | I | I | J | J | J | J | J | J | J | J | J | J | J |

**Figure 4.** Recombination and bin maps. (*A*) Aligned recombination maps of 150 rice RILs. Red, *ind*/*ind*; blue, *jap*/*jap*; yellow, *ind*/*jap*. (*B*) Aligned chromosome 1 of the first ten RILs. Scale indicates physical distance. A *vertical* line labels a recombination breakpoint. A region *between* two *vertical* lines across all RILs is recognized as a recombination bin. (*C*) Bin map of the 10 RILs.

missing any double-crossovers in the mapping population. Supplemental Figure 2B illustrates two examples, where double-crossovers were detected by the sequencing-based method but were not identified by the PCR-based method because of the limited marker density.

To evaluate the generality and feasibility of the method for various types of populations from different organisms, several technical issues need be discussed. First, genome sizes and abundance of SNPs between the mapping parents are important parameters for setting the level of resequencing throughput. To empirically determine an appropriate throughput level, we resequenced a randomly selected RIL using one entire lane of the Illumina GA, which yielded 0.32× coverage of the genome. A total of 123,302 SNPs were detected, giving an average of 1 SNP every 3.15 kb. The combination of 16 samples in one lane using three-base indexes would yield 0.02× coverage for each RIL and an SNP density of 1 per 50 kb, which should be sufficient for detecting recombination events in the population of 150 RILs. The average density of 1 SNP per 40 kb actually obtained for the 150 RILs was higher than expected. We also randomly sampled 1/2, 1/4, 1/8, and 1/16 of the SNPs from those generated by 0.32× sequencing of the RIL. The genotypes and recombination breakpoints were consistently identified for the various sample sizes, suggesting that sequencing 16 samples per lane was sufficient for this study.

Based on the SNP density, we chose the window size of 15 SNPs for genotyping, which covered on average 600 kb or 2.3 cM of rice chromosomes. We tested the effect of different window sizes on map construction and QTL analysis by using window sizes of 7, 11, 19, and 23 SNPs. The window sizes of 11, 19, and 23 yielded nearly identical results as the size of 15 in the identification of the largest QTL for plant height on chromosome 1 (Supplemental Fig. 3). However, the 7-SNP-window analysis yielded multiple QTL peaks around *sd1* with relatively low LOD values, suggesting that the small window size may lack adequate power for accurate genotyping. Evidently, the higher resequencing coverage permits the use of larger windows covering the same physical and genetic intervals and consequently more accurate mapping but is subjected to a trade-off of lower throughput and higher cost.

We then conducted a simulation to address the question of how much more work is required for mapping organisms with larger genome sizes and/or lower SNP densities. For mouse RILs, ~15× more sequences are required to reach the same mapping resolution than the rice RILs (Fig. 3C, right). Because the most updated Illumina GAII system has reached the level of throughput as high as 20G, which is 20× higher than the system that generates our rice data, time and cost required for mapping mouse inbred lines of the same size can be less now than that for this rice study. Furthermore, the effective sequence ratio should also increase as sequencing technology improves. In this study, about 40% of the 33-bp single-end reads that mapped to multiple locations of the parental genomes were discarded. By resampling the *japonica* genome sequences (see Methods), we estimated that fewer than 20% of reads would still be mapped to multiple genomic locations if they were 76-bp pair-end reads from the updated Illumina GAII system.

Because this genotyping method requires the availability of genome sequences of the mapping parents, the quality of the genome sequences becomes an important issue. Our simulation suggests that as long as one parent has high-quality genome sequences, the accuracy of genotype calling will be relatively insensitive to even relatively high error rates of the other parent (Fig. 3A), which can always be obtained relatively easily from genome resequencing. When the genome sequences of both parents are subjected to high error rates, e.g., both generated from genome resequencing, the accuracy of genotype calling could drop below an acceptable level more quickly. An $F_2$ population is more sensitive than RILs to higher error rates because it requires much more frequent calling between homozygous and heterozygous genotypes. Nevertheless, the accuracy can be improved rather substantially by even a small scale-up of resequencing coverage for the mapping populations (Fig. 3B), which is becoming increasingly easy to do as the throughput of the next-generation sequencers keeps increasing. Therefore, as sequencing technology improves, the sequence-based genotyping method, which has already looked promising, will continue to gain higher accuracy and efficiency.

The easy adjustment of resequencing throughput also allows us to obtain suitable levels of marker density and breakpoint resolution for addressing different questions with the least time and resource investment. Sequencing coverage can be scaled up any time for the entire or a part of the mapping population whenever new questions arise to require higher marker density or more precisely defined breakpoints. Particularly, recombination breakpoints can be determined very precisely using this method, theoretically within 1 kb for the rice RILs, given high enough resequencing coverage. Such a fine resolution will enable the detection of double-crossovers that usually remain unidentified by other types of genetic markers. This consequently improves the accuracy of QTL detection and enhances the efficiency and success rate of gene cloning. Precisely identified recombination breakpoints also allow the characterization of genome regions that exhibits unique genetic features such as recombination hot spots. Taken together, this genome-based methodology enabled by the next-generation sequencing technology will tremendously simplify and accelerate genetic analyses aiming at large-scale gene discovery and addressing a wide range of biological questions.

## Methods

### Plant material and DNA isolation

RILs were developed from a cross between *Oryza sativa* ssp. *indica* cv. 93-11 and ssp. *japonica* cv. Nipponbare followed by self-fertilization to $F_{11}$. The population was developed in the experimental field at China National Rice Research Institute in Hangzhou, Zhejiang Province. For genotyping, total genomic DNA was isolated from leaf tissues using the DNeasy Plant Mini Kit (Qiagen). For QTL analysis, plant height of a RIL was determined as the mean of measurements from five individuals.

### Multiplexed sequencing by Illumina GA

Genomic DNA of each RIL individual was fragmented to <500 bp by sonication (Bandelin, Sonopuls GM200). The fragments were treated with T4 DNA polymerase, T4 polynucleotide kinase, and Klenow DNA polymerase for end repairing, followed by treatment with Klenow fragment 3′–5′ exo and dATP to generate a protruding 3′ A for ligating with the adaptor carrying a three-base index. Sixteen 3-bp indexes, including AAT, ACT, AGT, ATT, CAT, CCT, CGT, CTT, GAT, GCT, GGT, GTT, TAT, TCT, TGT, and TTT, were linked to adapters as described in Cronn et al. (2008). The indexed DNA samples were run on 2% agarose gels, and fragments of 150–180 bp were recovered and purified. Each sample was then amplified by PCR for 18 cycles. DNAs of 16 RILs with different indexes were mixed in an equal molar concentration and loaded into one lane of the Illumina GA for 36-cycle sequencing, with the Illumina PhiX sample used as control. Image analysis and base calling were performed using Illumina GA processing pipeline V0.2.2.6. The "sol2sange" pipeline in the software MAQ (Li et al. 2008) was used to convert Illumina FASTQ to Sanger standard FASTQ format. A custom Perl script was written to sort sequences based on the 5′ indexes. Thirty-three-mer sequences were obtained after trimming the three-base indexes.

### SNP identification

The latest versions of BGI *indica* 93-11 contigs (ftp://ftp.genomics.org.cn/pub/ricedb/SynVs9311/9311/Sequence/Contig/) were aligned with International Rice Genome Sequencing Project (IRGSP) pseudomolecules of *japonica* cv. Nipponbare (Build 4.0, http://rgp.dna.affrc.go.jp/IRGSP/Build4/build4.html) using the software NUCMER (Kurtz et al. 2004). Candidate SNPs were identified using the DiffSeq program (with default setting of parameters) in the EMBOSS package (Rice et al. 2000). The original trace files of BGI 93-11 (ftp://ftp.ncbi.nih.gov/pub/TraceDB/) were adopted to remove low-quality SNPs using the SSAHASNP program in the ssaha2 package V1.0.9 (Ning et al. 2001). The custom-made *indica* cv. 93-11 pseudomolecules were generated by replacing the bases of the *japonica* cv. Nipponbare pseudomolecules with those of *indica* at the SNP sites.

The 33-mers of the RILs were aligned with the *indica* and *japonica* pseudomolecules using software SSAHA2 V2.0.0 (http://www.sanger.ac.uk/Software/analysis/SSAHA2/). An SNP was detected for a RIL when a 33-mer matched perfectly the unique sequence of one parent and had a 1-bp mismatch with that of the other parent, where an SNP had already been identified between the parents. The genotype of the RIL was then recognized at this SNP site. Low-quality sequences, including primarily short reads that matched multiple locations of either genome and that did not match perfectly with at least one of the parental genomes, were discarded.

To estimate the proportion of short reads matching multiple locations of the parental genome as the function of read length and end types (single vs. pair ends), sequences of different length and end types were randomly extracted from the Nipponbare genome with 0.02× coverage and then aligned with the genome sequences. The proportion of the short reads that matched multiple locations was calculated.

### Genotype calling

Given the SNP error rates of three genotypes, $E_{ind/ind}$, $E_{ind/jap}$, and $E_{jap/jap}$, the probability of finding *k japonica* SNPs in a window of *n* consecutive SNPs for each genotype follows a binomial distribution:

$$p_{ind/ind}(k) = \binom{n}{k} \times E_{ind/ind}{}^{k} \times (1 - E_{ind/ind})^{n-k} \tag{1}$$

$$p_{ind/jap}(k) = \binom{n}{k} \times \left(\frac{1 + E_{ind/jap}}{2}\right)^{k} \times \left(\frac{1 - E_{ind/jap}}{2}\right)^{n-k} \tag{2}$$

$$p_{jap/jap}(k) = \binom{n}{k} \times (1 - E_{jap/jap})^{k} \times E_{jap/jap}{}^{n-k} \tag{3}$$

Given the proportions of the three genotypes in the population, $\lambda_{ind/ind}$, $\lambda_{ind/jap}$, and $\lambda_{jap/jap}$, the expected probabilities of finding the genotypes for an observed *k* are:

$$P_{ind/ind}(k) = p_{ind/ind}(k) \times \lambda_{ind/ind} \tag{4}$$

$$P_{ind/jap}(k) = p_{ind/jap}(k) \times \lambda_{ind/jap} \tag{5}$$

$$P_{jap/jap}(k) = p_{jap/jap}(k) \times \lambda_{jap/jap} \tag{6}$$

A genotype is called based on the highest probability in a given window:

$$P_{\max}(k) = \max\{P_{ind/ind}(k), P_{ind/jap}(k), P_{jap/jap}(k)\} \tag{7}$$

The accuracy of genotyping calling is given by Bayes method:

$$P_n = \sum_{k=0}^{n} P_{\max}(k) \tag{8}$$

A computer program was developed for implementing these algorithms and conducting all analyses including constructing physical and bin maps from the Illumina GA sequence data. The program can be found in the Supplemental Material.

To estimate the amount of effective sequences ($S_e$) to be generated for a RIL, the following equation was used for the simulation:

$$S_e = \frac{g \times R}{D} \tag{9}$$

where variable R represents intended mapping resolution (the number of SNPs to be identified per Mb), variable D represents the SNP density between the two mapping parents (the number of SNPs per kb), and g is a constant representing the genome size of the mapping parents.

### Bin map construction and QTL analysis

The maps of the RILs were aligned and compared for their genotypes for a 100-kb interval. Adjacent 100-kb intervals with the same genotype across all RILs were combined into a recombination bin. The linkage map was constructed from the recombination bins serving as genetic markers using MAPMAKER/EXP V3.0 (Lincoln and Lander 1993). QTL were identified using composite interval mapping implemented in the software package Windows QTL Cartographer V2.5 (Wang et al. 2007). A 10-cM scan window was employed, and the likelihood ratio statistic was computed every 2 cM. LOD values and $R^2$ were determined based on likelihood ratio tests under a hypothesis allowing both additive and dominance effects. QTL were called for LOD values of 3.0 and higher.

## Acknowledgments

We thank Zemin Ning for assistance with sequence alignment, Chuanrang Zhu and Tingting Lu for computational assistance and helpful discussion, and Qifa Zhang and anonymous reviewers for valuable comments on the earlier versions of the paper. This work was supported by the Ministry of Science and Technology of China (the China Rice Functional Genomics Programs, Grant No. 2006AA10A102), the Chinese Academy of Sciences (Grant No. KSCX2-YW-N-024), and the National Natural Science Foundation of China (Grant No. 30821004).

## References

Craig, D.W., Pearson, J.V., Szelinger, S., Sekar, A., Redman, M., Corneveaux, J.J., Pawlowski, T.L., Laub, T., Nunn, G., Stephan, D.A., et al. 2008. Identification of genetic variants using bar-coded multiplexed sequencing. *Nat. Methods* **5:** 887–893.

Cronn, R., Liston, A., Parks, M., Gernandt, D.S., Shen, R., and Mockler, T. 2008. Multiplex sequencing of plant chloroplast genomes using Solexa sequencing-by-synthesis technology. *Nucleic Acids Res.* **36:** e122. doi: 10.1093/nar/gkn502.

Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. 2008. Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res.* **36:** e105. doi: 10.1093/nar/gkn425.

Frazer, K.A., Eskin, E., Kang, H.M., Bogue, M.A., Hinds, D.A., Beilharz, E.J., Gupta, R.V., Montgomery, J., Morenzoni, M.M., Nilsen, G.B., et al. 2007. A sequence-based variation map of 8.27 million SNPs in inbred mouse strains. *Nature* **448:** 1050–1053.

International Rice Genome Sequencing Project. 2005. The map-based sequence of the rice genome. *Nature* **436:** 793–800.

Jeremy, E., Jaroslav, J., Megan, S., Ambika, G., Bin, L., Hei, L., and David, G. 2008. Development and evaluation of a high-throughput, low-cost genotyping platform based on oligonucleotide microarrays in rice. *Plant Methods* **4:** 13. doi: 10.1186/1746-4811-4-13.

Kurtz, S., Phillippy, A., Delcher, A.L., Smoot, M., Shumway, M., Antonescu, C., and Salzberg, S.L. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* **5:** R12. doi: 10.1186/gb-2004-5-2-r12.

Li, H., Ruan, J., and Durbin, R. 2008. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18:** 1851–1858.

Lincoln, S.E. and Lander, S.L. 1993. *Mapmaker/exp 3.0 and mapmaker/qtl 1.1. technical report.* Whitehead Institute of Medical Research, Cambridge, MA.

Meaburn, E., Butcher, L.M., Schalkwyk, L.C., and Plomin, R. 2006. Genotyping pooled DNA using 100K SNP microarrays: A step towards genomewide association scans. *Nucleic Acids Res.* **34:** e2. doi: 10.1093/nar/gnj027.

Mouse Genome Sequencing Consortium. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* **420:** 520–562.

Ning, Z., Cox, A.J., and Mullikin, J.C. 2001. SSAHA: A fast search method for large DNA databases. *Genome Res.* **11:** 1725–1729.

Rice, P., Longden, I., and Bleasby, A. 2000. EMBOSS: The European molecular biology open software suite. *Trends Genet.* **16:** 276–277.

Sasaki, A., Ashikari, M., Ueguchi-Tanaka, M., Itoh, H., Nishimura, A., Swapan, D., Ishiyama, K., Saito, T., Kobayashi, M., Khush, G.S., et al. 2002. Green revolution: A mutant gibberellin-synthesis gene in rice. *Nature* **416:** 701–702.

Singer, T., Fan, Y., Chang, H.S., Zhu, T., Hazen, S.P., and Briggs, S.P. 2006. A high-resolution map of *Arabidopsis* recombinant inbred lines by whole-genome exon array hybridization. *PLoS Genet.* **2:** e144. doi: 10.1371/journal.pgen.0020144.

van Os, H., Andrzejewski, S., Bakker, E., Barrena, I., Bryan, G.J., Caromel, B., Ghareeb, B., Isidore, E., de Jong, W., van Koert, P., et al. 2006. Construction of a 10,000-marker ultradense genetic recombination map of potato: Providing a framework for accelerated gene isolation and a genomewide physical map. *Genetics* **173:** 1075–1087.

Wang, S., Basten, C.J., and Zeng, Z.B. 2007. *Windows QTL Cartographer 2.5*. Department of Statistics, North Carolina State University, Raleigh, NC.

Winzeler, E.A., Richards, D.R., Conway, A.R., Goldstein, A.L., Kalman, S., McCullough, M.J., McCusker, J.H., Stevens, D.A., Wodicka, L., Lockhart, D.J., et al. 1998. Direct allelic variation scanning of the yeast genome. *Science* **281:** 1194–1197.

Yu, J., Wang, J., Lin, W., Li, S.G., Li, H., Zhou, J., Ni, P.X., Dong, W., Hu, S.N., Zeng, C.Q., et al. 2005. The genomes of *Oryza sativa*: A history of duplications. *PLoS Biol.* **3:** 266–281. doi: 10.1371/journal.pbio.0030038.

*Received November 24, 2008; accepted in revised form March 23, 2009.*

# High-throughput genotyping by whole-genome resequencing

Xuehui Huang, Qi Feng, Qian Qian, et al.

*Genome Res.* 2009 19: 1068-1076 originally published online May 6, 2009
Access the most recent version at doi:10.1101/gr.089516.108