the plant journal
SEB Society for Experimental Biology

The Plant Journal (2013) **74,** 174–183 doi: 10.1111/tpj.12105

TECHNICAL ADVANCE/RESOURCE

QTL-seq: rapid mapping of quantitative trait loci in rice by whole genome resequencing of DNA from two bulked populations

Hiroki Takagi[1,2], Akira Abe[2,3], Kentaro Yoshida[1], Shunichi Kosugi[1], Satoshi Natsume[1], Chikako Mitsuoka[1], Aiko Uemura[1], Hiroe Utsushi[1], Muluneh Tamiru[1], Shohei Takuno[4], Hideki Innan[5], Liliana M. Cano[6], Sophien Kamoun[6] and Ryohei Terauchi[1,*]

[1]Iwate Biotechnology Research Center, Kitakami, Iwate, 024-0003, Japan,
[2]United Graduate School of Iwate University, Morioka, Iwate, 020-8550, Japan,
[3]Iwate Agricultural Research Center, Kitakami, Iwate, 024-0003, Japan,
[4]Department of Plant Sciences, University of California, Davis, CA 95616, USA,
[5]Graduate University for Advanced Studies, Hayama, Japan, and
[6]The Sainsbury Laboratory, Norwich Research Park, Norwich, UK

Received 7 September 2012; revised 13 December 2012; accepted 20 December 2012; published online 05 January 2013.
*For correspondence (e-mail terauchi@ibrc.or.jp).

SUMMARY

The majority of agronomically important crop traits are quantitative, meaning that they are controlled by multiple genes each with a small effect (quantitative trait loci, QTLs). Mapping and isolation of QTLs is important for efficient crop breeding by marker-assisted selection (MAS) and for a better understanding of the molecular mechanisms underlying the traits. However, since it requires the development and selection of DNA markers for linkage analysis, QTL analysis has been time-consuming and labor-intensive. Here we report the rapid identification of plant QTLs by whole-genome resequencing of DNAs from two populations each composed of 20–50 individuals showing extreme opposite trait values for a given phenotype in a segregating progeny. We propose to name this approach QTL-seq as applied to plant species. We applied QTL-seq to rice recombinant inbred lines and F_2 populations and successfully identified QTLs for important agronomic traits, such as partial resistance to the fungal rice blast disease and seedling vigor. Simulation study showed that QTL-seq is able to detect QTLs over wide ranges of experimental variables, and the method can be generally applied in population genomics studies to rapidly identify genomic regions that underwent artificial or natural selective sweeps.

Keywords: quantitative trait loci, breeding, whole genome sequencing, next generation sequencer, selective sweep, technical advance.

INTRODUCTION

The world's population has already exceeded 7 billion and is still growing, while the amount of land suitable for agriculture is decreasing due to a variety of factors such as rapid climate change. Therefore there is a great demand for efficient crop improvement to increase yield without further expanding farmland and damaging the environment (Godfray, 2010; David *et al*., 2011).

In crop plants, multiple genes each with a relatively minor effect control the majority of agronomically important traits. These genes are called quantitative trait loci (QTLs) (Falconer and Mackay, 1996). Identification of QTLs is an important task in plant breeding. Once a QTL controlling a favorable trait is mapped with closely linked DNA markers, it is introduced into an elite cultivar by crossing of the recurrent elite parent to the donor plant. Following each backcross, the progeny inheriting the desirable QTL are selected by using tightly linked DNA makers, a process known as marker-assisted selection (MAS; Ashikari and Matsuoka, 2006). Marker-assisted selection reduces the effort and time needed for phenotype evaluation of the progeny during successive rounds of selection, and also improves introgression breeding.

Traditionally QTLs have been identified by linkage analysis of progeny derived from a cross between parents showing contrasting phenotypes for a trait of interest. To perform linkage analysis, DNA markers capable of discriminating parental genomes are required. Due to this requirement, parents for crosses are selected from genetically distantly related cultivars. This entails that parents may be different in many QTLs controlling a given phenotype, complicating the isolation of individual loci. On the other hand, whenever closely related parents are used, identification of sufficient DNA markers for linkage analysis becomes a limiting step.

Bulked-segregant analysis (BSA) is an elegant method to identify DNA markers tightly linked to the causal gene for a given phenotype (Giovannoni *et al.*, 1991; Michelmore *et al.*, 1991). Following a cross between parental lines showing contrasting phenotypes, the resulting F_2 progeny are scored for segregation of the phenotype. Two bulked DNA samples are generated from the progeny showing contrasting phenotypes, and DNA markers exhibiting differences between the two bulks are screened. In the original reports, DNAs bulked from F_2 progeny were screened with restriction fragment polymorphisms (RFLPs) and random amplified polymorphic DNA (RAPD) markers to identify the markers linked to the traits of interest. Later, BSA was applied to identify QTLs (Mansur *et al.*, 1993; Darvasi and Soller, 1994), which is sometimes called 'selective DNA pooling'. However, in these analyses, the availability of DNA markers was the main factor limiting effectiveness of the methods. Furthermore, genotyping of each marker for the two bulked DNAs is still time-consuming and costly.

Recent development of whole genome sequencing has accelerated the analysis of QTLs in yeast, a model organism with a relatively small genome size (12.5 Mb). Ehrenreich *et al.* (2010) made a cross between two diploid yeast strains and obtained a large number of haploid progeny. They then applied BSA to select two populations with extreme phenotypes, and genotyped the bulked DNA with a single nucleotide polymorphism (SNP) microarray and whole genome sequencing, which successfully identified the location of QTLs involved in resistance to various chemical compounds. The proposed method is called X-QTL since an extremely large number of progeny were used in each bulk. Similar applications of whole-genome sequencing to BSA are reported in yeast with successful identification of QTLs for xylose utilization (Wenger *et al.*, 2010), heat tolerance (Parts *et al.*, 2011), and ethanol tolerance (Swinnen *et al.*, 2012). However, the application of whole genome sequencing to BSA for identifying QTLs in plant with much larger genome sizes than yeast has not been reported to date.

In this paper, we report plant QTL identification using whole-genome resequencing of two DNA bulks of progeny (each with 20–50 individuals) showing extreme phenotypic values by next-generation sequencing (NGS) technology. Since this approach has a wide applicability in QTL identification in plant species, including crops, we propose to name the method QTL-seq as specifically applied to plant species. Because it does not require DNA marker development and genotyping, the most time-consuming and costly procedure needed for the conventional QTL analysis, QTL-seq allows the rapid identification of QTLs.

RESULTS

Principle of QTL-seq

The principle of QTL-seq is shown in Figure 1 and is explained by taking rice as an example. QTL-seq combines bulked-segregant analysis (Giovannoni *et al.*, 1991; Michelmore *et al.*, 1991; Mansur *et al.*, 1993; Darvasi and Soller, 1994) and whole-genome resequencing for rapid identification of the genomic regions that differ between the two parents used in a genetic cross and also contribute to the higher and lower values of the traits of interest among the resulting progeny. For QTL mapping using QTL-seq, we first generate a mapping population by crossing two cultivars showing contrasting phenotypes for the traits of interest. In Figure 1, we assume that we are interested in plant height and that cultivars A and B have a low and high stature, respectively. Different kinds of mapping populations can be used for QTL-seq depending on the traits to be studied. Recombinant inbred lines (RILs) and doubled haploids (DH) show a high degree of homozygosity, and individuals in each line can be regarded as proxy clones that allow replicated measurements of the phenotype, and thus are suitable for detecting QTLs of minor effects. The advantage of using an F_2 population is the a short time required for its generation. However, no replicated measurements are possible for each genotype. As a result, the approach is not suitable for detecting minor effect QTLs.

After the progeny of a mapping population are measured for the focused trait, we score segregation of the phenotype. If the number of QTLs involved in the trait variation is multiple, frequency distribution of measured values will be close to the normal (Gaussian) distribution (Figure 1a). Here, we focused on the multiple progeny showing extreme phenotypes, i.e. those exhibiting the highest and the lowest extreme values. We sampled DNA from 10 to 20 individuals from each extremity and bulked them to generate 'Highest' bulk and 'Lowest' bulk. Each of the bulked DNAs was applied to whole genome resequencing with a > 6*x* genome coverage. We expect the bulked DNA to contain genomes from both parents in a 1:1 ratio for the majority of genomic regions. However, we should detect unequal representation of the genomes from the two parents in the genomic regions harboring QTL for the phenotypic difference between 'Highest' and 'Lowest' bulks.

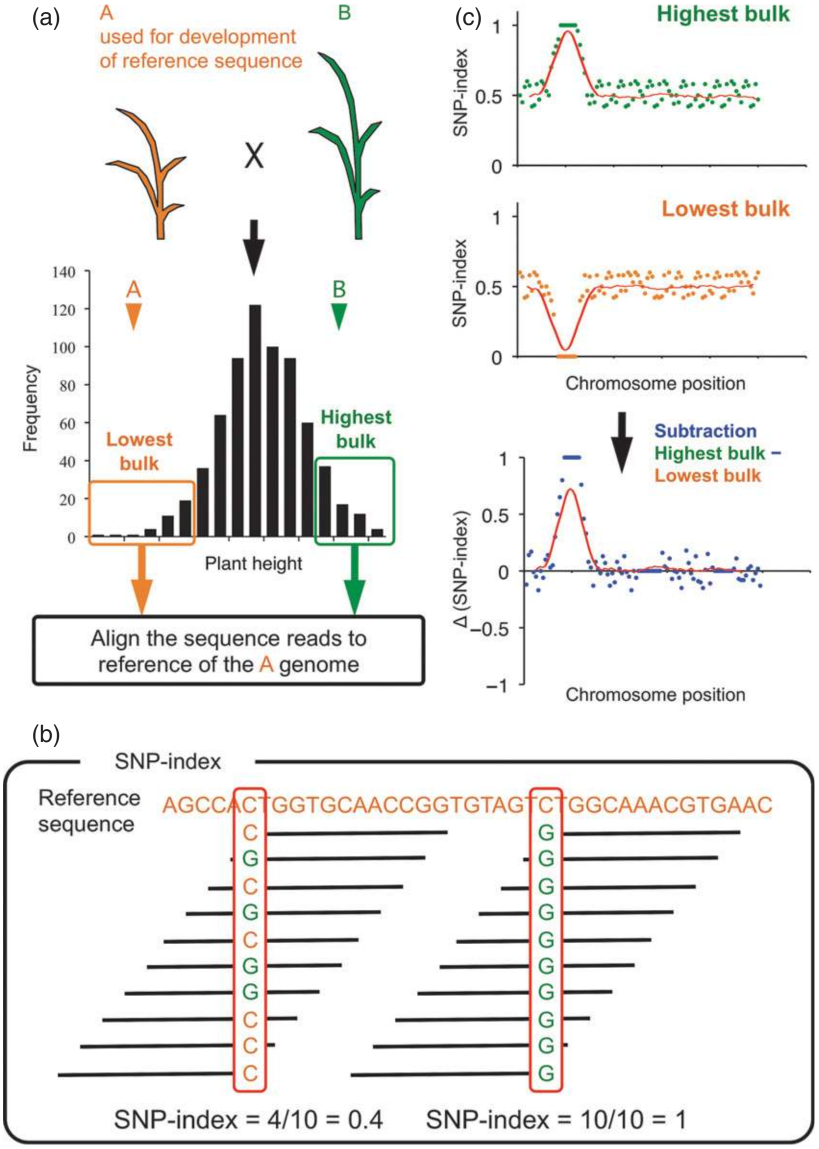

Figure 1. A simplified scheme of QTL-seq as applied to rice.

(a) Two inbred cultivars with contrasting phenotypes are crossed to generate F_2 progeny that are segregating for the trait value. In this example, parent A has low stature while parent B has high stature. Since multiple quantitative trait loci (QTLs) control plant height, frequency of plant height among the F_2 progeny follows the normal distribution. We select multiple progeny with highest and lowest stature, and bulk their DNAs to make 'Highest' and 'Lowest' bulk, respectively. These DNA bulks are applied to whole genome resequencing and aligned to the reference sequence of cultivar A to calculate the single nucleotide polymorphism (SNP)-index.

(b) Definition of SNP-index. Short reads generated by whole-genome sequencing are aligned to the reference sequence. If 10 short reads cover a given nucleotide position, the coverage of the site is 10. Among the 10 reads, if four contain a SNP different from the reference nucleotide, the SNP-index is defined as 0.4. On the other hand, if all the reads harbor a SNP different from the reference, the SNP-index is 1.0.

(c) Examples of SNP-index plot. The QTL can be identified as peaks or valleys of the SNP-index plot. Each spot corresponds to a SNP, and the *x*-axis corresponds to the chromosomal position. Lines are average values of SNP-index or Δ(SNP-index) drawn by sliding window analysis. Top: SNP-index plot of 'Highest' bulk. Middle: SNP-index plot of 'Lowest' bulk. Bottom: a plot of Δ(SNP-index).

To examine the relative amount of the genomes derived from the two parents, we evaluated the proportion of short reads corresponding to each of the two parental genomes that can be discriminated by single nucleotide polymorphisms (SNPs) available between the two. After aligning the sequence data to the reference sequence of either of the two parents, we counted the number (k) of short reads harboring SNPs that are different from the reference sequence. We defined the proportion of k in the total short reads (n) covering a particular genomic position ($=k\ n^{-1}$) as the SNP-index (Figure 1b; Abe *et al.*, 2012a). The SNP-index is 0 if the entire short reads contain genomic fragments from the parent that was used as a reference sequence. The SNP-index is 1 if all the short reads represent the genome from the other parent. A SNP-index of 0.5 means an equal contribution of both parents' genomes to the bulked progeny. Accordingly, the SNP-index is calculated for all the SNPs detected between the two parents, and the relationships between SNP-index and SNP position in the genome is graphically represented (Figure 1c). We carried out this procedure separately for the 'Highest' and 'Lowest' bulk sequences.

In practice, SNPs with SNP-index < 0.3 in both bulked sequences are filtered out during SNP calling because they cannot be discriminated from spurious SNPs caused by sequencing or alignment errors. However, if SNPs with a SNP-index of 0.3 or greater are present in only one of the two bulks, we consider them as real SNPs and assume their presence in the other bulk as well. In this case, we make use of the SNP-index value of the other bulked DNA even if it is <0.3 (see Experimental Procedures). By taking an average of SNP-indices of SNPs located in a given genomic interval, sliding window analysis can be applied to facilitate visualization of the graphs. We expect the SNP-index graphs of ‘Highest’ and ‘Lowest’ bulks to be identical for the genomic regions that are not relevant to the phenotypic difference between the two. However, the genomic regions harboring QTLs that contribute to the difference in the phenotype between the two bulks should exhibit unequal contributions from the two parental genomes. Furthermore, SNP-indices of these regions for ‘Highest’ and ‘Lowest’ bulks would appear as mirror images with respect to the line of SNP-index = 0.5. Such regions are expected to have a high probability of containing QTLs responsible for the trait difference between the ‘Highest’ and ‘Lowest’ bulks. Comparison of the two graphs is important to discern the QTLs from the genomic regions showing segregation distortion caused by reasons other than the imposed artificial selection (e.g. meiotic drive), and result in departure of the SNP-index from 0.5 in both bulks in the same direction. It is therefore convenient to combine the two graphs for ‘Highest’ and ‘Lowest’ bulks by subtracting the SNP-index value of the latter from the former to generate the graph of Δ(SNP-index) (Figure 1c). In this graph, Δ(SNP-index) = 1 if the bulked DNA comprises only parent B genome, Δ(SNP-index) = −1 if it is of parent A genome only and Δ(SNP-index) = 0 if both parents have the same SNP-indices at the genomic regions.

QTL-seq applied to RILs: detection of QTLs controlling partial resistance to rice blast in Nortai

We applied QTL-seq for the detection of QTLs involved in partial resistance of the rice cultivar Nortai against the fungal pathogen *Magnaporthe oryzae*, the causal agent of rice blast disease. Resistance of Nortai to *M. oryzae* race 037.1 does not seem to be mediated by typical R-genes; the hypersensitive response cannot be clearly distinguished and the trait is quantitative and difficult to measure. We crossed Nortai to the cultivar Hitomebore that is highly susceptible to the race 037.1 and obtained F_2 (Figure 2a). Each F_2 progeny was established as a line and brought to the F_7 generation by a single-seed descent method to generate a total of 241 RILs.

Using the 241 RILs, we carried out *M. oryzae* inoculation assay to assess the resistance of the progeny. Susceptibility of the progeny was measured and categorized to seven classes from class 4 (resistant) to class 10 (highly susceptible; Figure 2b). The inoculation assay was conducted four times over 4 years to ensure the correct scoring of each RIL (Figure S1 in Supporting Information). The average

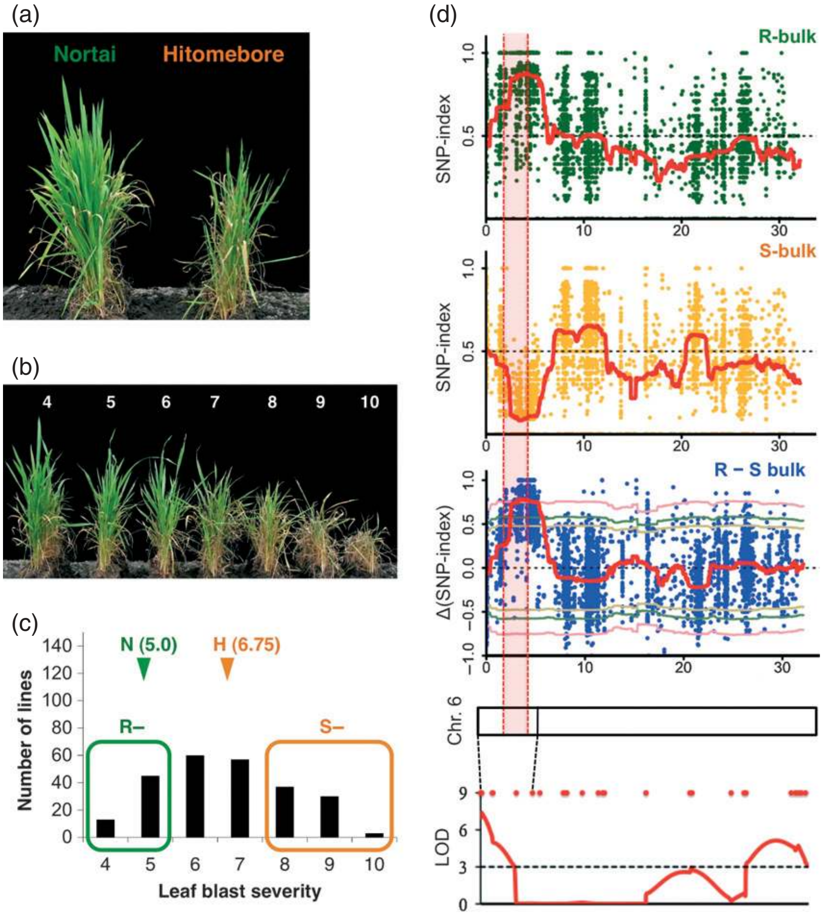

Figure 2. QTL-seq applied to rice recombinant inbred lines (RILs) identifies quantitative trait loci (QTLs) conferring partial blast resistance to the Nortai cultivar.
(a) Phenotype of two rice cultivars Nortai and Hitomebore 2 weeks after inoculation with a compatible race (race 037.1) of blast fungus. Nortai shows partial resistance whereas Hitomebore is susceptible.
(b) Scores (4, highly resistant, to 10, highly susceptible) assigned to different levels of partial resistance in RILs derived from a cross between Nortai and Hitomebore.
(c) Frequency distribution of partial resistance levels of 241 RILs. The *x*-axis corresponds to the level of partial resistance as given in (b). N and H indicate the average resistance level of Nortai and Hitomebore, respectively. The DNAs of RILs with resistance levels 4 and 5 were bulked to make resistance (R-) bulk, and those of levels 8–10 were bulked to make susceptible (S-) bulk.
(d) Single nucleotide polymorphism (SNP)-index plots of R-bulk (top) and S-bulk (next to the top), Δ(SNP-index) plot (next to the bottom) of chromosome 6 with statistical confidence intervals under the null hypothesis of no QTLs (gray, $P < 0.1$; green, $P < 0.05$; pink, $P < 0.01$) and log of odds (LOD) score plot of partial resistance QTLs as obtained by classical QTL analysis of 241 RILs (bottom).

score of Nortai and Hitomebore in four trials was 4.75 and 6.38, respectively. The frequency distributions of RILs falling into different classes is close to normal distribution, suggesting that multiple genes control the partial resistance of RILs (Figure 2c). We defined 20 RILs consistently showing high resistance (class 4 and 5) as Resistant (R-) progeny, and an additional set of 20 RILs consistently showing high susceptibility (classes 8, 9 and 10) as Susceptible (S-) progeny. Genomic DNA of R-progeny was bulked in an equal ratio to generate R-bulk DNA, and that of S-progeny was bulked to generate S-bulk DNA.

Each DNA bulk was subjected to whole-genome resequencing using an Illumina GAIIx sequencer. We obtained a total of 57.9 and 62.4 million sequence reads (each of 75 bp) from DNA bulk of R-progeny and S-progeny, respectively (Table S1). These reads were aligned to the reference sequence of the Hitomebore cultivar using BWA software (Li and Durbin, 2009). The average read depth was >6.88*x* in both bulked DNA (Table S1). A total of 161 563 SNPs were identified between Nortai and Hitomebore genomes (Table S2), and the SNP-index was calculated for each SNP (Figure 1b; Abe *et al.*, 2012a). Graphs showing relationships between SNP-index and genomic positions are given in Figures 2(d) and S2. We found highly contrasting patterns of SNP-index graphs for R-bulk and S-bulk in the region between 2.39 and 4.39 Mb on chromosome 6 as shown in Figure 2(d). The resistant RILs mainly had Nortai-type genomic segments in the 2.39 to 4.39 Mb region of chromosome 6, whereas susceptible RILs had Hitomebore-type genome in the same region, indicating that there is a major QTL differentiating Nortai and Hitomebore partial resistance located at this genomic region. Combining the information from the two graphs for R-bulk and S-bulk, we made a graph of Δ(SNP-index) whereby the Δ(SNP-index) = (SNP-index of R-bulk) – (SNP-index of S-bulk) (Figures 2d and S2). This revealed that most of the genomic regions show Δ(SNP-index) = 0, but some genomic regions exhibit positive or negative values of Δ(SNP-index). These may correspond to QTLs governing the difference between the R- and S-progeny. We calculated statistical confidence intervals of Δ(SNP-index) for all the SNP positions with given read depths under the null hypothesis of no QTLs, and plotted them along with Δ (SNP-index) (Experimental Procedures; Figures 2d and S2). The chance that Δ(SNP-index) becomes higher than 0.79 as observed for the chromosomal region of 2.39–4.39 Mb is $P < 0.01$ under the null hypothesis.

To verify the candidate QTLs detected by the QTL-seq method, we applied traditional QTL analysis to the same 241 RILs using SNP markers. A total of 425 SNP markers covering the genome were used for genotyping the 241 RILs by the Golden Gate Assay on the iSCAN platform (Illumina, http://www.illumina.com/). The data were analyzed by QTL-CARTOGRAPHER (Basten *et al.*, 2005) and LOD (log of odds) scores of linkage between SNP markers and the trait were obtained (Figure 2d**)**. The highest LOD score (LOD = 7.38) was observed in the interval of SNP makers located at 0 and 4.9 Mb. This interval corresponded to the genomic region identified by the QTL-seq method (Figure 2d). This result demonstrates that QTL-seq allows the rapid detection of QTLs using RILs. We named the interval 2.39–4.39 Mb on chromosome 6 of Nortai *qPi-nor1*(t) as the location of the partial resistance in Nortai.

Using RILs derived from two other cross-combinations of rice cultivars, we applied QTL-seq and identified the peaks of the Δ(SNP-index) plot presumably corresponding to a major QTL: a Δ(SNP-index) peak for lower grain amylose content in the cultivar Iwate96 as compared with Hitomebore (Figure S3), although this peak was not statistically significant ($0.05 < P < 0.1$). We identified four Δ(SNP-index) peaks for enhanced seedling vigor under low-temperature conditions of a cultivar Arroz da Terra as compared with Iwatekko (Figure S4). Of the four Δ(SNP-index) peaks detected for seedling vigor, three corresponded to QTLs previously identified by conventional QTL analysis [*qLTG3-2* and *qLTG11* (Fujino *et al.*, 2008); *qLTG3-1* (Miura *et al.*, 2001)]. The three Δ(SNP-index) peaks were statistically significant (the peak for *qLTG3-1*, $P < 0.01$; and the peaks for *qLTG3-2* and *qLTG11*, $P < 0.05$).

Application of QTL-seq to F_2 progeny

We further examined the possibility of applying QTL-seq to an F_2 population, which is much easier to generate than RILs of advanced generations. A *japonica* type cultivar Dunghan Shali is known to have a strong seedling vigor compared with Hitomebore (Figure 3a). We have recently fine-mapped a major QTL, *qPHS3-2*, on chromosome 3 that confers the seedling vigor in Dunghan Shali using conventional QTL analysis of RILs of the F_7 generation derived from a cross between Dunghan Shali and a *japonica* cultivar Kakehashi (Abe *et al.*, 2012b). The QTL most likely corresponds to a gene *OsGA20ox1*, a gene involved in gibberelllin (GA) byosynthesis (Abe *et al.*, 2012b; Yano *et al.*, 2012). Using Dunghan Shali, we addressed whether QTL-seq can detect *qPHS3-2* in the F_2 progeny derived from a cross between Dunghan Shali and Hitomebore. After crossing Dunghan Shali to Hitomebore, we obtained F_2 progeny. Selfed seeds of a total of 531 F_2 individuals were scored for their seedling height after 14 days of imbibition in water at 25°C. The variation in seedling height in seedling followed a normal distribution, indicative of the involvement of multiple genes in determining this character (Figure 3b). Two DNA bulks were prepared; the 50 tallest individuals as ‘H-bulk’ and the 50 shortest individuals as ‘L-bulk’, and were used for QTL-seq analysis (Figures 3c and S2). By examining the Δ(SNP-index) plot, we identified two genomic positions exhibiting the highest Δ(SNP-index) values: the region on chromosome 3 from 36.21 to

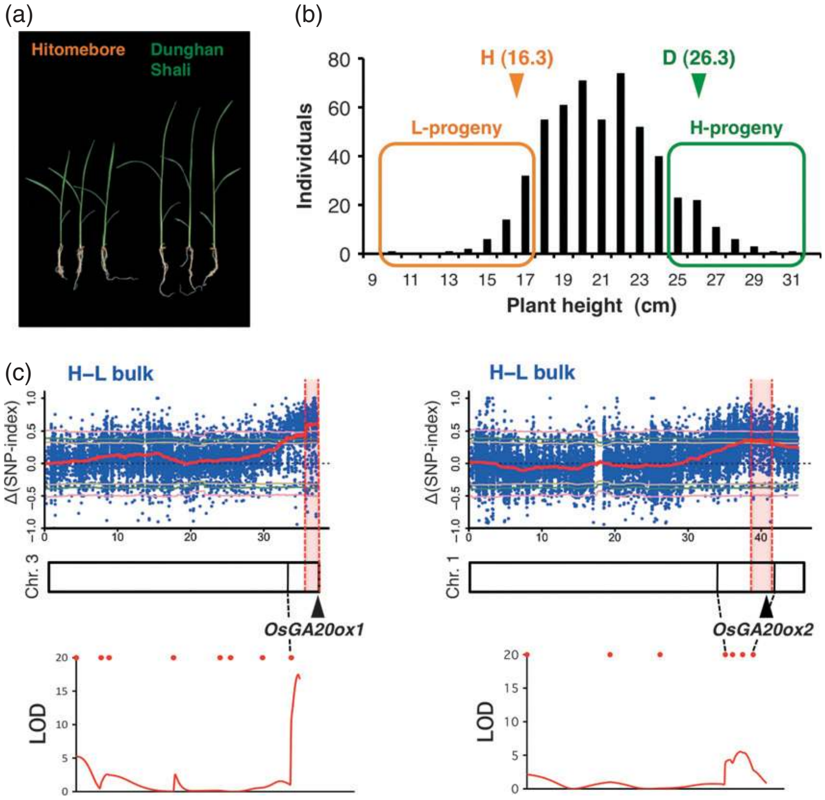

Figure 3. QTL-seq applied to rice F_2 progeny identifies quantitative trait loci (QTLs) involved in seedling vigor.
(a) Seedlings of Hitomebore and Dunghan Shali 10 days after water imbibition. Dunghan Shali shows higher seedling vigor compared with Hitomebore.
(b) Frequency distribution of seedling height in 531 F_2 progenies 14 days after water imbibition. H and D indicate the average seedling height of Hitomebore and Dunghun Shali, respectively. We selected 50 F_2 progeny shorter than 18 cm to make Low (L-) bulk and 50 progeny taller than 24 cm to make High (H-) bulk, and applied to QTL-seq using the Hitomebore reference genome sequence.
(c) Results of QTL-seq for chromosome 3 (left) and 1 (right). The Δ(SNP-index) plot (top) with statistical confidence intervals under the null hypothesis of no QTL (gray, $P < 0.1$; green, $P < 0.05$; pink, $P < 0.01$) and log of odds (LOD) score plot of QTL controlling plant height as obtained by classical QTL analysis of 250 recombinant inbred lines of the F_7 generation (bottom).

37.31 Mb with Δ(SNP-index) = 0.61 (statistical significance under the null hypothesis: $P < 0.01$) and the region on chromosome 1 from 39.08 to 41.08 Mb with Δ(SNP-index) = 0.67 ($P < 0.05$). This former position corresponded exactly to the reported *qPHS3-2*, most probably the locus of *OsGA20ox1.* Likewise, the latter position was also previously detected as a minor QTL (*qPHS-1*) (Abe *et al.*, 2012b). This result demonstrates that QTLs identified by conventional QTL mapping using RILs of F_7 generations could be successfully recovered by QTL-seq using the F_2 generation.

Simulation of QTL-seq

As shown above, QTL-seq successfully identified genomic regions controlling quantitative traits in the examples of rice RILs and F_2 families. More generally, we are interested in how experimental variables affect the performance of QTL-seq to faithfully detect QTLs. To this end, we carried out a computer simulation of QTL-seq by changing variables like (i) the contribution of QTLs to phenotypic variation, (ii) the percentage of individuals to be selected, (iii) the read depth, and (iv) the dominance effect of the QTL locus on phenotype. We assumed that the rice genome size is roughly 360 Mb and the recombination rate is 4 cM Mb^{-1}. We also postulated that 150 000 SNPs between the two parents are distributed with equal intervals (i.e. a SNP every 2.4 kb). For the QTL-seq process, it is assumed that from among all progeny individuals of F_2 or F_7 generations, we select p% each of progeny with opposite extreme trait values to make 'Highest' and 'Lowest' bulks, and we sample n random alleles from each bulk to represent the depth of sequencing. Using these n alleles, we calculated Δ(SNP-index). Since we routinely take an average of Δ(SNP-index) of 10 consecutive SNPs to obtain a sliding window value m, we evaluated the behavior of m by simulations. With 10 000 replications of the simulation, we found that the 99% cutoff value of $|m|$ for SNPs that are not selected (null distribution) in F_2 depends on the percentages of individuals in the bulk (p) and read depth of the focused region (Figure 4a, left). The intervals of values of m become narrower as the coverage and the percentage of individuals in each bulk (p) increases (for our application to rice, the 99% cutoff of $|m|$ would be 0.29 given $n = 10$ and $p = 0.15$). We also applied the same simulation to RILs of the F_7 generation (Figure 4a, right). The null distribution of $|m|$ is wider than that for an F_2 population.

We next explored the power to detect a QTL. In practice, we place a QTL in the simulated genome assuming that the relative contribution of the QTL to the total phenotype variation is given by Qp. Then, we evaluated the power as the proportion of the simulation replications with $|m|$ around the QTL larger than the 99% cutoff value obtained earlier. In this power simulation, bi-allelic states were allowed at the focal QTL, and for the dominance effect we considered two cases, codominance and complete dominance. For Qp,

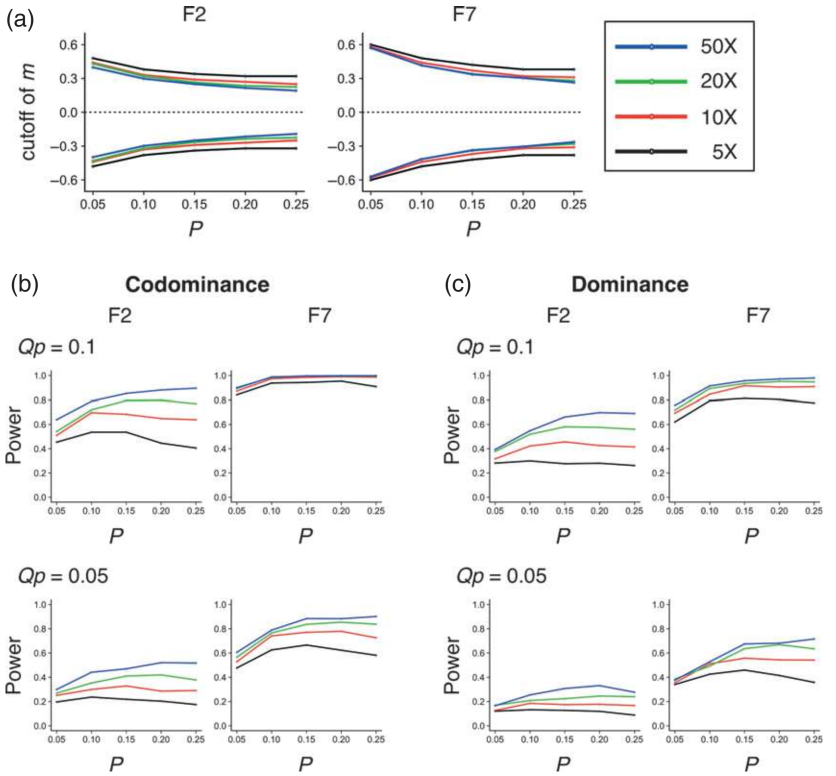

Figure 4. Simulation reveals the capability of QTL-seq to detect quantitative trait loci (QTLs) in a wide range of values of experimental variables.

(a) Ninety-nine per cent intervals of the null distribution of *m* statistics [average value of Δ (SNP-index) of 10 consecutive single nucleotide polymorphisms (SNPs)]. The *x*- and *y*-axes represent the percentage of individuals in each bulk (*p*) and the *m* value, respectively. The results for F_2 progeny (left) and F_7 recombinant inbred line progeny (right) are shown. Different read depths (50×, 20×, 10× and 5×) are indicated by different colors (inset).

(b), (c) The power of QTL-seq for detecting QTLs in the cases of codominance (b) and dominance (c). Two values of the QTL effect [Qp = 0.1 (top) and Qp = 0.05 (bottom)], as well as two types of populations (F_2 and F_7) were tested.

two values (0.05 and 0.1) were used. Note that Qp is the relative contribution of this QTL to the total phenotype variation, which includes everything other than the genetic effect of the focal QTL (that is, the environmental factors and the genetic contributions from other QTLs that are not specified here). Figure 4(b) and (c) shows the results for the cases of codominance and complete dominance, respectively. We found that the larger read depth increases the power in all cases. The power is higher when Qp = 0.1 than the cases of Qp = 0.05. It appears that higher power is expected when the QTL allele is codominant (additive) as compared with complete dominance, and there is an optimum value for the percentage of individuals in each bulk (*p*). When the value is small, the power is low, probably because there would be too much sampling variance. As *p* increases, the power increases, but it starts to decrease when the sample size is so large that many individuals with intermediate phenotype are included in the bulking. We found higher power in F_7 RILs than in F_2 populations, and thus concluded that in QTL-seq application to the F_2 populations, p = 0.15 and n = 20 would be a reasonable choice, and that the method has reasonable power to detect QTLs with a relative contribution of roughly 10% (Qp = 0.1).

DISCUSSION

Two types of genetic variations, the ones derived from artificial mutagenesis and those naturally occurring in landraces and wild crop relatives, have been used in plant breeding. Mutant lines generated by artificial mutagenesis are valuable for isolating agronomically important genes. To this end, we have recently developed MutMap, an efficient method to identify the causal mutation of a given phenotype by whole genome resequencing of the bulked DNA of progeny showing mutant phenotype (Abe *et al.*, 2012a). Although MutMap is a powerful technique, crop breeding has mostly depended on genetic variations available among different cultivars and species in what is called QTL breeding. This is in part because naturally occurring variants harbor a potentially larger repertoire of useful alleles than the artificially generated mutants due to the larger number of mutations accumulated over long time in nature. Therefore, analysis of the QTL variations among natural variants is important for enhancing breeding by isolating useful alleles of the genes controlling agronomically important traits (Yano, 2001). However, conventional QTL analysis is a laborious process requiring the development of DNA markers and the generation of a large number of advanced generation progeny. Here we demonstrated the successful application of whole-genome resequencing for detecting rice QTLs for agronomically important traits, including partial resistance and seedling vigor, using RILs and F_2 populations, respectively. The major advantage of QTL-seq is that it does not necessitate DNA marker development and marker genotyping for mapping purposes. The SNPs available between the parental lines serve as such markers, thus reducing the cost and time required for marker development and genotyping. Furthermore, the use of SNP-index allows accurate evaluation of the frequencies of parental alleles in a subset of progeny of a given genomic position. These two key attributes make QTL-seq an attractive method for quick and cost-effective identification of QTLs.

Bulked-segregant analysis was first applied to facilitate the linkage analysis of discrete characters in F_2 populations (Giovannoni *et al.*, 1991; Michelmore *et al.*, 1991). In these studies, F_2 progeny showing two discrete characters were isolated, and DNA from the F_2 individuals were pooled to make two DNA bulks corresponding to the two character types. After a battery of DNA markers including RAPD markers (Williams *et al.*, 1990) were tested for these two bulked DNAs, markers showing differences between the two DNA bulks were selected to represent the DNA markers linked to the gene(s) responsible for the difference in the characters. This original bulked-segregant method was later extended to QTL analysis. After RILs or F_2 were scored for the phenotypes, progeny showing extreme opposite phenotypes were selected, and these DNAs were separately bulked to find DNA markers showing linkage with the phenotypic differences ('selective DNA pooling'; Mansur *et al.*, 1993; Darvasi and Soller, 1994). This latter method is in principle similar to QTL-seq, but requires DNA marker development and testing of bulked DNA with each marker, both time-consuming and labor-intensive processes that are circumvented by QTL-seq. Consequently, QTL-seq is much more rapidly performed. QTL-seq also allows an accurate quantitative evaluation of the genomic contribution from the two parents to the bulked DNAs by using SNP-index, whereas the conventional method has to rely on analog assessment of marker states, e.g. the relative strength of intensity of DNA amplicons after PCR amplification of the markers. Therefore, we believe that QTL-seq is quicker and has a much higher power than the previous methods used for QTL identification. Applications of whole genome sequencing to two DNA bulks of progeny with extreme phenotypes have been reported in yeast (X-QTL; Ehrenreich *et al.*, 2010; Wenger *et al.*, 2010; Parts *et al.*, 2011; Swinnen *et al.*, 2012), and its statistical property applied to yeast was also addressed (Magwene *et al.*, 2011). QTL-seq has the same principle as these methods. However, QTL-seq is the first application of a similar method in plant species with a much larger genome size (rice, 380 Mb), and we demonstrated that it can be carried out with a significantly smaller number of progeny (20–50) in each bulk than the methods previously reported in yeast. In crop species, bulking of an extremely large number of progeny is not practical.

QTL-seq applied to seedling vigor in rice demonstrated that this method successfully identifies QTL in an F_2 generation, which is a much earlier generation than the F_7 one that we used for conventional QTL analysis based on RILs. Our simulation analysis showed that if the phenotypic effect of the focal QTL accounts for more than 10% of the entire variation and if the read depth is more than or equal to 20 its genomic position may be readily detected by QTL-seq even in the F_2 generation. We also demonstrated that QTL-seq is applicable to progeny obtained from crosses made between genetically closely related cultivars. The rice cultivars Nortai and Hitomebore and Dunghan Shali used in the current experiments all belong to *japonica* species, and DNA polymorphisms among them are low, making DNA marker development difficult in the conventional scheme of QTL analysis.

We envisage that QTL-seq can be applied to any population for detecting genomic regions that underwent artificial or natural selection. For instance, a population of a species is distributed over a certain environmental gradient (high temperature versus low temperature). We could then make two DNA bulks: one from multiple individuals from high temperature zones and the other from low temperature zones. Sequence reads from these two DNA bulks are compared to a reference sequence, and Δ(SNP-index) is calculated for all the genomic regions. The regions showing higher Δ(SNP-index) than the background genome should point to the regions responsible for the adaptation of population to high/low temperature. In this regard, QTL-seq could be perceived as a general method for detecting genomic regions showing signatures of recent selective sweep by whole-genome resequencing of DNAs from two groups of individuals that underwent recent artificial or natural selection in the opposite directions.

In view of the recent rapid development in sequencing technology, we foresee that methods that make use of whole-genome sequencing-based techniques, including QTL-seq, MutMap (Abe *et al.*, 2012a), SHOREmap (Schneeberger *et al.*, 2009), NGM (Austin *et al.*, 2011) and others (Mokry *et al.*, 2011; Trick *et al.*, 2012), will dramatically accelerate crop improvement in a cost-effective manner. These and other related technologies that take full advantages of the rapidly declining cost of genome sequencing are expected to significantly contribute to the on-going efforts aimed at addressing the world food security problem by reducing breeding time.

EXPERIMENTAL PROCEDURES

Evaluation of partial resistance of RILs

To evaluate the partial resistance of RILs to leaf blast disease we conducted upland nursery trials in 2006 and 2011 at Iwate Agricultural Research Center in Kitakami, Iwate, Japan. Overall, a total of four independent inoculation assays were carried out. For each RIL, about 200 seeds were sown in single 40-cm long rows that were spaced at 10 cm apart. For use as inoculum, seedlings of a highly susceptible cultivar Moukoto were grown on both sides of each block of RIL rows. Nitrogen was applied at the rate of 20 kg per 1000 m^2 as a basal fertilizer. Disease severity was visually scored according to the procedure of Asaga (1981).

Whole-genome sequence of bulked DNA

Bulked DNA samples were prepared by mixing an equal ratio of DNA extracted from 100 mg of fresh rice leaves as previously described (Abe *et al.*, 2012a). The library for Illumina sequencing was constructed from 5 μg of DNA sample and sequenced by 76 cycles on an Illumina Genome Analyzer IIx as described in Abe *et al.* (2012a). The short reads in which more than 10% of sequenced nucleotide exhibited a phred quality score of <30 were excluded from the analysis that followed.

Alignment of short reads to the reference sequence and sliding window analysis

To identify the QTL, we aligned the short reads obtained from the two DNA bulks to the reference genome of the cultivar Hitomebore (DDBJ Sequence Read Archive, DRA000809) using BWA software (Li and Durbin, 2009) [Correction added on 13 March 2013 after original online publication: DRA000499 was changed to DRA000809]. Alignment files were converted to SAM/BAM files using SAMtools (Li *et al.*, 2009), and applied to the SNP-calling filter 'Coval' we previously developed (SK *et al.*, in preparation; Abe *et al.*, 2012a) to increase SNP-calling accuracy. SNP-index was calculated for all the SNP positions. We excluded SNP positions with SNP-index of <0.3 and read depth <7 from the two sequences, as these may represent spurious SNPs called due to sequencing and/or alignment errors. However, if SNPs with SNP-index $\geq$ 0.3 were present in only one of the sequences obtained from the two DNA bulks (bulk A), we considered them as real SNPs and assumed their presence in the other bulk (bulk B) too. In this case we used the SNP-index value of the bulk B DNA even if it was <0.3. For positions in the genome where the entire short reads match the reference sequence, we assign a SNP-index of 0. Sliding window analysis was applied to SNP-index plots with 2 Mb window size and 10 kb increment. We calculated the average SNP-index of the SNPs located in the window (*m*) and used it for the sliding window plot. If the number of SNPs within the 2 Mb window was <10, we skipped the interval for the analysis. Use of *m* for sliding window analysis after taking the average of 10 SNP-indices was important to reduce the noise in the plot (Figure S5).

To generate confidence intervals of the SNP-index value under the null hypothesis of no QTL, we carried out computer simulation. We first made two bulks of progeny with a given number of individuals by random sampling. From each bulk, a given number of alleles corresponding to the read depth were sampled. We then calculated SNP-index for each bulk, and derived Δ(SNP-index). This process was repeated 10 000 times for each read depth and confidence intervals were generated (Figure S6). These intervals were plotted for all the genomic regions that have variable read depths.

The SNP genotyping of RILs

We used the Illumina GA IIx sequencer to obtain the Nortai genome sequence, which was compared with the Hitomebore whole-genome sequence to identify SNPs via a filter pipeline. The identified SNPs were applied to the Illumina® Assay Design Tool to design the Oligo Pool Assay (OPA) for the GoldenGate Genotyping Assay (Illumina). The DNA was extracted from 50 mg of fresh rice leaves using the DNeasy 96 Plant Kit (Qiagen, http://www.qiagen.com/) and was quantified using the Quant-iT PicoGreen dsDNA Reagent and Kits (Invitrogen, http://www.invitrogen.com/). The designed OPA and 250 ng of DNA were used for the preparation of bead chips according to the protocol for the GoldenGate Genotyping Assay. The bead chips were scanned by iSCAN and the data were analyzed by GenomeStudio (Illumina).

Computer simulation

In order to obtain the null distribution of *m*, we simulated the RIL construction process according to the single seed descent (SSD). method. We set the genomic parameters to be roughly consistent with rice. That is, the genome size is set to about 360 Mb and the recombination rate at 4 cM Mb^{-1}. We postulated that 150 000 SNPs between the two parents are distributed at equal intervals (i.e. a SNP every 2.4 kb). The number of individuals in a progeny is assumed to be $N = 200$. The breeding process was continued to the F_7 generation, and the QTL-seq process was applied to the F_2 and F_7 generations independently. In the QTL-seq process it is assumed that from among all progeny individuals of F_2 or F_7 generations we select *p*% each of progeny with opposite extreme trait values to make 'Highest' and 'Lowest' bulks. Each bulk is sequenced to depth *n*, so that the SNP data we obtain will be a random set of *n* alleles from *Np* individuals, where replacement is allowed. We simulated 10 000 replications of this process, from which the null distribution of *m* for the F_2 and F_7 generations were obtained. We were also interested in the distribution of *m* in the region encompassing the SNP that is responsible to the focal phenotype. For this purpose, we modified the simulation such that a QTL is placed in the simulated genome. At the QTL, there were two alleles, and the genetic contribution of this QTL relative to the total phenotype variation was given by *Qp*. Although this model includes only one QTL, it does not mean that there is only one QTL in the genome. The remaining contribution with proportion 1 – *Qp* represents all factors including the genetic contributions of other multiple QTLs and environmental variables. With 10 000 replications of the simulations under this simple model, we obtained the distribution of *m* under various parameter sets, from which the power to detect QTLs was computed as the proportion of replications with *m* out of the 99% cutoff values (see above).

ACKNOWLEDGEMENTS

This study was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences (PROBRAIN), the Ministry of Agriculture, Forestry, and Fisheries of Japan (Genomics for Agricultural Innovation PMI-0010) and Grant-in-aid for Scientific Research from the Ministry of Education, Cultures, Sports and Technology, Japan to HK and RT (Grant-in-Aid for Scientific Research on Innovative Areas 23113009) and JSPS KAKENHI to RT (Grant No. 24248004) . We thank Shigeru Kuroda for general support of the work.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version of this article.

Figure S1. Frequency distributions of partial resistance levels of the 241 recombinant inbred lines derived from a cross between Nortai and Hitomebore over four independent trials carried out in 2006 and 2011.

Figure S2. Single nucleotide polymorphism (SNP)-index and Δ(SNP-index) plots for 12 chromosomes of rice bulked DNA.

Figure S3. QTL-seq applied to recombinant inbred lines derived from a cross between the cultivar 'Iwate96' and 'Hitomebore' segregating in grain amylose content.

Figure S4. QTL-seq applied to recombinant inbred lines derived from a cross between the cultivar 'Arroz da Terra' and 'Iwatekko' segregating in germination rate at low temperature condition.

Figure S5. Flow chart of QTL-seq analysis.

Figure S6. Simulation test for deriving confidence intervals of Δ (SNP-index) under the null hypothesis (no quantitative trait locus).

Table S1. Summary of Illumina GAIIx sequencing for Nortai × Hitomebore recombinant inbred lines and Hitomebore × Dunghan Shali F_2.

Table S2. Number of single nucleotide polymorphisms detected between Hitomebore and 13 rice cultivars.

REFERENCES

Abe, A., Kosugi, S., Yoshida, K. *et al.* (2012a) Genome sequencing reveals agronomically-important loci in rice from mutant populations. *Nat. Biotechnol.* **30**, 174–178.

Abe, A., Takagi, H., Fujibe, T., Aya, K., Kojima, M., Sakakibara, H., Uemura, A., Matsuoka, M. and Terauchi, R. (2012b) OsGA20ox1, a candidate gene for a major QTL controlling seedling vigor in rice. *Theor. Appl. Genet.* **125**, 647–657.

Asaga, K. (1981) A procedure for evaluating field resistance to blast in rice varieties. *J. Cent. Agric. Exp. Stn.* **35**, 51–138.

Ashikari, M. and Matsuoka, M. (2006) Identification, isolation and pyramiding of quantitative trait loci for rice breeding. *Trends Plant Sci.* **11**, 344–350.

Austin, R.S., Vidaurre, D., Stamatiou, G. *et al.* (2011) Next-generation mapping of Arabidopsis genes. *Plant J.* **67**, 715–725.

Basten, C.J., Weir, B.S. and Zeng, Z.B. (2005) *QTL Cartographer, Version 1.17: A Reference Manual and Tutorial for QTL Mapping*. Raleigh: North Carolina State University.

Darvasi, A. and Soller, M. (1994) Selective DNA pooling for determination of linkage between a molecular marker and a quantitative trait locus. *Genetics*, **138**, 1365–1373.

David, T., Christian, B., Jason, H. and Belinda, L.B. (2011) Global food demand and the sustainable intensification of agriculture. *Proc. Natl Acad. Sci. USA*, **108**, 20260–20264.

Ehrenreich, M.I., Torabi, N., Jia, Y., Kent, J., Martis, S., Shapiro, A.J., Gresham, D., Caudy, A.A. and Kruglyak, L. (2010) Dissection of genetically complex traits with extremely large pools of yeast segregants. *Nature*, **446**, 1039–1042.

Falconer, D.S. and Mackay, T.F.C. (1996) *Introduction to Quantitative Genetics*, 4th edn. London: Prentice Hall.

Fujino, K., Sekiguchi, H., Matsuda, Y., Sugimoto, K., Ono, K. and Yano, M. (2008) Molecular identification of a major quantitative trait locus, qLTG3–1, controlling low-temperature germinability in rice. *Proc. Natl Acad. Sci. USA*, **105**, 12623–12628.

Giovannoni, J.J., Wing, R.A., Ganal, M.W. and Tanksley, S.D. (1991) Isolation of molecular markers from specific chromosome intervals using DNA pools from existing mapping populations. *Nucleic Acids Res.* **19**, 6553–6558.

Godfray, H.C.J. (2010) Food security: The challenge of feeding 9 billion people. *Science*, **327**, 812–818.

Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R.; 1000.Genome.Project.Data.Processing.Subgroup. (2009) The sequence alignment/map format and samtools. *Bioinformatics*, **25**, 2078–2079.

Magwene, P.M., Willis, J.H. and Kelly, J.K. (2011) The statistics of bulk segregant analysis using next generation sequencing. *PLoS Comput. Biol.* **7**, e1002255.

Mansur, L.M., Orf, J. and Lark, K.G. (1993) Determining the linkage of quantitative trait loci to RFLP markers using extreme phenotypes of recombinant inbreds of soybean (Glycine max L. Merr.). *Theor. Appl. Genet.* **86**, 914–918.

Michelmore, R.W., Paran, I. and Kesseli, R.V. (1991) Identification of markers linked to disease resistance genes by bulked segregant analysis: A rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl Acad. Sci. USA*, **88**, 9828–9832.

Miura, K., Lin, S.Y., Yano, M. and Nagamine, T. (2001) Mapping Quantitative Trait Loci Controlling Low Temperature Germinability in Rice(Oryza sativa L.). *Breed. Sci.* **51**, 293–299.

Mokry, M., Nijman, I.J., van Dijken, A., Benjamins, R., Heidstra, R., Scheres, B. and Cuppen, E. (2011) Identification of factors required for meristem function in Arabidopsis using a novel next generation sequencing fast forward genetics approach. *BMC Genomics*, **12**, 256.

Parts, L., Cubillos, A.F., Warringer, J. *et al.* (2011) Revealing the genetic structure of a trait by sequencing a population under selection. *Genome Res.* **21**, 1131–1138.

Schneeberger, K., Ossowski, S., Lanz, C., Juul, T., Petersen, A.H., Nielsen, K.L., Jorgensen, J.-E., Weigel, D. and Andersen, S.U. (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods*, **6**, 550–551.

Swinnen, S., Schaerlaekens, K., Pais, T. *et al.* (2012) Identification of novel causative genes determining the complex trait of high ethanol tolerance in yeast using pooled-segregant whole-genome sequence analysis. *Genome Res.* **22**, 975–984.

Trick, M., Adamski, N.M., Mugford, S.G., Jiang, C.-C., Febrer, M. and Uauy, C. (2012) Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploidy wheat. *BMC Plant Biol.* **12**, 14.

Wenger, W.J., Schwartz, K. and Sherlock, G. (2010) Bulk segregant analysis by high-throughput sequencing reveals a novel xylose utilization gene from *Saccharomyces cerevisiae*. *PLoS Genet.* **6**, e1000942.

Williams, J.G.K., Kubelik, A.R., Livak, K.J., Rafalski, J.A. and Tingey, S.V. (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. *Nucleic Acids Res.* **18**, 6531–6535.

Yano, M. (2001) Genetic and molecular dissection of naturally occurring variation. *Curr. Opin. Plant Biol.* **4**, 130–135.

Yano, K., Takashi, T., Nagamatsu, S., Kojima, M., Sakakibara, H., Kitano, H., Matsuoka, M. and Aya, K. (2012) Efficacy of microarray profiling data combined with QTL mapping for the identification of a QTL gene controlling the initial growth rate in rice. *Plant Cell Physiol.* **53**, 729–739.