Next generation sequencing (NGS)-based methods for simultaneous mapping and identification of candidate gene and QTLs in rice

(イネにおける次世代シーケンサーを用いた遺伝子およびQTL同定法の確立)

2014.3
岩手大学大学院
連合農学研究科
寒冷圏生命システム学専攻
(岩手大学)

高木 宏樹
Table of contents

Chapter 1: General introduction 1
Chapter 2: QTL-seq 7
  Summary 8
  Introduction 9
  Materials and Methods 12
  Results 18
  Discussion 43
Chapter 3: MutMap-Gap 47
  Summary 48
  Introduction 49
  Materials and Methods 50
  Results 54
  Discussion 75
Chapter 4: MutMap+ 77
  Summary 78
  Introduction 79
  Materials and Methods 82
  Results 84
  Discussion 104
Chapter 5: General discussion 107
Chapter 6: General conclusion 115
References 119
Acknowledgments 128
Chapter 1

General Introduction
The world population exceeded 7 billion in 2012, and is projected to continue growing in the future. In order to meet the demand of this ever-increasing population, more than 70% increase in food production relative to current levels is required by 2050 (Tester and Langridge, 2010). To this end, there is a greater demand for accelerated breeding of crops to achieve sustainable yield increases without further expanding farmlands and damaging the environment (Godfray, 2010; David et al., 2011).

One of the most efficient methods in crop breeding is marker-assisted selection (MAS) (Ashikari and Matsuoka, 2006). In MAS, once a gene or quantitative trait locus (QTL) controlling a favorable trait is mapped with closely linked DNA markers, the genomic interval harboring the gene/QTL 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 gene/QTL are selected using the tightly linked DNA markers. Although MAS is a very useful method to reduce the effort and time needed for phenotypic evaluation of successive progeny, the identification of gene/QTLs of interest and the development of DNA markers tightly linked to such genomic regions are important prerequisites. Therefore, the rapid identification of a gene or QTL of interest is a challenging task for any endeavor that attempts to accelerate crop breeding via MAS.

Recent advances in next generation sequencing (NGS)-based methods such as SHOREmap (Schneeberger et al., 2009), NGM (Austin et al., 2011) and the MutMap method developed by our group (Abe et al. 2012a) have accelerated the isolation of candidate genes based on artificially induced mutations. The common feature of these methods is that they primarily target those traits that are controlled by single gene. Nevertheless, the majority of agronomically important traits in crops such as yield, as well as tolerance to abiotic and biotic stresses are controlled by multiple genes each with a relatively minor effect. These genes are called quantitative trait loci (QTLs) (Falconer
and Mackay, 1996). Accordingly, crop breeding mostly depends on genetic variations occurring in nature in landraces and wild crop relatives (Yano, 2001). This is partly because naturally occurring variants harbor potentially larger repertoire of useful alleles than the artificially generated mutants due to the larger number of mutations accumulated in nature over a longer time. Therefore, the analysis of QTL variations among natural variants is important for enhancing breeding. Previously, there was no report of an NGS-based method that is effective for QTL mapping and thus reduces the time required for QTL analysis. This PhD thesis primarily reports the development and application of an NGS-based method for QTL mapping using rice as an example.

Traditionally, QTLs have been identified by linkage analysis of progeny derived from a cross between parents showing contrasting phenotypes for a trait of interest using DNA markers capable of discriminating the parental genomes. This demand for polymorphic DNA markers is a major limiting factor in the application of QTL analysis. When progeny derived from a cross between genetically distant cultivars is used for linkage analysis, developing DNA markers that can discriminate the mapping parents is a relatively straightforward procedure. Nevertheless, phenotyping the progeny is often difficult because expression of the phenotype of interest could be confounded by additional natural variations present between the two parents. On the other hand, crossing closely related parents makes the scoring of the resulting progeny easier, while the development of sufficient DNA markers for linkage analysis becomes the limiting factor, making genotyping of the mapping progeny with each DNA marker time-consuming and costly. This requirement for sufficient DNA markers for the analysis has been the major reason behind the relatively longer time associated with QTL mapping and identification of candidate genes.

Using whole genome sequencing of two bulked samples obtained from progeny showing extreme values for the phenotype under consideration in a
segregating population, we have developed a novel WGS-based method for rapid identification of QTLs that we named QTL-seq. QTL-seq circumvents the requirement for developing DNA markers for linkage analysis, significantly reducing the time needed for QTL mapping. I confirmed the versatility of QTL-seq by applying it to rice recombinant inbred lines and F2 populations, and successfully identified QTLs for important agronomic traits, such as partial resistance to the fungal rice blast disease and seedling vigor. Details of the QTL-seq method are discussed in Chapter 2.

The MutMap method we recently developed has been powerful for the identification of genes responsible for agronomically important traits in rice (Abe et al., 2012a). For MutMap application, the mutant showing a phenotype of interest is first crossed to the parental line that was used for the original mutagenesis, and the resulting F2 progeny are scored for the phenotype. Then, DNA from 20-40 F2 progeny showing the mutant phenotype is bulk sequenced, and the resulting short reads are aligned to the “reference sequence” of the parental line. Following the identification of SNPs between the mutant-type bulk and parental line sequences, SNP-index is calculated at each SNP position. In MutMap, SNP-index is defined as the ratio of mutant-type short reads to the total short reads covering a particular SNP position (Abe et al., 2012a). SNP-indices are graphically plotted for all chromosomes over the entire genome, and the genomic interval with a unique peak of SNP-index = 1 represents the region harboring the causative mutation responsible for the phenotype being considered.

We have so far successfully applied MutMap in rice and identified novel genes involved in the control of important agronomic traits, and the method is now a routine protocol in our lab. Unequivocal scoring of phenotypes in segregating F2 progeny even when the phenotypic changes are only subtle is a major advantage of MutMap, while the method also allows the utilization of SNPs incorporated by EMS mutagenesis as DNA markers for simultaneous
mapping and isolation of the gene of interest. However, there are the cases in which MutMap cannot be directly applied for mapping and identification of gene. Therefore, an additional component of this PhD sturdy has been the development of NGS-based methods that improve the original MutMap protocol in order to make it more versatile and expand the conditions under which it can be applied for mapping and isolation of candidate genes. Accordingly, we have developed two novel NGS-based methods that address the limitations of the original MutMap method.

The first novel method that has been developed is MutMap-Gap. The availability of a parental line “reference sequence” is an important prerequisite for the application of MutMap. In the original protocol, these “reference sequences” are developed based on the publically available highly accurate reference genomes of representative cultivars/lines of the species of interest; e.g. Nipponbare cultivar in rice (Oryza sativa) and Col ecotype in Arabidopsis thaliana. To generate a “reference sequence” for the parental line being studied, NGS short reads of the parental line are first aligned to the reference genome, and then the SNPs available between the two sequences are identified. This is followed by replacing the reference genome specific SNPs with those SNPs specific to the parental line sequence to reconstitute “reference sequence” of the latter.

In the process of generating the parental line “reference sequence”, the genomic regions displaying significant structural variation between the parental line and the cultivar originally used for developing the reference genome are not taken in to consideration. Consequently, MutMap cannot identify causative mutations that are located within the parental line-specific genomic regions. MutMap-Gap is a powerful method that has been developed to address this limitation of MutMap, and combines MutMap analysis and de novo assembly of gap regions to identify genes of interest located in genomic regions not represented by reference genomes. In Chapter 3, details of the MutMap-Gap
protocol are provided using an example from its application in the isolation of the rice resistance gene \( Pii \).

The second novel method reported here is \textbf{MutMap+}. The application of MutMap, as well as the newly developed MutMap-Gap, relies on the availability of segregating progeny resulting from the cross between a mutant of interest and the parental line used for mutagenesis. This prerequisite of a genetic cross limits the utility of MutMap to address mutations leading to early development lethality, sterility, or generally hamper crossing, as well as restricts its application in crop species where artificial crossing is difficult. In Chapter 4, I describe the MutMap+ method that relies on the sequencing of two bulked samples obtained from the segregating M3 population established from a heterozygous M2 individual, and thus circumvents the requirement for crossing.

I believe that NGS-based methods of QTL-seq, MutMap-Gap, and MutMap+ reported here, together with the original MutMap method developed by our group, will open the way for the rapid identification of agronomically important genes and QTLs, thereby contributing enormously to accelerated breeding of crops.
Chapter 2
QTL-seq
Rapid mapping of quantitative trait loci in rice by whole genome resequencing of two bulked DNA

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: QTL). QTL mapping and isolation is important for efficient crop breeding by marker-assisted selection (MAS) and for a better understanding of the molecular mechanisms underlying the traits. Since it requires the development and selection of DNA markers for linkage analysis, QTL analysis is however time consuming and labor intensive. Here we report a method for rapid identification of plant QTL by whole genome resequencing of two bulked DNAs of 20-50 individuals showing extreme opposite trait values for a given phenotype in a segregating progeny, which we named QTL-seq. We applied QTL-seq to rice recombinant inbred lines (RILs) and F2 populations and successfully identified QTL 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 QTL 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.

Introduction

The world population has already exceeded 7 billions and still continues to grow, while the 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 (QTL) (Falconer and Mackay, 1996). Identification of QTL is an important task in plant breeding. Once the 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). MAS reduces the effort and time needed for phenotype evaluation of the progeny during successive rounds of selection, as well as improves introgression breeding.

Traditionally QTL 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 QTL 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 F2 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, bulked DNAs of F2 progeny were screened with restriction fragment polymorphisms (RFLPs) and random amplified polymorphic DNAs (RAPDs) markers to identify the markers linked to the traits of interest. Later, bulked-segregant analysis 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 DNA is still time-consuming and costly.

Recent development of whole genome sequencing has harnessing the analysis of QTLs in yeast, a model organism with a relatively small genome size (12.5Mb). 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 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 QTL 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 bulked DNAs of progeny (each 20 ~ 50 individuals) showing extreme phenotypic values by next-generation sequencing (NGS) technology. Since this approach has a profound applicability in plant species including crops, we name the method QTL-seq as 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 QTL.
Materials and Methods

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. Overall, a total of four independent inoculation assays were carried out. For each RIL, about 200 seeds were sown in a 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/1000 m2 as a basal fertilizer. Disease severity was visually scored according to the procedure of Asaga (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 five micrograms 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 was less than 30 of phred quality score were excluded from the analysis that followed.

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

To identify the QTL, we aligned the short reads from bulked DNA to the reference genome of the cultivar Hitomebore (DDBJ Sequence Read Archive: DRA000499) using BWA software (Li and Durbin, 2009). Alignment files were converted to SAM/BAM files using SAMtools (Li et al., 2009), and applied to the filter to increase accuracy as described by Abe et al. (2012a). SNP-index was calculated for all the SNP positions. We excluded SNPs
position at which both bulked DNA had SNP-index of < 0.3 and read depth was less than 7, as these may represent spurious SNPs called due to sequencing and/or alignment errors. However, if SNPs with SNP-index ≥ 0.3 were present in only one (bulk A) of the two bulk DNAs (bulks A and B), we considered them as real SNPs and assumed their presence in the other bulk DNA (bulk B). In this case we used the SNP-index value of the bulk B DNA even if it is less than 0.3. For positions in the genome where the entire short reads match the reference sequence, we assign SNP-index = 0. Sliding widow analysis was applied to SNP-index plots with 2 Mb window size and 10 kb increment. We calculated average SNP-index of the SNPs located in the window \( m \) and used it for sliding window plot. If the number of SNPs within 2 Mb window was less than 10, we skipped the interval for the analysis. Use of \( m \) for sliding window analysis after taking average of 10 SNP-indices was important to reduce the noise in the plot (Figure 2-1).

To generate confidence intervals of SNP-index value under the null hypothesis of no QTL, we carried out computer simulation. We first made two pools of progeny with given number of individuals by random sampling. From each pool, a given number of alleles corresponding to the read depth were sampled. We then calculated SNP-index and \( \Delta \) (SNP-index) calculated for each pool. This process was repeated 10,000 times and confidence intervals were generated (Figure 2-2). These intervals were plotted for all the genomic regions that have variable read depths.

**SNP genotyping of RILs**

We applied illumina GA IIx sequencer to obtain Nortai genome sequence, which was compared with Hitomebore whole genome sequence to identify SNPs via a filter pipeline. The identified SNPs were applied to illumina® Assay Design Tool to design the OPA for the GoldenGate Genotypeing Assay (illumina, San Diego, CA, USA). DNA was extracted from 50 mg of fresh rice
leaves using the DNeasy 96 Plant Kit (QIAGEN Sciences, Maryland, USA) and was quantified using the Quant-iT PicoGreen dsDNA Reagent and Kits (Invitrogen, Oregon, USA). The designed OPA and 250 nanograms of the DNA were used for 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, San Diego, CA, USA).

**Computer simulation**

In order to obtain the null distribution of \( m \), we simulated the RILs construction process according to the SSD method. We set the genomic parameters to be roughly consistent with rice. That is, the genome size is set at about 360 Mb and the recombination rate at 4 cM/Mb. 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 F7 generations, and the QTL-seq process was applied to the F2 and F7 generations independently. In the QTL-seq process, it is assumed that from among all progeny individuals of F2 or F7 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 F2 and F7 generations were obtained. We are 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 above simulation such that a QTL is placed in the simulated genome. At the QTL, there are two alleles, and the genetic contribution of this QTL relative to the total phenotype variation is 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 factors. 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 QTL was computed as the proportion of replications with \( m \) out of the 99% cutoff values (see above).
Figure 2-1. Flow chart of QTL-seq.
Red box indicate filtering point of illumina short reads. yellow box indicate filtering point following calculation of SNP-index.
Figure 2-2. Flow chart of Simulation test for null distribution in QTL-seq.

**Simulation for random sampling**

We assume the progenies were obtained by crossing two cultivars whose genotypes were represented by AA and BB, respectively. The number of individuals in both progeny is equal and corresponded to the QTL-seq samples presented in the main text. The genotype of the selected individuals is decided at a random position in the genome. The type of progeny is decided as either RILs or F2 as per the examples presented in the main text.

**Simulation for calculating Δ(SNP-index)**

SNP-index is simulated by the probability distribution of SNP-index, which corresponded to the binomial distribution of B (given read depth, B allele frequency). Δ(SNP-index) is calculated by subtraction of Progeny 2 SNP-index from Progeny 1 SNP-index.

**Calculation of null distribution**

The 90, 95 and 99% confidence intervals of Δ(SNP-index) are calculated based on the result obtained from 10,000 times replication. We use this confidential interval as null distribution of the non-selected bulked DNA.
Results

Principle of QTL-seq

The principle of QTL-seq is shown in Figure 2-3 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 as well as 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 2-3, we assume that we are interested in plant height and that cultivars A and B are low and high in 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, thus are suitable for detecting QTL of minor effects. The advantage of using an F2 population is the 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 QTL.

After the progeny of a mapping population are measured for the focused trait, we score segregation of the phenotype. If the number of QTL involved in the trait variation is multiple, frequency distribution of measured values will be close to the Normal (Gaussian) distribution (Figure 2-3a). 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 10-20 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 >10 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.

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 parents. After alignment of 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 read \( n \) covering a particular genomic position \( = k/n \) as SNP-index \( (Abe \ et \ al., \ 2012a) \). SNP-index is 0 if the entire short reads contain genomic fragments from the parent that was used as a reference sequence. SNP-index is 1 if all the short reads represent the genome from the other parent. SNP-index of 0.5 means equal contribution of both parents’ genomes to the bulked progeny. Accordingly, 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 \( (Figure \ 2-3b) \). We carried out this procedure separately for the “Highest” and “Lowest” bulk DNAs.

In practice, SNPs with SNP-index < 0.3 in both bulked DNAs are filtered out during SNP calling because they cannot be discriminated from spurious SNPs caused by sequencing or alignment errors. However, if SNPs with SNP-index 0.3 or greater are present in only one of the two bulked DNAs, we consider them as real SNPs and assume their presence in the other bulked DNA. In this case, we make use of the SNP-index value of the other bulked DNA even if it is less than 0.3 \( (see \ Materials \ and \ Methods) \). 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 QTL 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 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 2-3b). In this graph, Δ(SNP-index) = 1 if the bulked DNA comprises only of 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.
Figure 2-3. A simplified scheme of QTL-seq as applied to rice.
(a) Two inbred cultivars with contrasting phenotypes are crossed to generate F2 progeny that are segregating for the trait value. In this example, parent A has low stature while parent B has high stature. Since multiple QTL control plant height, frequency of plant height among the F2 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 are applied to whole genome resequencing and aligned to the reference sequence of cultivar A to calculate SNP-index. (b) Examples of SNP-index plot. QTL can be identified as peaks or valleys of SNP-index plot. Each spot corresponds to a SNP, and x-axis corresponds to 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).
**QTL-seq applied to RILs: Detection of QTL controlling partial resistance to rice blast in Nortai**

We applied QTL-seq for the detection of QTL 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 F2 (Figure 2-4). Each F2 progeny was established as a line and brought to F7 generation by 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 the seven classes from class 4 (resistant) to class 10 (highly susceptible; Figure 2-5). The inoculation assay was conducted four times over four years to ensure the correct scoring of each RIL. The average score of Nortai and Hitomebore in four trials was 4.75 and 6.38, respectively. The frequency distributions of RILs falling into different classes are close to Normal distribution, suggesting that multiple genes control the partial resistance of RILs. 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 (class 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.

![Figure 2-4. Phenotype of two rice cultivars Nortai and Hitomebore two weeks after inoculation with a compatible race (race 037.1) of blast fungus. Nortai shows partial resistance whereas Hitomebore is susceptible.](image)
Figure 2-5. Partial resistance levels of 241 RILs.
(a) Scores (4: highly resistant, ~ 10: highly susceptible) assigned to different levels of partial resistance in RILs derived from a cross between Nortai and Hitomebore.
(b) Frequency distributions of partial resistance levels of the 241 RILs derived from a cross between Nortai and Hitomebore over four independent trials carried out in 2006 and 2011. x-axis: resistance level (4: highly resistant; 10: highly susceptible). y-axis: number of lines falling into each level. N and H indicate the average value of Nortai and Hitomebore, respectively.
Each Bulked DNA 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 75bp) from the bulked DNA of R-progeny and S-progeny, respectively (Table 2-1). These reads were aligned to the reference sequence of Hitomebore cultivar using the BWA software (Li and Durbin, 2009). The average read depth was >6.88x in both bulked DNA (Table 2-1). A total of 161,563 SNPs were identified between Nortai and Hitomebore genomes (Table 2-2), and SNP-index was obtained for each SNP. Graphs showing relationships between SNP-index and genomic positions are given in Figure 2-6 and Figure 2-7. We found highly contrasting patterns of SNP-index graphs for R-bulk and S-bulk in the region between 2.39 Mb and 4.39 Mb on chromosome 6 as shown in Figure 2-6. The resistant RILs tend to have Nortai-type genomic segments in the 2.39 to 4.39 Mb region of chromosome 6, whereas susceptible RILs tend to have 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 of 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) (Figure 2-6 and Figure 2-7). This revealed that most part 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 QTL, and plotted them along with Δ(SNP-index) (Materials and Methods: Figure 2-2). The chance that Δ(SNP-index) becomes higher than 0.79 is $P < 0.01$ under the null hypothesis for the chromosomal region of 2.39-4.39Mb.
Table 2-1.
Summary of Illumina GAIIx sequencing for Nortai x Hitomebore RILs and Hitomebore x Dunghan Shali F2.*

<table>
<thead>
<tr>
<th>Type of mapping population</th>
<th>Pollen parent</th>
<th>Seed parent</th>
<th>Sample</th>
<th>Number of bulked lines</th>
<th>Number of reads</th>
<th>Genomic coverage (%) a</th>
<th>Mean depth b</th>
<th>Accession number in DDBJ</th>
</tr>
</thead>
<tbody>
<tr>
<td>RILs</td>
<td>Nortai</td>
<td>Hitomebore</td>
<td>R-bulk</td>
<td>20</td>
<td>57,901,696</td>
<td>91.73</td>
<td>6.93</td>
<td>DRA000809</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>S-bulk</td>
<td>20</td>
<td>62,378,836</td>
<td>89.69</td>
<td>6.88</td>
<td>DRA000809</td>
</tr>
<tr>
<td>F2</td>
<td>Hitomebore</td>
<td>Dunghan Shali</td>
<td>H-bulk</td>
<td>50</td>
<td>15,054,688</td>
<td>96.45</td>
<td>19.35</td>
<td>DRA000809</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>L-bulk</td>
<td>50</td>
<td>15,522,748</td>
<td>96.85</td>
<td>19.54</td>
<td>DRA000809</td>
</tr>
</tbody>
</table>

* Sequence reads of each sample were aligned to the Hitomebore consensus sequence as described in the Material and Methods.

a Percentage of total genomic region of Hitomebore reference sequence aligned by shot reads.
b Average of read depth in the whole genome.

Figure 2-6.
QTL-seq applied to rice RILs identifies QTL conferring partial blast resistance to Nortai cultivar.
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 null hypothesis of no QTL (gray: P < 0.1; green: P < 0.05; pink: P < 0.01) and LOD score plot of partial resistance QTL as obtained by classical QTL analysis of 241 RILs (bottom).
Figure 2-7.
Figure 2-7.
Figure 2-7.
**Figure 2-7.**
QTL-seq results of RILs derived from a cross between Hitomebore and Nortai.
Red lines indicate the sliding window average of 2Mb interval with 10 kb increment for SNP-index and Δ(SNP-index).
(a) SNP-index for bulked DNA of the RILs showing higher level of partial resistance (R-bulk) derived from a cross between Nortai and Hitomebore. (b) The SNP-index for bulked DNA of the RILs showing low level of partial resistance (S-bulk) derived from a cross between Nortai and Hitomebore. (c) The Δ(SNP index) plot obtained by subtraction of S-bulk SNP-index from R-bulk SNP-index for RILs obtained from a cross between Nortai and Hitomebore.

**Table 2-2.**
Number of reliable SNPs detected in the 13 rice cultivar with reference to the Hitomebore sequence.

<table>
<thead>
<tr>
<th>Cultivar name</th>
<th>Number of SNPs</th>
<th>spp.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dunghan Shali</td>
<td>96,543</td>
<td><em>japonica</em></td>
</tr>
<tr>
<td>Nipponbare</td>
<td>101,330</td>
<td><em>japonica</em></td>
</tr>
<tr>
<td>Nortai</td>
<td>161,563</td>
<td><em>japonica</em></td>
</tr>
<tr>
<td>Jaguary</td>
<td>474,026</td>
<td>Tropical <em>japonica</em></td>
</tr>
<tr>
<td>Urasan1</td>
<td>587,615</td>
<td>Tropical <em>japonica</em></td>
</tr>
<tr>
<td>Rexmont</td>
<td>700,485</td>
<td>Tropical <em>japonica</em></td>
</tr>
<tr>
<td>TUPA729</td>
<td>1,045,537</td>
<td>Tropical <em>japonica</em></td>
</tr>
<tr>
<td>TUPA121-3</td>
<td>1,451,016</td>
<td><em>indica</em></td>
</tr>
<tr>
<td>Kasalath</td>
<td>1,558,719</td>
<td><em>indica</em></td>
</tr>
<tr>
<td>Kaluheenatl</td>
<td>1,661,984</td>
<td><em>indica</em></td>
</tr>
<tr>
<td>Surjamukhi</td>
<td>1,714,339</td>
<td><em>indica</em></td>
</tr>
<tr>
<td>Ratul</td>
<td>1,768,420</td>
<td><em>indica</em></td>
</tr>
<tr>
<td>Badaridhan</td>
<td>1,862,230</td>
<td><em>indica</em></td>
</tr>
</tbody>
</table>

1 Sequence reads of each sample were aligned to the Hitomebore consensus sequence as described in the Material and Methods.
2 The number of SNPs defined as genome positions with ≥80% of the reads covering the positions containing SNP different from the reference nucleotide. Sequence reads were filtered by Coval filter (Kosugi et al., 2013) for selection of high quality read.
3 Classification of cultivar types according to: http://www.gene.affrc.go.jp/databases-core_collections_wr_en.php
To verify the candidate QTL detected by 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, San Diego, CA, USA). The data was 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 2-6). The highest LOD score (LOD = 7.38) was observed in the interval of SNP makers located at 0 Mb and 4.9 Mb. This interval corresponded to the genomic region identified by QTL-seq method (Figure 2-6). This result demonstrates that QTL-seq allows the rapid detection of QTL 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 successfully identified the peaks of $\Delta$(SNP-index) plot presumably corresponding to a major QTL: a $\Delta$(SNP-index) peak for lower grain amylose content in the cultivar Iwate96 as compared to Hitomebore (Figure 2-8). We also identified four $\Delta$(SNP-index) peaks for enhanced seedling vigor under low temperature condition of a cultivar Arroz da Terra as compared to Iwatekko (Figure 2-9). Of the four $\Delta$(SNP-index) peaks detected for seedling vigor, three corresponded to QTL previously identified by conventional QTL analysis ($qlTG3\text{-}2$ and $qlTG11$ (Fujino et al., 2008); $qlTG3\text{-}1$ (Miura et al., 2001)). The three $\Delta$(SNP-index) peaks were statistically significant (the peak for $qlTG3\text{-}1$, $P < 0.01$; and the peaks for $qlTG3\text{-}2$ and $qlTG11$, $P < 0.05$)
Frequency distribution of amylose content in 160 F4 progeny derived from a cross between Iwate96 and Hitomebore. Both cultivars belong to spp. japonica. We selected 15 each individual showing lower grain amylose content (L-bulk) and higher grain amylose content (H-bulk), and applied to QTL-seq using Hitomebore reference sequence. Grain amylose content was measured by Auto analyzer (BL TEC K.K.).

Figure 2-8.
QTL-seq applied to RILs derived from a cross between the cultivar “Iwate96” and “Hitomebore” segregating in grain amylose content.

Red arrow head indicate the position of QTL that we detected by QTL-seq.
Frequency distribution of germination rate of 200 F7 progeny derived from a cross between Iwatekko (I) and Arroz da Terra (A) after 8 days of absorption of water at 13°C.

Pink-color shaded regions correspond to QTL detected by QTL-seq in the current study. Black arrow head indicates the position of cloned gene (qLTG3-1 (10)) White arrow heads indicate the positions of reported QTL (11).

Figure 2-9. QTL-seq applied to RILs derived from a cross between the cultivar “Arroz da Terra” and “Iwatekko” segregating in germination rate at low temperature condition.
Application of QTL-seq to F2 progeny

We further examined the possibility of applying QTL-seq to F2 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 2-10). We have recently fine-mapped a major QTL, \textit{qPHS3-2}, on chromosome 3 that confers the seedling vigor in Dunghan Shali using conventional QTL analysis of RILs of F7 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 \textit{OsGA20ox1}, a gene involved in gibberellin (GA) byosynthesis (Abe et al., 2012b; Yano et al., 2012). Using Dunghan Shali, we addressed whether QTL-seq can detect \textit{qPHS3-2} in the F2 progeny derived from a cross between Dunghan Shali and Hitomebore. After crossing Dunghan Shali to Hitomebore, we obtained F2 progeny. Selfed seeds of a total of 531 F2 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 Normal distribution, indicative of multiple genes involvement in determining this character (Figure 2-10b). Two bulked DNAs were prepared; 50 individuals showing taller seedling height as “H-bulk” and 50 individuals with shorter plant height as “L-bulk”, and were used for QTL-seq analysis (Figure 2-11 and Figure 2-12). By examining Δ(SNP-index) plot, we identified two genomic positions exhibiting the highest Δ(SNP-index) values: chromosome 3 region from 36.21 Mb to 37.31 Mb with Δ(SNP-index) = 0.61 (Statistical significance under the null hypothesis: \( P < 0.01 \)) and chromosome 1 region from 39.08 Mb to 41.08 Mb with Δ(SNP-index) = 0.67 (\( P < 0.05 \)). This former position corresponds exactly to the reported \textit{qPHS3-2}, most probably the locus of \textit{OsGA20ox1}. Likewise, the latter position was also previously detected as a minor QTL (\textit{qPHS-1}) (Abe et al., 2012b). This result demonstrates that QTL identified by conventional QTL mapping using RILs of F7 generations could be successfully recovered by QTL-seq using F2
Figure 2-10.
Seedlings height of Hitomebore, Duanghan Shali and F2 progeny derived from crossing between Hitomebore and Duanghan Shali.
(a) Seedlings of Hitomebore and Duanghan Shali 10 days after water imbibition. Duanghan Shali shows higher seedling vigor as compared to Hitomebore. (b) Frequency distribution of seedling height in 531 F2 progenies 14 days after water imbibition. H and D indicate the average seedling height of Hitomebore and Duanghan Shali, respectively. We selected 50 F2 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 Hitomebore reference genome sequence.
Figure 2-11.
The result of QTL-seq applied to rice F2 progeny derived from a cross between Hitomebore and Dunghan Shali for identification of QTL involved in seedling vigor. Results of QTL-seq for chromosome 3 (left) and 1 (right). Δ(SNP-index) plot (top) with statistical confidence intervals under null hypothesis of no QTL (gray: $P < 0.1$; green: $P < 0.05$; pink: $P < 0.01$) and LOD score plot of QTL controlling plant height as obtained by classical QTL analysis of 250 RILs of F7 generation (bottom).
Figure 2-12
Figure 2-12
Figure 2-12
Figure 2-12.
QTL-seq results of F2 F2 progeny derived from a crosses between Hitomebore and Dunghan Shali.
Red lines indicate the sliding window average of 2Mb interval with 10 kb increment for SNP-index and Δ(SNP-index).
(a) SNP-index plot for bulked DNA of F2 showing larger seedling height (H-bulk) derived from a cross between Hitomebore and Dunghan Shali. (b) SNP-index plot for bulked DNA of F2 showing lower seedling height (L-bulk) derived from a cross between Hitomebore and Dunghan Shali. (c) The Δ(SNP index) plot obtained by subtraction of L-bulk SNP-index from H-bulk SNP-index for F2 obtained from a cross between Hitomebore and Dunghan Shali.
**Simulation of QTL-seq**

As shown above, QTL-seq successfully identified genomic regions controlling quantitative traits in the examples of rice RIL and F2 families. More generally, we are interested in how experimental variables affect the performance of QTL-seq to faithfully detect QTL. To this end, we carried out a computer simulation of QTL-seq by changing variables like (i) contribution of QTL on phenotypic variation, (ii) percentage of individuals to be selected, (iii) read depth, (iv) 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. 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 F2 or F7 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 \( \Delta(\text{SNP-index}) \). Since we routinely take an average of \( \Delta(\text{SNP-index}) \) 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 F2 depends on the percentages of individuals in the bulk \((p)\) and read depth of the focused region (Figure 2-13). The intervals of values \( m \) 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 F7 generation (Figure 2-13). The null distribution of \( |m| \) is wider than that for an F2 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 are allowed at the focal QTL, and for the dominance effect, we consider two cases, codominance and complete dominance. For \(Qp\), 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 2-13 and 4c show 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 to the 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 it increases, the power increases, but it starts decreasing when the sample size is so large that many individuals with intermediate phenotype are included in the bulking. We found higher power in F7 RILs than in F2 populations, and thus conclude that in QTL-seq application to the F2 populations, \(p=0.15\) and \(n=20\) would be a reasonable choice, and that the method have reasonable power to detect QTL with a relative contribution of roughly 10% \((Qp=0.1)\).
Figure 2-13. Simulation reveals capability of QTL-seq for detecting QTL in wide range of values of experimental variables
(a) 99% intervals of the null distribution of $m$ statistics (average value of $\Delta$(SNP-index) of 10 consecutive SNPs). The $x$- and $y$-axis represent the percentage of individuals in each bulk ($p$) and the $m$ value, respectively. The results for F2 progeny (left) and F7 RIL progeny (right) are shown. Different read depths (50x, 20x, 10x and 5x) are indicated by different colors (inset). (b) and (c) The power of QTL-seq for detecting QTL in the cases of codominance (b) and dominance (c). Two values of QTL effect [$Q_p=0.1$ (top) and $Q_p=0.05$ (bottom)], as well as two types of populations (F2 and F7) were tested.
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 a causal mutation for a given phenotype by whole genome resequencing of 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 QTL for agronomically important traits, including partial resistance and seedling vigor, using RILs and F2 populations, respectively. The major advantage of QTL-seq is that it does not necessitate DNA marker development and marker genotyping for mapping purpose. 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 for a given genomic position. These two key attributes make QTL-seq an attractive method for quick and
cost-effective identification of QTL.

Bulked-segregant analysis was first applied to facilitate the linkage analysis of discrete characters in F2 populations (Giovannoni et al., 1991; Michelmore et al., 1991). In these studies, F2 progeny showing two discrete characters were isolated, and DNA from the F2 individuals were pooled to make two bulked DNA 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 DNA, markers showing difference between the two bulked DNA 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 F2 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 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 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. 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 available for QTL identification. Applications of whole genome sequencing to two bulked DNAs 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., 2012). QTL-seq has the same principle with the reported 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 reported previously.

QTL-seq applied to seedling vigor in rice demonstrated that this method successfully identifies QTL in F2 generation, which is a much earlier generation than F7 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 entire variation and if the read depth is more than or equal 20, its genomic position may be readily detected by QTL-seq even in F2 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 ssp. japonica, 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 vs. low temperature). We could then make two bulks of DNAs: one from multiple individuals from the high temperature and the other from the low temperature zones. Sequence reads from these two DNA bulks are compared to a reference sequence, and \( \Delta(\text{SNP-index}) \) is calculated for all the genomic regions. The regions showing higher \( \Delta(\text{SNP-index}) \) 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 based on whole-genome sequencing-based techniques, including QTL-seq and MutMap, 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 the breeding time.
MutMap-Gap

Whole genome resequencing of mutant F2 progeny bulk combined with *de novo* assembly of gap regions identifies the rice blast resistance gene *Pii*.

Summary

Next generation sequencing allows the identification of mutations responsible for phenotypes in mutants by whole genome resequencing and alignment to a reference. However, when the resequenced cultivar/line displays significant structural variation from the reference genome, mutations in the genome regions missing from the reference (gaps) cannot be identified by simple alignment. Here we report a method named “MutMap-Gap” that involves delineating a candidate region harboring a mutation of interest by the recently reported MutMap method, followed by *de novo* assembly, alignment, and identification of the mutation in the gaps. We applied MutMap-Gap to isolate the blast resistant gene *Pii* from the rice cultivar Hitomebore using mutant lines that lost *Pii* function. MutMap-Gap should prove useful for cloning genes that exhibit significant structural variations; such as disease resistance genes of the nucleotide-binding site and leucine rich repeat (NBS-LRR) class.

Introduction

The development in DNA sequencing technologies has enabled whole genome sequencing (WGS) of plant species to become routine. One of the applications of WGS is the identification of causative mutations responsible for mutant phenotypes of interest. For this purpose, a series of methods have been reported including SHOREmap (Schneeberger et al., 2009), NGM (Austin et al., 2011) and MutMap (Abe et al., 2012a).

In MutMap (Abe et al., 2012a), a recessive mutant with a phenotype of interest is crossed to the parental line used for the mutagenesis and DNAs from multiple individuals of mutant F2 progeny are bulk-sequenced. The resulting short reads are aligned to the “reference sequence” constructed for the parental line, and the alignment result is used to infer the genomic location of causal mutation responsible for the phenotype. The prerequisite of MutMap is that the genomic fragment spanning the causative mutation is present in the parental “reference sequence”. However, this is not guaranteed since the parental line “reference sequence” is constructed based on the “reference genome” of a representative cultivar/line of the species; e.g. Nipponbare cultivar in rice, Oryza sativa, and Col ecotype in Arabidopsis thaliana, by resequencing the parental line genome and replacement of nucleotides of “reference genome” with those of parental line at all the single nucleotide polymorphism (SNP) positions between the two lines. Consequently, MutMap cannot identify mutations located within the parental line-specific genomic regions (gaps) that are missing from the “reference genome”. To identify the mutation in such gap regions, we introduce MutMap-Gap, a combination of MutMap and targeted de novo assembly of gap regions.
Materials and Methods

Constructing the Hitomebore ‘reference sequence’ based on the Nipponbare ‘reference genome’

To generate the northern Japanese rice (*Oryza sativa*) cv Hitomebore parental line ‘reference sequence’, we aligned 12.25 Gb of Hitomebore wild-type (WT) parental line sequence reads obtained by illumina (San Diego, CA, USA) sequencing to the Nipponbare reference genome (build five genome sequence; [http://rapdblegacy.dna.affrc.go.jp/download/index.html](http://rapdblegacy.dna.affrc.go.jp/download/index.html)) by BWA (Li and Durbin, 2009) as described in Abe *et al.* (2012a). The Hitomebore ‘reference sequence’ was constructed by replacing Nipponbare nucleotides with those of Hitomebore at the 124,968 SNP positions that were identified between the two cultivars.

Whole genome sequencing of bulked DNA from F2 progeny and MutMap analysis

DNA samples were prepared by bulking DNA extracted from leaves of mutant F2 individuals as described previously (Abe *et al.* 2012a). The library for the illumina GAIIx sequencer was prepared from 5 lg DNA samples. Short reads were aligned to the Hitomebore ‘reference sequence’ using BWA software (Li and Durbin, 2009). Alignment files were converted to SAM/BAM files using SAMtools (Li *et al.* 2009), and the aligned short reads were filtered by Coval (Kosugi *et al.* 2013) to improve SNP calling accuracy. The SNP index was calculated as described previously (Abe *et al.* 2012a). The details are given in Figure 3-1. Significant peak of SNP-index were decided by generating confidence interval of SNP-index value under the null hypothesis of randomly bulked F2 progeny Figure 3-2. Sliding window analysis was applied with a 4 Mb window size and a 10 kb increment.
De novo assembly

A mate-pair library of the Hitomebore WT genome for contig scaffolding was prepared using the Mate Pair Library Prep Kit (illumina) and de novo assembly was performed with CLC (http://www.clcbio.com) software. The generated contigs were used to generate scaffolds using mate-pair sequence reads of Hitomebore WT DNA with SSPACE (Boetzer et al. 2011). Mutant screen Using a total of 3033 Hitomebore mutant lines generated by ethylmethanesulfonate (EMS) mutagenesis, we carried out a spray inoculation test with the rice blast fungus Magnaporthe oryzae isolate TH68-126 (race 033.1) carrying AVR-Pii (Yoshida et al. 2009) and incompatible with Hitomebore (with Pii).
Figure 3-1.
Summary of Illumina GAIIx sequencing and alignment data.
Figure 3-2. Simulation test for null distribution in MutMap-Gap.
Flow chart showing the simulation test for null distribution. (c) and (d) The result of simulation test at each read depth based on 10,000 replications. The x- and y-axis represent the read depth and SNP-index, respectively. The green and yellow lines represent 95 and 99% confidence interval for null distribution, respectively.

Simulation for random sampling
Here, we assume that the F2 progenies were derived from a cross between mutant “M” and its parental cultivar “P”. In addition, the genotypes of “P” and “M” at a given position are represented by AA and aa, respectively. For simulating random bulking of mutant F2 individuals, the frequency of “a” allele was decided at a random position in the genome. The number of alleles simulated corresponds to the MutMap samples described in the main text.

Simulation for calculating SNP-index
SNP-index is simulated by the probability distribution of SNP-index, which corresponded to the binomial distribution of B (at a given read depth, and “a” allele frequency).

Calculation of null distribution
The 95 and 99% confidence intervals of SNP-index are calculated based on the result obtained from 10,000 replications. We use this confidential interval as null distribution of the non-selected bulked DNA.
Results

Principle of MutMap-Gap

MutMap-Gap is a WGS-based method developed for the identification of the causative nucleotide change of a given mutant phenotype in a genomic region that is missing from the reference genome. MutMap-Gap combines the previously reported MutMap method (Abe et al. 2012a) with gap filling by de novo assembly. This is followed by identification of causative mutation responsible for the mutant phenotype in the assembled gap region, as described in Figure 3-3 and 2-4 using rice as an example.

First, mutant lines are developed by mutagenesis of the parental cultivar "P" with e.g. EMS. Given that cultivar “P” is different from the cultivar "Ref" for which accurate genome sequence is available, we first need to generate a “reference genome sequence” of the cultivar “P”. For this purpose, we resequence the cultivar “P” wild type (WT) and align the resulting reads to “Ref” genome sequence, which is the publically available reference genome. Following this step, nucleotides of “Ref” genome are replaced with those of the cultivar “P” at all the SNP positions identified between “P” and “Ref” (Figure 3-3a). Although the majority of sequence reads obtained for “P” are expected to be aligned to the “Ref” genome, the short reads derived from “P”-specific genomic region cannot be aligned to “Ref” genome, and thus are collected as unmapped reads.

Assume that we are interested in a mutant line “M” generated in cultivar “P” background, and the causal mutation for the phenotype under consideration resides in a “P”-specific genomic region (Figure 3-3b). Since the genomic region containing the causal mutation is not represented in the reference genome, simply aligning the “M” reads to the “reference sequence” cannot identify it. However, we can recover the target region by the combined use of MutMap (Abe et al., 2012a) and gap filling by de novo assembly. To this end, we first
delineate the approximate position of the causative mutation by MutMap. For MutMap analysis, "M" is backcrossed to "P" to generate an F2 progeny. Among the segregating F2 progeny, individuals showing the mutant phenotype are selected and bulked in equal ratio, and the bulked-DNA is subjected to whole-genome resequencing.

Figure 3-3. MutMap applied to a rice cultivar differentiated from the one used for generating the reference genome cannot identify mutations in the region missing from the reference.
To obtain the SNP information of the mutant, its short reads are aligned to the cultivar “P” “reference sequence” (Figure 3-3b). From the alignment data, we identify SNPs over the entire genome and calculate SNP-index representing the frequency of mutant type reads out of the total reads aligned at each SNP position. SNP-index graphs are plotted to graphically show the relationship between SNP-index and chromosomal positions for all the 12 chromosomes of rice. The SNPs in the candidate region close to the causal mutation are expected to have a higher SNP-index (SNP-index ~ 1), whereas those in unlinked regions show SNP-index around 0.5. Finding a peak of SNP-index in the graph allows the identification of the approximate genomic interval harboring the causal mutation (Abe et al., 2012a). However, as our candidate interval is located within the gap region, it is not possible to identify the causal mutation by MutMap alone (Figure 3-3c).

To target such a mutation located in the gap region, we apply *de novo* assembly to reconstruct the "P" genome sequence within the target interval delineated by the initial MutMap step (Figure 3-4). For *de novo* assembly, we utilize two types of sequence reads derived from “P” WT. These are the reads that could be aligned to the target interval region as delineated by MutMap, as well as those that could not be aligned to the reference sequence of the cultivar "Ref" (Figure 3-3a). The unmapped reads represent the unique sequences present in the cultivar “P” but missing in "Ref", and are assumed to contain the short reads corresponding to the gap regions within our candidate genomic internal in "P". Using these two types of reads, we perform *de novo* assembly using CLC (http://www.clcbio.com), and SSPACE (Boetzer et al., 2011) software to recover scaffolds presumably located in the target interval (Figure 3-4a).

Finally, the short reads derived from the bulked DNA of mutant type F2 are aligned to the combined sequences of the scaffolds produced by the *de novo* assembly and "P" cultivar “reference sequence” (Figure 3-4b). This procedure
allows the identification of SNPs residing within the newly generated scaffolds, for which SNP-index is calculated. The SNPs showing SNP-index = 1 are the likely candidates for locating the causative mutation for the mutant phenotype.

Figure 3-4. MutMap-Gap fills the gap within a target genome region delineated by MutMap with de novo assembly.
(a) Using a combination of the unmapped reads collected in the previous step (Figure 3-3a) and the short reads aligned to the target region (Figure 3-3c), de novo assembly is performed to reconstruct the sequence of the target interval. “P+scaffolds” reference sequence is used for alignment purpose in the following step. (b) Short reads derived from the bulk-DNA of mutant F2 progeny are aligned to the “P+scaffolds” reference.
**MutMap-Gap identifies the rice Pii, a blast resistance gene**

As a proof of principle, we applied MutMap-Gap for the identification of the rice blast (*M. oryzae*) resistance (R-) gene *Pii*. *Pii* confers resistance to rice against the blast pathogen harboring the corresponding *AVR-Pii* gene. The complete genome sequence of the rice ssp. japonica cultivar Nipponbare was published in 2005 (International Rice Genome Sequencing Project). However, Nipponbare is known to lack *Pii*, indicating that this cultivar cannot be directly used for cloning of the *Pii* gene.

To isolate *Pii*, we used a Northern Japanese rice cultivar “Hitomebore” known to harbor the *Pii* gene. We have been working on this cultivar over a decade and have already generated a total of about 15,000 EMS mutagenized lines (Abe *et al.*, 2012a; Rakshit *et al.*, 2010). Whole-genome resequencing of Hitomebore wild-type (WT) plant using an illumina GAIIx sequencer and replacing Nipponbare nucleotides with those of Hitomebore at all the SNP positions (124,968 positions) identified between these two cultivars allowed us to develop Hitomebore “reference sequence” (Hit-WT sequence). Of the 389Mb Nipponbare genome, about 358Mb (92%) were covered by the 10.71 Gb Hitomebore sequence reads generated, corresponding on average to 27.5x coverage of Nipponbare genome. Nevertheless, about 31Mb (8%) of Nipponbare genome could not be aligned (Table 3-1). These unaligned regions may either represent the regions that are difficult to sequence due to e.g. high GC contents or correspond to highly divergent sequences between Hitomebore and Nipponbare. On the other hand, we found short reads amounting to a total of 251 Mb that were unmapped to Nipponbare reference genome sequence. These short reads may be derived from Hitomebore-specific genomic regions not present in Nipponbare.
Table 3-1. Summary of Illumina GAIIx sequencing and alignment data.

<table>
<thead>
<tr>
<th>Sample name</th>
<th>No. of mutant F2 plants bulked</th>
<th>Total read (Gb) (^a)</th>
<th>Alignment data</th>
<th>Mean depth</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Mapped (Gb)(^b)</td>
<td>Unmapped (Gb)</td>
</tr>
<tr>
<td>Hitomebore(^c)</td>
<td>-</td>
<td>12.25</td>
<td>10.71</td>
<td>0.251</td>
</tr>
<tr>
<td>Hit5948(^d)</td>
<td>17</td>
<td>2.45</td>
<td>2.28</td>
<td>0.07</td>
</tr>
<tr>
<td>Hit6780(^d)</td>
<td>24</td>
<td>2.87</td>
<td>2.70</td>
<td>0.04</td>
</tr>
</tbody>
</table>

\(^a\) The short reads in which more than 10% of nucleotides scored less than 30 for phred quality were excluded from the analysis. Additionally, only those short reads having their corresponding pair reads were used for alignment.

\(^b\) The short reads produced by PCR duplication were removed for the alignment.

\(^c\) Hitomebore short reads were aligned to the Nipponbare reference genome (IRGSP build 5).

\(^d\) The short reads generated for the two mutant-bulks were aligned to the Hitomebore “reference sequence” that was developed by replacement of Hitomebore SNPs with those of Nipponbare.

For screening of Hitomebore mutants that lost \(Pii\) function, we carried out spray inoculation test using an incompatible blast race 033.1 having \(AVR-Pii\) on a total of 3,033 EMS-mutagenized Hitomebore lines (Figure 3-5a). Accordingly, we identified two independent \(Pii\) deficient candidate mutants, Hit5948 and Hit6780, which showed susceptible phenotypes following inoculation with the blast race 033.1 (Figure 3-5b). For mapping the causal mutations by MutMap, we independently crossed the two mutants to Hitomebore WT and generated two F2 progeny. The Hit5948\(\times\)WT F2 progeny segregated 61 (WT) to 17 (mutant-type) phenotypes, whereas the Hit6780\(\times\)WT F2 progeny segregated 88 (WT) to 24 (mutant-types) phenotypes. In both cases, the segregation conformed to the 3:1 ratio (Chi square test: \(\chi^2 = 0.43\), ns in Hit5948\(\times\)WT F2; \(\chi^2 = 0.76\) ns in Hit6780\(\times\)WT F2), indicating that in both cases the mutant phenotypes were caused by single recessive mutations (Figure 3-5c).

For MutMap analysis, we bulked the DNA of mutant-type F2 progeny (17 individuals for Hit5984 and 24 individuals for Hit6780) and applied to whole genome sequencing by illumina GAIIx DNA sequencer. We carried out 75bp paired-end sequencing and obtained 2.45 Gbp and 2.87 Gbp sequence reads for
the DNA samples obtained from Hit5948 and Hit6780, respectively (Table 3-1). The sequence reads were aligned to Hitomebore (Hit-WT) “reference sequence” and SNPs were identified. For all SNP positions, SNP-index was calculated and graphs relating SNP position and SNP-index were generated for further analysis (Figure 3-6, Figure 3-7).

MutMap applied to Hit5948 revealed a peak of SNP-index value in the genomic interval of 7.88 Mb to 11.98 Mb on chromosome 9. Of the four SNPs with “SNP index = 1” identified in the candidate region, SNP-10290916 was localized in the second exon of the gene Os09t0327600-01 gene predicted in Nipponbare reference genome (Table 3-2). This SNP represented a nonsense mutation causing amino-acid change from Trp (TGG) to a stop codon (TGA) at the 226th-aa residue. Os09t0327600-01 encodes a protein with nucleotide binding site and leucine rich repeat (NBS-LRR) domain that are conserved in plant resistance genes (Jones and Dangle, 2001; Zhou et al., 2004). Therefore, we hypothesized that Os09t0327600-01 homolog in Hitomebore functions as Pii, although Nipponbare lacks functional Pii. A scrutiny of Os09t0327600-01 of Nipponbare showed that this gene encodes a truncated R-protein that is likely non-functional.
Figure 3-5. Screening Pii mutants and the segregation test in F2 progeny.
(a) A total of 3,033 Hitomebore EMS mutant lines were screened for their resistance response by inoculation with M. oryzae isolate 033.1 that contains AVR-Pii avirulence genes. (b) The cultivar Hitomebore contains Pii R-gene, and thus shows resistance to M. oryzae isolate 033.1. Whereas the two mutants, Hit5948 and Hit6780 are susceptible, suggesting they have lost Pii resistance. (c) The segregation of two progenies generated by crossing the two Pii mutants Hit5948 and Hit6780 back to Hitomebore wild-type for resistance (Wild type) and susceptible (Mutant type) phenotypes confirmed to 3:1 ration, suggesting the phenotype is governed by a single recessive gene. The resistance response is assessed by punch or spot inoculation of M. oryzae isolate 033.1.

Figure 3-6. MutMap reveals genomic regions harboring candidate mutations of Hit5984 and Hit6780, two mutants that lost Pii resistance.
MutMap results of Hit5948 and Hit6780 for chromosome 9. Blue dots indicate SNP-index values at a given SNP position. Red lines represent the sliding window average SNP-index values of 4 Mb interval with 10 kb increment. The green lines show the 95% confidence limit of SNP-index value under the null hypothesis of SNP-index = 0.5. Chromosomal regions delineated by grey color indicate the genomic regions presumably harboring the causal mutations.
(a) MutMap analysis of Hit5948.

Figure 3-7.
(b) MutMap analysis of Hit6780.

Figure 3-7.
SNP-index plots for the 12 chromosomes of rice was generated following MutMap analysis of bulked mutant DNA samples with the statistical confidence intervals under the null hypothesis of bulked DNA without any selection. The red line indicates the sliding window average of 4 Mb interval with 10 kb increment for SNP-index. Green and yellow lines represent the sliding window average of 4 Mb interval with 10 kb increment for the 95% and 99% statistical confidence intervals under the null hypothesis of DNA bulking without any selection, respectively. (a) MutMap analysis of Hit5948. (b) MutMap analysis of Hit6780.

Table 3-2.
The candidate SNPs having SNP-index of 1 within the region that exhibited statistically significant (Fisher's exact test: $P<0.05$) differences between Hit5948 mutant- and wild-type bulk sequences a.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Position</th>
<th>Reference base</th>
<th>Altered base</th>
<th>Depth</th>
<th>Mutated gene</th>
<th>intron or exon</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>9726667</td>
<td>C</td>
<td>T</td>
<td>8</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>10290916</td>
<td>C</td>
<td>T</td>
<td>7</td>
<td>Os09t0327600 exon W (TGG) to stop (TAG)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>11319160</td>
<td>C</td>
<td>T</td>
<td>6</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>11659480</td>
<td>C</td>
<td>T</td>
<td>5</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
</tbody>
</table>

a The sequence reads were aligned to Hitomebore “reference sequence” developed in this study.

Table 3-3.
The candidate SNPs having SNP-index of 1 within the region that exhibited statistically significant (Fisher's exact test: $P<0.05$) differences between Hit6780 mutant- and wild-type bulk sequences a.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Position</th>
<th>Reference base</th>
<th>Altered base</th>
<th>Depth</th>
<th>Mutated gene</th>
<th>Intron or exon</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>9</td>
<td>7752759</td>
<td>C</td>
<td>T</td>
<td>6</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>9923713</td>
<td>C</td>
<td>T</td>
<td>9</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>9933054</td>
<td>C</td>
<td>T</td>
<td>8</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>10147622</td>
<td>C</td>
<td>T</td>
<td>7</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>10288490</td>
<td>C</td>
<td>T</td>
<td>11</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>10560211</td>
<td>C</td>
<td>T</td>
<td>15</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>10706016</td>
<td>C</td>
<td>T</td>
<td>6</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>10854185</td>
<td>C</td>
<td>T</td>
<td>12</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>11098043</td>
<td>C</td>
<td>T</td>
<td>7</td>
<td>No hit</td>
<td>--</td>
<td>--</td>
</tr>
<tr>
<td>9</td>
<td>11495413</td>
<td>C</td>
<td>T</td>
<td>12</td>
<td>Os09t0346900 Intron</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>

a The sequence reads were aligned to Hitomebore “reference sequence” developed in this study.
A similar analysis applied to Hit6780 using MutMap identified a candidate genomic region likely harboring the causative mutation in the 7.18 Mb - 13.05 Mb interval on chromosome 9, an identical region to which the causal mutation of Hit5948 was mapped. Although ten SNPs with SNP-index = 1 were identified in the region, none represented non-synonymous mutations. No SNP was detected in Os09t0327600-01, the candidate gene for Hit5948 (Table 3-3). We hypothesized that the causative mutation of Hit6780 is located in the Hitomebore-specific genomic region, which is not represented in the reference genome sequence of Nipponbare (Figure 3-8a).

To identify the causative mutation of Hit6780, we applied MutMap-Gap analysis. Accordingly, we retrieved all the short reads (4,849,550 reads) of Hitomebore WT mapped to 7.18 Mb - 13.05 Mb region on chromosome 9 and combined them with the Hitomebore WT reads (3,358,005 reads) that could not be mapped to the Nipponbare reference genome. These combined short reads were used for de novo assembly with CLC software (http://www.clcbio.com), generating a total of 2241 contigs that are over 1 kb in size. For scaffolding of the contigs with the SSPACE (Boetzer et al., 2011) software, about 46,989,929 mate-pair short reads were used, generating about 852 scaffolds with a minimum size of 1 kb (Figure 3-9). The scaffolds were then combined with Hitomebore “reference sequence” (Hit-WT), and used as a reference for aligning the short reads derived from bulked-DNA of mutant type F2 of Hit6780 as in MutMap analysis. Of the 852 scaffolds generated, only two harbored SNPs with SNP-index = 1. Gene prediction by GENSCAN (http://genes.mit.edu/GENSCAN.html) using the 2 scaffolds revealed that one of the SNPs corresponded to an intergenic region and only the SNP in scaffold No. 7 (length = 63,355 bp) was localized within the splicing junction of a gene we tentatively named hit7 (Figure 3-10). This Hit6780 mutation causes miss-spliced mRNA probably giving rise to a truncated protein (Figure 3-10). The hit7 gene contains NBS-LRR domain, implying that this SNP the likely
causal mutation of Hit6780.

We compared the DNA sequences of hit7 with that of Os09t0327600-01, and found a high similarity in the region where the candidate SNP of Hit5948 was detected (Figure 3-8a,b). Accordingly, we assumed that the causative mutations of Hit5948 and Hit6780 are located within the same gene. The conventional MutMap could detect the Hit5984 mutation that was localized in the region where sequence similarity was very high between hit7 of Hitomebore and Os09t0327600-01 of Nipponbare. On the other hand, the causal SNP of Hit6780 could not be undetected by MutMap because it is localized in the region missing in Nipponbare reference genome. The Nipponbare Os09t0327600-01 gene seems to have lost R-gene function by truncation of the region corresponding to C-terminus of the protein (Figure 3-8a,b).
Figure 3-8. MutMap-Gap identifies the causal mutation of Hit6780.
(a) A snapshot of IGV genome viewer of the chromosomal location spanning the 10289000-10293000 region of chromosome 9 in the Nipponbare reference genome region showing the predicted Os09t0327600-01 gene aligned with short reads derived from Hitomebore WT. The short reads cover only the predicted exon regions of Os09t0327600-01. (b) A snapshot of IGV genome viewer showing the predicted hit7 gene region spanning the scaffold 7 generated by de novo assembly. The predicted gene structure is given by boxes (Exons and UTRs) and lines (introns). The alignment was made with short reads generated from Hitomebore WT. The hatched area in the coding region of hit7 (second exon) shows >95% similarity with the Os09t0327600-01 region in (a). Red arrows indicate the positions of the candidate mutations in Hit5948 (left; position 1783) and Hit6780 (right; position 2567). The genomic position of Hit5984 mutation is conserved between Hitomebore and Nipponbare, whereas that of Hit6780 is not. (c) Confirmation of the candidate SNPs by Sanger sequencing. DNA sequencing peak chromatograms of genomic DNA in the regions around the mutations for Hit5948 (left) and Hit6780 (right) as compared with Hitomebore WT. Arrows indicate the mutated nucleotides.
Summary of scaffolds generated by scaffolding of the contigs using illumina mate-pair reads. A total of 852 scaffolds that are over 1kb in size were selected for the MutMap-Gap analysis. The red line indicates the scaffold N50 value.

(a) The structure of *Pii* gene is indicated by grey boxes (5’ and 3’ UTRs), black boxes (exons) and lines (introns). The inverted black arrow indicates the mutated nucleotide position of Hit6780. Red arrows represent locations of the primer pairs used for RT-PCR analysis. The primer sequences are provided below the gene. (b) The RT-PCR result showing the expected 431 bp and 618 bp fragments for Hitomebore wild type and Hit6780, respectively.
To check whether both of Hit5948 and Hit6780 have mutations in the same gene, we carried out complementation test by crossing Hit5948 with Hit6780. As expected, the phenotype of F1 plants harboring the Hit5948 and Hit6780 mutations in heterozygous state (Figure 3-11a) could not be recovered to the wild type phenotype (Figure 3-11b), confirming that Hit5984 and Hit6780 mutants have defects in the same gene hit7. We further tested the association between presence/absence of $P_{ii}$ phenotype and cleaved amplified polymorphic sequence (CAPS) marker polymorphism discriminating hit7 and Os09t0327600-01 sequences using a total of 30 rice cultivars. A complete association was observed between the $P_{ii}$ phenotypes and CAPS patterns, supporting the identity of $P_{ii}$ gene as hit7 (Figure 3-12a,b). We accordingly renamed hit7 as $P_{ii}$.

The $P_{ii}$ gene is composed of five exons, and the 3078 bp coding sequence spanning the start and stop codons encodes a 1028 amino acid protein predicted as a putative nucleotide-binding site leucine-rich repeat (NBS-LRR) type R-protein (Figure 3-12c), which is typical of the majority of disease resistance genes in plants (Belkhadir et al., 2004; Zhou et al., 2004; McHale et al., 2006).
Figure 3-11. Allelism test of Hit5984 and Hit6780 mutations.
(a) F1 plants were obtained from a cross between Hit5984 and Hit6780. Genomic DNA was extracted from F1 plants, and two genomic regions spanning the mutations in Hit5984 (position 1783) and Hit6780 (position 2567) were amplified and sequenced. As expected, F1 plants showed heterozygous peaks of T/C and G/A at the 1783th and 2567th positions, respectively. (b) Results of punch inoculation test of fro Hitomebore WT and the F1 plant with an incompatible race (033.1). The F1 remained susceptible to the race 033.1, suggesting that Hit5984 and Hit6780 have defects in the same gene. Bar = 1 cm.
Figure 3-12.
Figure 3.12.
Fig. 3-12.
Genomic structure of *Pii* determined by 5-RACE and 3-RACE PCR.
(a) Purple boxes, red boxes and black lines represent the 3′ and 5′ UTRs, exons, and introns of the *Pii* gene, respectively (top). The 3078 bp coding is indicated at the bottom. (b) The gDNA sequence of *Pii* showing the 5′ and 3′ UTRS (purple color), exons (red color), introns (black) and the mutated residue in Hit5948 (green bar). (c) *Pii* is predicted to encode a 1026-aa protein that contains an NB-ARC domain (191–460 aa residues) and LRR domain (581–855 aa residues).
**Discussion**

Successful identification of a causative nucleotide change responsible for a mutant phenotype by whole genome resequencing depends on the availability of a good reference genome sequence. If the resequenced mutant belongs to a cultivar/line which is structurally different from the one used for constructing the reference genome, it is possible that the causative mutation may reside in the genomic region specific to resequenced cultivar but absent from reference genome. In the past, such regions have been addressed by generating BAC library of the cultivar/line used for mutant screen, followed by screening of the BAC clones by DNA probes tightly linked to the target region and by sequencing of these clones. This procedure has been laborious and time consuming. MutMap-Gap as presented here circumvents such process, and offers a fast alternative approach to obtain sequence information for a genomic region of interest that is missing from the reference genome. The power of MutMap-Gap resides in the quick delineation of the target genomic region harboring the causal mutation by MutMap, followed by local *de novo* assembly with short reads corresponding to the target region.

An advantage of MutMap is that it substitutes the conventional marker-based linkage analysis and accelerates the identification of target regions by a single WGS of bulked DNA of mutant-type F2 progeny. Instead of screening BAC library, MutMap-Gap then allows gap filling by *de novo* assembly of short reads that are unmapped to reference genome as well as using a subset of mapped short reads derived from the target region. The latter only serves the purpose of enhancing the *de novo* assembly of the target region. This targeted *de novo* assembly of a narrower genomic region delineated by MutMap is the core of MutMap-Gap method. Rice blast is a destructive and widespread disease caused by an ascomycete pathogen *M. oryzae* and accounts for significant yield losses worldwide. For efficient marker-assisted breeding
of rice blast resistance, the identification of R-genes is important. R-genes, which mostly encode the NBS-LRR class proteins, have a unique mode of evolution and represent a highly divergent group in plants (Clark et al., 2007). The successful isolation of the rice blast resistant Pi2 gene from Hitomebore whose “reference sequence” is constitute based the rice Nipponbare reference genome lacking Pi2 demonstrates the power of MutMap-Gap.

In conclusion, WGS-based gene isolation methods including SHOREmap (Schneeberger et al., 2009), NGM (Austin et al., 2011) and MutMap (Abe et al., 2012a) promise to accelerate the identification and cloning of plant genes of importance to agriculture. Ideally, these methods need to be applied to elite crop cultivars to make them immediately available to breeders (Abe et al., 2012a). However, in most cases, a high quality genome sequence of the elite cultivar is not available, and this needs to be reconstructed by comparison to a reference genome as in the case of the Northern Japanese rice cultivar Hitomebore. Such genome sequence reconstruction approaches would miss regions with structural variations, which are common in certain genes such those for disease resistance. As MutMap-Gap addresses this issue, we envisage this method will further broaden the opportunity to use WGS for isolating novel plant genes by quick forward genetics approaches.
Chapter 4
Summary

Advances in genome sequencing technologies have enabled researchers and breeders to rapidly associate phenotypic variation to genome sequence differences. We recently took advantage of next-generation sequencing technology to develop MutMap, a method that allows rapid identification of causal nucleotide changes of rice mutants by whole genome resequencing of pooled DNA of mutant F2 progeny derived from crosses made between candidate mutants and the parental line. Here we describe MutMap+, a versatile extension of MutMap, that identifies causal mutations by comparing SNP frequencies of bulked DNA of mutant and wild-type progeny of M3 generation derived from selfing of an M2 heterozygous individual. Notably, MutMap+ does not necessitate artificial crossing between mutants and the wild-type parental line. This method is therefore suitable for identifying mutations that cause early development lethality, sterility, or generally hamper crossing. Furthermore, MutMap+ is potentially useful for gene isolation in crops that are recalcitrant to artificial crosses.


*These authors equally contributed to this work.
Introduction

Recent developments in next-generation sequencing (NGS) technologies are revolutionizing various aspects of genomics. In particular, NGS-based whole genome analysis of organisms is becoming routine. Draft genome sequences of a large number of organisms including non-model species are rapidly accumulating thanks to the tremendous advances made in *de novo* sequencing techniques and the development of highly efficient assembly software (Zerbino and Birney 2008; Gnerre et al., 2011). Once whole genome draft sequences become available, resequencing of multiple individuals of the same species allows rapid identification of genomic variations, contributing to genetic analyses of medical conditions in human as well as important traits in crops, animals and microbes (Schneeberger et al., 2009; Birkeland et al., 2010; Arnold et al., 2011; Austin et al., 2011; Abe et al., 2012a; Gonzaga-Jauregui et al., 2012).

Whole genome sequencing technology promises to dramatically impact crop improvement in this era of looming food crisis and an ever-increasing world population. To exploit genome sequencing in crop breeding, we recently developed a method called MutMap and applied it to the identification of rice genes responsible for agronomically important traits (Abe et al., 2012a). For MutMap application, we generated over 12,000 mutant lines of M2-M5 generations by ethylmethane sulfonate (EMS) mutagenesis of a Northern Japanese rice (*Oryza sativa* ssp. japonica) cultivar “Hitomebore” (Rakshit et al., 2010). The mutant lines show wide variations of phenotypes in agronomically important traits like plant height, grain number and disease tolerance. Using MutMap, we identified the unique genomic positions that harbor mutations causing pale green leaves and semi-dwarfism. Because it relies on crosses to the parental line and eliminates the need for wide-crosses to genetically unrelated lines, MutMap is particularly useful for identifying genes that
determine quantitative minor effect phenotypes, which is a major challenge in crop improvement (Abe et al., 2012a).

MutMap is based on the crossing of a mutant of interest to the parental line that was used for the mutagenesis, followed by selfing of F1 individuals to generate F2 progeny. If the phenotype is caused by a single recessive mutation, the F2 population is expected to segregate 3:1 for the wild-type and mutant progeny. DNA from about 20 F2 individuals showing the mutant phenotype is pooled in an equal ratio and subjected to illumina whole genome sequencing with depth of more than 10x coverage. The short reads are then aligned to the reference genome sequence constructed for the cultivar used for the mutagenesis, to Hitomebore reference sequence in our case. Since the causal SNP is shared by F2 mutant progeny, all the re-sequenced short reads covering such SNP should have a nucleotide different from the reference sequence. In contrast, SNPs that are not relevant to the phenotype under consideration should segregate in 1:1 ratio among the F2 progeny. Consequently, about half of the short reads covering such positions a nucleotide that is different from the reference genome. To quantify the proportion of short reads having nucleotides different from the reference sequence (SNPs), we developed the concept of SNP-index. If the entire short reads covering a particular genomic position share a SNP that differs from the reference, the SNP-index is defined as 1, whereas if only half of the short reads share such as SNP, then the SNP-index is 0.5. The SNP-index is calculated for all the SNPs incorporated by mutagenesis, and the relationship between SNP-index and genomic position is graphically plotted. The genomic region showing a unique SNP-index peak (SNP-index =1) corresponds to the position of the causal mutation responsible for phenotype of our candidate mutant.

The MutMap method is based on selecting mutants of interest at M3-M5 generations, and crossing them to the parental line, followed by evaluation of phenotypes in segregating F2 progeny. Accordingly, mutants that are not
amenable for crossing, e.g. mutants with early development lethality or sterility, are not suitable for MutMap application. To address this problem, we developed MutMap+, a versatile extension of MutMap that is based on selfing of heterozygous plants showing wild-type phenotype and identified in M2 progeny segregating for wild-type and a mutant phenotype of interest that is recessive homozygous. The resulting M3 population segregates for the target mutation and can be directly used for whole-genome sequencing. Since MutMap+ circumvents the necessity of backcrossing to the wild-type parent, it is suitable for the isolation of mutations in genes that hamper artificial crossing. Thus MutMap+ expands the applications of MutMap for the genetic improvement of rice and other crop plants.
Materials and Methods

Plant Materials

The rice (*Oryza sativa* L.) mutants used in this study, Hit9188 and Hit11440, were identified in M2 mutant population generated by ethyl-methanesulfonate (EMS) treatment of immature embryo of an elite japonica cultivar Hitomebore. Details of the mutagenesis protocol are provided elsewhere (Rakshit *et al.* 2010). For identification of candidate mutants for MutMap+ application, about 10 individuals from each M2 line were germinated and grown in the paddy field. Of these, the lines that segregated for mutant and wild-type phenotypes were selected. Growing about 100 seeds from each of the M2 mutant plants showing wild-type phenotypes allowed identification of heterozygous M2 individual plants, which were further used for establishing segregating M3 progeny utilized for making the wild-type and mutant bulks for illumina sequencing.

Whole Genome Sequence of Bulked DNA

Genomic DNA extracted from 100 mg of fresh leaf sample of each M3 individual selected using the DNeasy Plant Mini Kit (QIAGEN Sciences) was mixed in an equal ratio to make the bulk-DNA used for sequencing. The library for illumina sequencing was constructed from five micrograms of DNA sample and sequenced for 76 cycles on an illumina Genome Analyzer IIx as described in Abe *et al.* (2012a).

Alignment of Short Reads to Reference Sequence and SNP Calling

We aligned the short reads from bulked DNA (DRA001007) to the reference genome of the cultivar Hitomebore (DRA000499) using BWA (Burrows-Wheeler Aligner) software (Li and Durbin 2009). 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 (Kosugi et al. 2013; Abe et al. 2012a) to increase SNP-calling accuracy. Following SNP-index calculation for all SNP positions, we excluded SNPs with SNP-index of < 0.3 from the analysis as they may represent spurious SNPs called due to sequencing and/or alignment errors. For calculating Δ(SNP-index), we only used the SNPs detected in both bulked DNAs.

**Sliding Window Analysis**

Sliding widow analysis was applied with 4 Mb window size and 10 kb increment. In this sliding window analysis, we calculated average SNP-index and average P-value in Fisher's exact test for the SNPs located in the window. Finally we mapped the causal mutation in the window exhibiting the average P-value < 0.05.

**RNAi Experiments**

For RNAi analysis targeting the *OsNAP6* gene, a 300 bp *OsNAP6* partial fragment was amplified from wild-type Hitomebore cDNA by PCR and subcloned into the binary vector pANDA (Miki and Shimamoto 2004). Binary vectors were introduced into Agrobacterium strain EHA105, which as used to transform Hitomebore plants as previously described (Toki et al., 2006).
Results

*Principle of MutMap+*

The principle of MutMap+ is explained in Figure 4-1 using rice as an example. We applied EMS mutagenesis to rice immature embryo immediately after fertilization, and the embryos were allowed to develop to seeds. These seeds were sown to generate M1 plants, in which the majority of EMS mutations were expected to be in the heterozygous state (Figure 4-1a). Although phenotypes of dominant or semi-dominant mutations could be observed at this generation, M2 seeds obtained from selfing of M1 plants were advanced to M2 generation as our focus was on recessive mutations. We then selected those M2 progeny that segregated 3:1 for wild-type and mutant phenotypes, respectively (Figure 4-1b). For application of the MutMap+ scheme, we planted ~10 M2 progeny per line in the field, and scored phenotypic segregation (Figure 4-1b). For the lines segregating for a phenotype of interest, we expected that 2/3 of the wild-type siblings are heterozygous for the causal mutation. These wild-type M2 siblings were carefully grown to generate as many M3 seeds as possible that were harvested separately from individual plants. Over 80 M3 seeds derived from each M2 parent were then sown in the field to observe the segregation of phenotype. M3 families that segregated for the wild-type and mutant phenotypes were selected for further analysis (Figure 4-1c).

Next, we made two bulks of DNA, one from 20-40 M3 individuals showing the mutant phenotype and another from 20-40 M3 individuals with the wild-type phenotype, both of which were derived from selfing of a single heterozygous M2 plant (Figure 4-1c). These two bulks of DNA were separately sequenced by illumina GAIIx sequencer, and aligned to the reference sequence of the parental cultivar. For each bulk, we generated SNP-index vs SNP genomic position graphs (Figure 4-1d).
There are two reasons for the occurrence of genomic regions with SNP-index = 1 in the SNP-index plots obtained for the mutant DNA pool (Figure 4-1d). The first is that the region actually harbors the causal mutation for the phenotype. The second is because SNPs irrelevant to the phenotype become fixed to homoyzous state in the M2 generation and thus are shared by all M3 plants. This is expected to happen in 50% of the total genomic region. It is not possible to discriminate between the two types of SNPs by simply looking at the SNP-index plot of the mutant bulk. However, it is possible to identify those SNPs with SNP-index = 1 that are fixed to homozygous state by comparing the SNP-index plots of the wild-type and mutant bulks. Regions showing SNP-index = 1 by random fixation of SNPs in the M2 generation should be shared between the two bulks, whereas the region harboring the causal mutation should be specific to the mutant bulk. Regions with SNP-index of 1 resulting from the two types of SNPs should be mutually exclusive since the M2 individual used for generating the M3 progeny sequenced is heterozygous for the region harboring the causal mutation. In practice, we scan the SNP-index plots of the mutant- and wild-type bulks to find regions with SNP-index = 1 specific to the mutant bulk (Figure 4-1d). To visualize this data, we subtract the SNP-index of the wild-type bulk from that of mutant-type bulk for each SNP to obtain a Δ(SNP-index) (Figure 4-1e). Δ(SNP-index) value should be around 0 for most of the genome, but it is significantly positive in the region harboring the causal mutation. To assess the statistical significance of the Δ(SNP-index) values, we apply Fisher’s exact test (Fisher 1935).
Figure 4-1. A simplified scheme of MutMap+.

(a) After EMS mutagenesis of rice at immature embryo stage, the resulting seeds are used to establish M1 generation. In this generation, most of mutations are in the heterozygous state. (b) M2 progeny obtained from a self-fertilized M1 plant segregate in wild-type (indicated by green color) and mutant (brown color) phenotypes. Here we focus on a wild-type heterozygous plant. (c) The heterozygous M2 plant is selfed to obtain M3 progeny that segregate to wild- and mutant-type progeny. Genomic DNA from 20-40 mutant-type and an equal number of wild-type M3 progeny are separately bulked, and subjected to whole-genome sequencing. The resulting short reads are aligned to reference sequence of the cultivar used for mutagenesis. (d) SNP-index is obtained for each SNP, and plots relating SNP-index and chromosome positions are made for both the mutant- and wild-type M3 bulks separately, and the two SNP-index plots are compared to identify the region with SNP-index = 1 that is specific to the mutant bulk. (e) We can also evaluate Δ(SNP-index) plot, which is obtained by subtracting SNP-index value of wild-type bulk from that of mutant-type bulk. Genomic region harboring the causal mutation should have positive Δ(SNP-index) values.
As a proof of concept, we applied MutMap+ to a rice mutant Hit9188 derived from EMS mutagenesis of cultivar Hitomebore. This mutant, identified among ten M2 plants that originated from a self-fertilized M1 individual planted in the paddy field, is characterized by dwarfism and pale green leaves followed by premature death 3 weeks after germination (Figure 4-2). Thus, the mutant could not be crossed to the parental line Hitomebore to apply the classical MutMap analysis. As a result, wild-type M2 siblings of the mutant were grown to maturity and allowed to self-fertilize. We obtained over 100 seeds/plant, which were sown separately to generate M3 progeny that were further assessed for phenotypic segregation. One M3 progeny segregated 167 (WT) to 56 (mutant-type), conforming to the 3:1 ratio (Chi square test: \( \chi^2 = 0.0015, \text{NS} \)). This indicated that the mutant phenotype is caused by a single recessive mutation. We then made two bulks of DNA; a mutant bulk of 40 mutant-type progeny and a wild-type bulk of 40 wild-type progeny. The mutant bulk and the wild-type bulk were separately sequenced to generate illumina GAIIx 75bp paired end reads. We obtained 95,552,424 and 57,925,150 short reads corresponding to 7.17 Gbp and 4.34 Gbp sequence reads for the mutant and wild-type bulks, respectively (Table 4-1).

The illumina short reads obtained for the two DNA bulks (DDBJ Sequence Read Archive: DRA001007) were separately aligned to the reference sequence of Hitomebore (DRA000499) and the SNP-index was calculated for each SNP (Abe et al., 2012a). The relationship between the SNP-index and chromosome position was plotted for the 12 rice chromosomes as shown in Figure 4-3a and 3-3b. To help visualize the patterns of SNP-index, we calculated the sliding window average SNP-index values of 4 Mb interval with 10 kb increment. The SNP-index plots were very similar between the mutant and wild-type bulks across the entire genome (Figure 4-3a and 3-3b). However, we identified a single unique region on chromosome 1 (between 0 and 5.23Mb) with a peak of
SNP-index close to 1 in the mutant bulk that is missing from the wild-type bulk (Figure 4-3b, Figure 4-3a and 3-3b), and this could be readily visualized using a Δ(SNP-index) graph (Figure 4-3c). As expected, Δ(SNP-index) was close to 0 across the genome, but within the unique genomic region identified on chromosome 1, 0-5.23Mb, its value was greater than zero. This was the only region that exhibited a SNP index difference of > 0 that is statistically significant between the mutant and wild-type bulks (Fisher’s exact test: $P < 0.05$).

**Figure 4-2.**
Early senescence and premature death of Hit9188 plants.
Phenotype of Hitomebore wild-type and mutant plants (a) 14 DAS (days after sowing), (b) 18 DAS and (c) 22 DAS. Bar = 10 cm.

**Table 4-1.**
Output summary of Illumina GAIIx sequencing of Hit9188 and Hit11440 M3 bulks.

<table>
<thead>
<tr>
<th>Mutant</th>
<th>Sample</th>
<th>Number lines bulked</th>
<th>Number of reads</th>
<th>Total sequence (Gb)</th>
<th>Genome coverage (%)</th>
<th>Mean depth$	extsuperscript{c}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hit9188</td>
<td>Mutant-type bulk</td>
<td>40</td>
<td>95,552,424</td>
<td>7.17</td>
<td>97.4</td>
<td>16.5</td>
</tr>
<tr>
<td></td>
<td>Wild-type bulk</td>
<td>40</td>
<td>57,925,150</td>
<td>4.34</td>
<td>95.2</td>
<td>9.9</td>
</tr>
<tr>
<td>Hit11440</td>
<td>Mutant-type bulk</td>
<td>20</td>
<td>67,392,352</td>
<td>5.05</td>
<td>96.7</td>
<td>12</td>
</tr>
<tr>
<td></td>
<td>Wild-type bulk</td>
<td>20</td>
<td>62,529,426</td>
<td>4.68</td>
<td>97.8</td>
<td>11.4</td>
</tr>
</tbody>
</table>

$	extsuperscript{a}$Sequence reads of each sample were aligned to the Hitomebore consensus sequence as described in the Material and Methods.

$	extsuperscript{b}$Percentage of total genomic region of Hitomebore reference sequence aligned by short reads.

$	extsuperscript{c}$Average read depth over the whole genome.
Figure 4-3.
Figure 4-3.
(c) $\Delta$(SNP-index) plot obtained by subtraction of Wt-bulk SNP-index from Mt-bulk SNP-index

Figure 4-3.
After identifying the region with SNP-index = 1 that is specific to the mutant bulk, we scrutinized in detail the SNPs therein (Chromosome 1, 0-5.23 Mb). Accordingly, we found a total of 8 SNPs with SNP index = 1 in the region for the mutant bulk (Table 4-2). Of these, only two SNP (nucleotide 1234738 and 1503571) were nonsynonymous causing amino acid changes (Table 4-2). The first mutation (nucleotide 1234738), a C to T transition, causes amino acid change from proline to leucine in Os01g0121800, a gene encoding a glycosyl tranferase family 14 protein (Table 4-2). The second SNP (nucleotide 1503571) causes amino acid changes in a 482 amino acid protein encoded by Os01g0127300 (Figure 4-4a). This SNP (C in Hitomebore to T in Hit9188: G to A in sense-strand) alters amino acid 378 from alanine to threonine (Figure 4-4b). Because of the type of proteins encoded by the two genes harboring the mutations detected in Hit9188 and the phenotype of the mutant, we suspected that the second mutation is the most likely candidate, and thus focused on this SNP for further analysis. We used amplicon sequencing with the Sanger method to confirm that this SNP at position 1503571 is G in Hitomebore wild-type, A in the mutant bulk, and a mixture of G and A in the wild-type bulk (Figure 4-4c).
Table 2. The candidate SNPs having SNP-index of 1 within the region that exhibited statistically significant (Fisher's exact test: $P<0.05$) differences between Hit9188 mutant- and wild-type bulk sequences.

<table>
<thead>
<tr>
<th>Chr.</th>
<th>Position</th>
<th>Reference base</th>
<th>Altered base</th>
<th>Depth</th>
<th>Mutated gene</th>
<th>Amino acid change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>974469</td>
<td>G</td>
<td>A</td>
<td>6</td>
<td>No hit</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1234738</td>
<td>C</td>
<td>T</td>
<td>26</td>
<td>Os01t0121800 (Glycosyl transferase, family 14)</td>
<td>P to L</td>
</tr>
<tr>
<td>1</td>
<td>1379378</td>
<td>C</td>
<td>T</td>
<td>5</td>
<td>No hit</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1503571</td>
<td>C</td>
<td>T</td>
<td>18</td>
<td>Os01t0127300 (SuFBD family)</td>
<td>A to T</td>
</tr>
<tr>
<td>1</td>
<td>1523129</td>
<td>C</td>
<td>T</td>
<td>27</td>
<td>No hit</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>1674496</td>
<td>C</td>
<td>T</td>
<td>20</td>
<td>Os01t0130000 (Cation efflux protein family)</td>
<td>None (intom)</td>
</tr>
<tr>
<td>1</td>
<td>3025549</td>
<td>C</td>
<td>T</td>
<td>6</td>
<td>No hit</td>
<td></td>
</tr>
<tr>
<td>1</td>
<td>3485142</td>
<td>C</td>
<td>T</td>
<td>5</td>
<td>No hit</td>
<td></td>
</tr>
</tbody>
</table>

*Hitomebore sequence (DDBJ000499) was used as a reference.

Figure 4-4. The causal SNP for the Hit9188 early stage lethality mutant.

(a) Genomic location and structure of the gene, Os01t0127300-01, harboring a nucleotide change in the mutant Hit9188. (b) Amino acid sequence of the protein encoded by Os01t0127300-01 and the predicted amino acid change caused by the mutation in Hit9188. (c) DNA sequencing peak chromatograms of the region of Os01t0127300-01 gene harboring SNP showing the wild-type G in Hitomeore wild type (top), mutant A in mutant-bulk DNA (middle), and the G/A mixture in wild-type bulk DNA (bottom).
The Os01g0127300 gene is predicted to encode a member of the SufBD protein and is homologous to the *Arabidopsis thaliana* NAP6 (*Non-intrinsic ABC protein 6*) gene encoding a SufD protein (Hjorth *et al.*, 2005). We thus named the gene and the protein as *OsNAP6* and OsNAP6, respectively. To determine whether the mutation detected in *OsNAP6* is responsible for the Hit9188 mutant phenotypes, we used RNA interference to knockdown the expression of *OsNAP6*. Accordingly, we cloned a 300-bp sequence spanning the last exon and 3′-UTR of *OsNAP6* into pANDA vector (Miki and Shimamoto 2004) generating pANDA-OsNAP6 (Figure 4-5a). Callus of a rice cultivar Hitomebore was transformed with pANDA-OsNAP6 by Agrobacterium tumefaciens. The rice transgenic lines (RNAi-OsNAP6) with reduced expression of *OsNAP6* as determined by quantitative RT-PCR (Figure 4-5b) exhibited extreme dwarfism, pale green leaves and early death, which is similar to phenotypes of the Hit9188 mutant (Figure 4-5c). From this result, we concluded that the mutation detected in *OsNAP6* is responsible for the Hit9188 developmental phenotypes.

*OsNAP6* shares considerable similarity with proteins from multiple plant species and the mutated alanine residue, which is located within the predicted putative β-helix domain, is conserved (Figure 4-6) (Hjorth *et al.*, 2005). The *A. thaliana nap6* T-DNA insertion mutant shows various developmental phenotypes such as reduced chlorophyll content, shorter roots, a considerable proportion of abnormal embryos, and altered thylakoids (Hjorth *et al.*, 2005).
Figure 4-5. RNA interference of OsNAP6 corresponding to Os01t0127300-01 results in the developmental phenotype identical to the Hit9188 mutant. (a) Gene structure and a scheme of gene silencing construct used for RNAi of OsNAP6 gene (Os01t0127300-01) in rice. (b) Results of real-time quantitative reverse transcription (RT)-PCR showing the relative expression level of OsNAP6 in rice plants transformed with OsNAP6-RNAi construct (RNAi) and empty vector (Empty). Asterisks indicate significant differences (Student’s t-test, ** P < 0.01) (c) Phenotype of leaf blade (top) and seedlings (bottom) of OsNAP6 RNAi transgenic plant as compared to the Hit9188 mutant.
Figure 4-6.
Alignment of NAP6 homologs from multiple plant species.

Identical and similar amino acids are indicated in black and gray backgrounds, respectively. The position of the mutated residue in Hit9188 is given by the red box.
Seedlings with reduced chlorophyll content that eventually die out from about three weeks after sowing also characterize the Hit9188 mutant reported here (Figure 4-2). The SUF (mobilization of sulfur) system, which contains six proteins (SufA, SufB, SufC, SufD, SufS and SufE), is one of the three systems involved in the biosynthesis, assembly, maturation and repair of iron-sulfur (Fe-S) cluster that is involved in numerous important biological processes (Xu and Moller 2004; Fontecave et al., 2005; Wollers et al., 2010). We further applied MutMap+ to a second mutant Hit11440 that is characterized by albino seedlings that eventually die from about two weeks after germination (Figure 4-7). M3 seeds harvested from a heterozygous M2 individual identified among 10 M2 plants were used to establish 104 M3 progeny that segregated 81 (wild-type) to 23 (mutant), confirming to 3:1 ratio (Chi square test: $\chi^2 = 0.462$, ns). For MutMap+ application, we prepared two DNA bulks; the first bulk composed of 20 M3 mutant albino progeny and the second bulk composed of 20 wild-type M3 progeny mixed in equal proportion and sequenced to a depth of 12 and 11.4 fold coverage, respectively (Table 4-1). SNP-indices were calculated for both bulks and plotted for all the 12 rice chromosomes. Comparison of SNP-index plots of the two bulks as well as generation and further evaluation of $\Delta$(SNP-index) plots identified a 4.34 Mb region (0.32 Mb - 4.66 Mb) on chromosome 8 that is unique to the mutant bulk (Figure 4-8).
Figure 4-8.
Figure 4-8.
(c) \(\Delta \text{(SNP-index)}\) plot obtained by subtraction of Wt-bulk SNP-index from Mt-bulk SNP-index

Figure 4-8.
We assessed the SNPs within the candidate region on chromosome 8 in detail and identified 10 SNPs with SNP index of 1 (Table 4-3). Of these, only a single SNP causing a G (Hitomebore) to A (Hit11440) transition, which corresponded to a C to T substitution due to opposite orientation of the gene, at the 2,178,176th nucleotide position was non-synonymous. The SNP introduced a premature stop codon at the 137th residue (CAG-Gln to TAG-Stop) of Os08g0139100, a gene predicted to encode a protein similar to the chloroplast precursor DAG protein (Figure 4-9). The protein contains a predicted 27-aa chloroplast transit peptide at its N-terminus (http://www.cbs.dtu.dk/services/SignalP/) and shares considerable homology with DAG proteins from multiple plant species (Figure 4-10). The mutated glutamine residue is also highly conserved among the species considered. The DAG locus was previously isolated in Antirrhinum majus by transposon tagging, and the corresponding dag (differentiation and greening) mutant is characterized by a slow growth and variegated leaves (Chatterjee et al. 1996; Chatterjee and Martin 1997), and is among the many examples of variegation mutants caused by transposable element activity (Yu et al., 2007). The expression of DAG is reported to be essential for chloroplast development, and is required for the expression of nuclear genes such as CAB and RBCS that affect the chloroplast, as well as for the expression of the plastid gene RPBO (Chatterjee et al., 1996). Taken together, our finding and the already known function of DAG protein strongly suggest that the mutation identified in Os08g0139100 is the most likely candidate responsible for the albino and early death phenotypes of Hit11440.
Figure 4-9. The candidate gene for Hit11440.
(a) Genomic location and structure of the candidate gene, Os08t0139100-01, harboring a nucleotide change in the Hit11440 mutant. (b) The predicted 299-amino acid sequence of the protein encoded by Os08t0139100-01. The mutated Glutamine (Q) residue is indicated in red. (c) Sanger sequencing confirms the candidate SNP identified by illumina whole genome re-sequencing. Peak chromatograms showing the region of Os08t0139100-01 gene harboring the candidate SNP. The wild-type C in Hitomeore wild type (top), the mutated T in mutant-bulk DNA (middle), and the C/T mixture in wild-type bulk DNA (bottom) are indicted by the black arrow.
Figure 4-10. Alignment of DAG proteins from multiple plant species.

The predicted 27-aa chloroplast residue in rice is indicated by orange line, and the highly conserved mutated Glutamine (Q) residue in Hit11440 is indicated with the red box. NCBI reference numbers of the sequences used: *Oryza sativa* (NP_001060965.1), *Brachypodium distachyon* (XP_003573360.1), *Hordeum vulgare* (BAJ99034.1), *Sorghum bicolore* (XP_002445021.1), *Zea mays* (ACN27869.1), *Arabidopsis thaliana* (AAM65001.1), *Medicago truncatula* (XP_002316698.1), *Lotus japonicus* (XP_002518590.1), and *Vitis vinifera* (XP_002283211.1).
Discussion

MutMap as well as MutMap-Gap can generally be extended to other crop plants provided that the prerequisite of one backcross to the parental line is met. Here we describe MutMap+, which we applied to rice mutants to rapidly identify the causal mutations by whole genome resequencing and comparison of SNP-index plots of two bulked DNA obtained from a segregating M3 progeny. Notably, these experiments did not involve crossing between the candidate mutants and wild-type plants, therefore providing new applications that were not possible with the original MutMap protocol.

The fact that no crossing is required is a significant advantage of MutMap+. Genetic analysis of early stage lethality or sterility genes has remained time consuming. For instance, heterozygous sibling of a mutant of interest is usually crossed to a wild-type plant of a distantly related cultivar to obtain F1, which is selfed to generate F2. If the F2 segregates for the phenotype under investigation, one could map the mutation with DNA markers. In this traditional scheme, at least a cross and selfing followed by linkage analysis with DNA markers is needed. The recently reported method of SNP-Ratio Mapping (SRM) (Lindner et al., 2012) allows rapid identification of lethal alleles with the help of next generation sequencing, but the method still requires crossing. MutMap+ circumvents the need for both crossing and linkage analysis, and therefore can be implemented over a shorter time span. We foresee that this technique can systematically be applied to rapidly identify the genes involved in the control of early plant development and fertility.

The application of MutMap+ is not restricted to genes associated with lethality and sterility as described above. This technique could also be widely used to identify genes involved in all types of agronomically important traits. The prerequisite is the availability of draft genome sequences to generate accurate SNP-index plots. The fact that MutMap+ allows gene isolation
without crossing has an additional important practical implication. There are many crop species for which efficient techniques of artificial crossing have not been established. For instance, in many millet species including foxtail millet (*Setaria italica*), fonio (*Digitaria exilis*) and tef (*Eragrostis tef*), it is not easy to carry out crosses due to small flower sizes and difficulty in emasculation. MutMap+ can be easily applied to these species that are important to agriculture in developing countries. In view of the projected progress in *de novo* sequencing technologies and the ever-decreasing sequencing costs, we believe that MutMap+ is poised to be applicable for the improvement orphan crops. The improvement of neglected crops species and varieties should contribute to world food security by enhancing locally adapted agricultural systems.

One of the key advantages of the MutMap approach is the ability to rapidly map quantitative traits in crop genomes, a limiting feature in many breeding programs. This advantage is conferred by the fact that the crosses are made to the wild-type plant with only the EMS-derived SNPs (on average 1,500 per genome in our rice mutant population) used in genetic mapping. MutMap+ offers the same advantage as the classical MutMap protocol. Because it is based on selfing, it enables precise and robust phenotyping of minor effect traits.
Chapter 5

General Discussion
To meet the expanding food demands of the rapidly growing world population, accelerating the breeding for traits of agronomic importance including yields, quality and stress tolerance is vital. As described in this study, QTL-seq as well as the MutMap derivatives MutMap-Gap and MutMap+ are powerful tools to rapidly identify genomic regions harboring agronomically important traits. The most important feature common to the NGS-based methods presented in this PhD study is the fact that the development of DNA marker is not a prerequisite for linkage analysis, as whole genome sequencing allows the utilization of all SNPs distributed over the entire genome as DNA markers.

By circumventing the need for DNA marker development for mapping, the application of NGS-based methods provide breeders with two advantages over traditional linkage analysis associated with map-based cloning and QTL analysis. First, it reduces the cost, time and labor needed for developing DNA markers as well as for genotyping of mapping progeny. In QTL-seq, MutMap-Gap and MutMap+, only one or two bulked DNA samples prepared from segregating population are applied to whole genome sequencing. In rice, the crop used as an example for application of the methods developed in this study, enough sequencing depth (10× coverage of the genome) for the analysis of 7 samples could be secured by the illumina GAIIx platform with an output of 35-50 Gb sequence data in 10 days. The latest Illumine sequencing platform, Hiseq2500, has an output 500-600 Gb sequence data in 10 days, opening the
possibility that QTL-seq as well as MutMap-Gap and MutMap+ can be applied to crops with larger genomes such as maize, sorghum, soybean, barley, and wheat.

Second, QTL-seq, MutMap-Gap and MutMap+ could be applied to identify the genomic region harboring alleles/mutations causing subtle phenotypic changes, as they allow crosses to be made between closely related cultivars/lines. When progeny resulting from distance crosses are used for mapping, minor alterations in quantitative traits or phenotypes cannot be adequately measured because the large number of QTLs difference between the mapping parents could affect the target trait directly and/or indirectly. For example, QTL-seq successfully identified the QTL for amylose content of rice grain as described in Chapter 2 (Figure 2-8). In this experiment, I had to optimize the growing conditions to obtain a uniform growth until harvest. This is because amylose content is affected by different plant factors including heading date, plant height, photosynthesis, etc. If a progeny from distance crosses is used in this experiment, accurate evaluation of amylose content is not easy because it would be difficult to optimize the many growth factors that might affect grain quality.

QTL-seq is a versatile technique that can be potentially applied to improve the efficiency and speed of the commonly known breeding techniques as well. So far, the analysis of complicated quantitative traits, such as the amylose content example briefly discussed above, is done using chromosome
segment substitution lines (CSSLs), in which a particular chromosomal segment from a donor line is substituted into the genetic background of a single recurrent line, leading to the substituted segments covering all chromosomes. However, generation of CSSLs requires cost, time and effort, and DNA markers are needed for selection of each line harboring a particular substituted segment. Our QTL-seq method circumvents the need for the development of CSSLs by using progeny obtained from crosses made between genetically closely related cultivar/lines. Such progeny allows the observation of clear phenotypes by minimalizing the affect of other QTLs that might directly and/or indirectly affect the phenotype of interest.

QTL pyramiding based on MAS is another useful strategy for efficient crop breeding (Servin et al., 2004). In this approach, QTLs of interest are combined by crossing NILs harboring the QTLs in a common genetic background. As Ashikari et al. (2005) demonstrated in breeding the rice cultivar harboring two desirable QTLs for grain number and plant height, pyramiding breeding allows the immediate addition of desirable QTLs from NILs into the target cultivar/line without introgression of undesirable. But, the main bottleneck with the speed by which NILs that can be used for pyramiding are developed is the identification of QTLs controlling agronomically important traits, the step that can now be readily accelerated by QTL-seq.

When breeding crops for biotic/abiotic stress tolerance, the use of naturally occurring variations may be more attractive than artificial mutagenized
genes. This is because natural variations are often the result of selection by artificial and/or natural environment changes, making them invaluable resources to improve traits related to biotic and abiotic stresses. The QTL-seq method I established in this study provides a powerful tool whereby the candidate genes for a given phenotype are readily delineated to less than 2-Mb region, allowing the application of MAS.

Until recently, bacterial artificial chromosome (BAC)-based physical maps generated by Sanger sequencing have been predominantly used to construct reference genomes of model species such as rice and Arabidopsis (Arabidopsis Genome Initiative, 2000; International Rice Genome Sequencing Project IRGSP, 2005). Because of the cost and time associated with this approach, the sequencing of other economically important crop species with bigger genomes has been a slow process. Consequently, only the model plants of rice and Arabidopsis have been used for identifying genes and determining their molecular function in plants. The recent advances in next generation sequencing technologies have however dramatically expanded the number of species that can be sequenced, leading to a gradual increase of the species for which whole genome sequences are now available (Bevan and Uauy, 2013). This has made genomes of an increasingly more number of species accessible for analysis using the NGS-based methods described here. Accordingly, I hope that our QTL-seq, as well MutMap and its derivatives MtMap-Gap and MutMap+ can be readily applied to more species for identifying novel genes and
accelerating crop breeding. As the application of these methods is extended to more species and used under different experimental conditions, I believe that technical improvements to the methods can be made to ensure their maximum utility.

At Iwate Biotechnology Research Center (IBRC), I am currently expanding the utility of our NGS-based methods for accelerating breeding of rice. To this end, I recently initiated a project for the application of MutMap to breeding for salinity tolerance. Soil salinity is one of the major abiotic stresses reducing agricultural production. Currently 45 million ha, corresponding to 20% of irrigated land worldwide, is affected by salt accumulation (Munns et al., 2008). Although soil salinity is not a major limiting factor to crop production in Japan because of the availability of enough rainfall, it has now become an issue in some regions following the major earthquake and tsunami of March 2011 that devastated vast coastal areas of Northern Japan leading to the death or missing of almost 19,000 people and contamination of about ~20,000 ha of rice paddy field by seawater (Chagué-Goff et al., 2012, Kang et al., 2013).

To provide assistance to rice growers who have to deal with paddy fields inundated by seawater, I initiated a breeding program to develop rice cultivars with improved salinity tolerance. Accordingly, I first screened our EMS-mutant lines in the elite northern Japan cultivar Hitomebore background for salinity tolerance, and identified a mutant line that showed enhanced tolerance to salt. I have successfully identified the causative mutation by
MutMap. And I’m currently using the candidate mutation for MAS to remove the unnecessary mutations. A detailed functional analysis of the isolated gene is now in progress to understand the mechanism of salt tolerance in rice. Furthermore, the performance of the mutant under the normal field conditions is being tested. The successful isolation of the gene for salt tolerance is just one example among our ongoing research programs that illustrates the significance of our NGS-based methods for accelerated crop breeding.
Chapter 6

General Conclusion
In this study, I reported the development of whole-genome sequencing-based methods, namely QTL-seq, MutMap-Gap, and MutMap+, which allow the rapid identification QTLs and genes controlling important traits in crops.

QTL-seq allows the rapid identification of QTLs based on the sequencing of two bulked DNA samples obtained from the progeny showing extreme values for the phenotype under consideration in a segregating population. Compared with the conventional QTL analysis that depends on DNA marker development for linkage analysis that entails cost and labor, QTL-seq drastically reduces the cost, time and labor work for QTL mapping.

The MutMap-Gap and MutMap+ methods reported here have been developed to address the limitation of MutMap and broaden the application of this NGS-based method. Because the reference sequence used in MutMap is developed based the publically available genome sequence of the species being studied, the method fails to identify mutations that are located in the genomic region missing from the reference genome. This is the often the case when the variety or cultivar being analyzed shows significant structural variations from the one used for developing the reference genome for the species. MutMap-Gap addresses this limitation by the combination of MutMap analysis and de novo assembly of gap regions. Consequently, MutMap-Gap could be applied to even the mutation located within the parental line specific genomic regions.
For the application of both the MutMap and MutMap-Gap methods for genome analysis, we have to first make crosses between the mutant of interest and the parental line used for mutagenesis. As a result, these methods are not suited for the identification of mutations that do not allow artificial crossing. MutMap+ extend the availability of MutMap by comparing the SNP-index values generated for two bulked DNAs obtained from mutant and wild type progeny of M3 generation derived from selfing of an M2 heterozygous individual. Accordingly, MutMap+ doesn’t involve artificial crossing between mutants and the wild type parental line. Therefore, MutMap+ is suitable for the identification of mutations that result either in early development lethality, sterility, or generally hamper crossing. Both the MutMap based novel methods should broaden the opportunity to use WGS for gene isolation in the various species and mutant phenotype.

Due to the recent advances in next generation sequencing technologies, there has been a significant decline in the cost of whole genome sequencing. In parallel, draft sequences of a large number of organisms including those of economically important crop species are accumulating rapidly. For example, despite its large genome size, the barely genome sequence has been made public recently (International Barley Genome Sequencing Consortium, 2012). It is therefore expected that whole genome sequence of more and more species will be accessed more easily and routinely in the near future regardless of their genome size and complexity. Once whole genome sequence became available,
QTL-seq, MutMap-Gap and MutMap+ should make significant contributions for rapid identification of QTL and gene in such species.

I envisage that the methods described in this study will contribute to forward genetic approach, and identify QTLs and genes accelerate the efficient breeding. To this end, the pipelines for the QTL-seq, MutMap, MutMap-Gap, and MutMap+ methods are now made available for the public at our web site (http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap).
References


Lindner H, Raissig MT, Sailer C, Shimosato-Asano H, Bruggmann R and


Acknowledgements

本研究をまとめるにあたり、岩手大学大学院連合農学研究科教授（公益財団法人 岩手生物工学研究センター研究部長）寺内良平博士には、懇切丁寧、かつ熱い御指導、御鞭撻を終始賜りました。心よりの深謝を申し上げます。副査を引き受けてくださった 岩手大学農学部附属寒冷バイオフロンティア研究センター教授 上村松生博士、弘前大学農学生命科学部教授 石川隆二博士には貴重な御指導と御助言を賜りました。また、帯広畜産大学地域環境学研究部門教授 加藤清明博士には貴重な御指導と御助言を賜りました。総合研究大学院大学先導科学研究科准教授 印南秀樹博士には QTL-seq 法の理論にかかる統計的な解析を賜りました。厚くお礼申し上げます。

岩手県農業研究センターの皆様には、試験および調査の実施にあたり多大なる御協力をいただきました。作物研究室 菅原浩視氏には、圃場作業において有意義な御指摘、御助言を賜りました。皆様に感謝の意を表します。

また、試験の遂行にあたり多大なる御指導、御助言、御協力を賜りました岩手生物工学研究センターの皆様に感謝の意を表します。特に本論文をまとめるにあたり多くの御助言を賜りました Tamiru Mulu 博士には厚く御礼申し上げます。

平成 26 年 3 月 高木 宏樹