Gaurav Sablok · Sunil Kumar
Saneyoshi Ueno · Jimmy Kuo
Claudio Varotto *Editors*

# Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches

Springer

# Chapter 3
# Whole Genome Sequencing to Identify Genes and QTL in Rice

**Ryohei Terauchi, Akira Abe, Hiroki Takagi, Muluneh Tamiru, Rym Fekih, Satoshi Natsume, Hiroki Yaegashi, Shunichi Kosugi, Hiroyuki Kanzaki, Hideo Matsumura, Hiromasa Saitoh, Kentaro Yoshida, Liliana Cano, and Sophien Kamoun**

## Overview of Genetic Analysis for Identifying Genes

One of the major purposes of genetic analysis is to infer an alteration in the genetic material that is responsible for the phenotypic changes observed in an organism under study. This has been routinely addressed by genetic association studies. For example, let's assume that we have a population of individuals segregating in two phenotypic variants, and that our interest is to identify the gene responsible for this phenotypic difference. For this purpose, we first divide the population into two phenotypic groups and then look for the genetic variation that shows statistically significant association with the groups. Since genes are arranged linearly on chromosomes, two loci that are physically close to each other are more likely inherited together, whereas distantly located loci tend to be inherited independently due to recombination occurring between the two loci (Bateson et al. 1905; Morgan 1910; Lobo and Shaw 2008). Therefore, once an association is identified between a genetic variation and the phenotype under investigation, we infer that physical location of the causative gene controlling the phenotype is close (linked) to the identified genetic variation. A common practice is to first identify as many genetic variations

R. Terauchi, Ph.D. (✉) • A. Abe, Ph.D. • H. Takagi, Ph.D. • M. Tamiru, Ph.D.
R. Fekih, Ph.D. • S. Natsume • H. Yaegashi • H. Kanzaki, Ph.D. • H. Saitoh, Ph.D.
Division of Genomics and Breeding, Iwate Biotechnology Research Center,
Narita 22-174-4, Kitakami 024-0003, Iwate, Japan
e-mail: terauchi@ibrc.or.jp

S. Kosugi, Ph.D.
Kazusa DNA Research Institute, Kisarazu, Chiba, Japan

H. Matsumura, Ph.D.
Gene Research Center, Shinshu University, Ueda, Nagano, Japan

K. Yoshida, Ph.D. • L. Cano, Ph.D. • S. Kamoun, Ph.D.
The Sainsbury Laboratory, Norwich Research Park, Norwich, Norfork, UK



G. Sablok et al. (eds.), *Advances in the Understanding of Biological Sciences Using Next Generation Sequencing (NGS) Approaches*,
DOI 10.1007/978-3-319-17157-9_3

segregating among the individuals of the study, and use these variations as "genetic markers" to test their association with the phenotype. Following identification of genetic markers that show association with a phenotype, we explore their vicinity to identify the very genetic change that is responsible for the phenotypic variation.

Two major approaches have been largely employed in genetic association studies. The first is applied to progeny derived from a cross between known parents; therefore, it is most widely used for gene isolation from crop species that are amenable to artificial crossing. Typically, crossing of two inbred parental lines results in F1, which is self-fertilized to generate F2 progeny. Using a large number of F2 progeny segregating for a particular phenotype, the association between the phenotype and genetic markers is examined. This approach addresses linkage (co-segregation) of phenotypes and markers from the parents to progeny, thus is usually called "linkage study." The second genetic association approach does not involve crossing, and is applied to a population of individuals with unknown relationships to each other. This approach is commonly called "association study," and whole genome association study (WGAS) has been widely used for gene identification in humans and other organisms. In WGAS, the population is divided into "case" and "control" groups to reveal markers that are associated with the "case"/"control" dichotomy. Each approach has its advantages and disadvantages. Linkage analysis is usually carried out over two generations (parents–offspring). As a result, the number of recombination occurring between the two generations is limited. In contrast, association analysis depends on individuals whose common ancestor traces back a large number of generations, ensuring that the number of recombinations among the individuals under study is large. The difference in the number of recombinations affects performance of the analysis. Owing to a higher level of linkage disequilibrium (LD), linkage analysis can be powerful in finding an approximate position of a causative gene using a small number of markers, but requires additional efforts in identifying the causative gene itself. On the other hand, low levels of LD make association analysis less suited for inferring the approximate position of causative genes, but it is more powerful in identifying the causative gene, provided that a large number of markers are available. Consequently, the combination of linkage analysis and association study has proved powerful as the two approaches complement each other.

## Genetic Markers to Become Obsolete?

For linkage analysis and association study, the availability of a large number of genetic markers is a prerequisite for successful analysis. As a result, researchers in the field of genetics have devoted considerable amount of time and resources over the years to develop such markers (Avise 1994). Development of DNA technology in the 1970s enabled the use of Restriction Fragment Length Polymorphisms (RFLP) markers. This was followed by the invention of Polymerase Chain Reaction (PCR) and discovery of ubiquitous distribution of di- tri-nucleotide simple sequence repeats (SSR) in eukaryotic genomes, which allowed the application of highly

variable markers called SSR or microsatellite markers. Development and advances in automated DNA sequencing technology enabled the identification of unlimited numbers of single nucleotide polymorphisms (SNPs) in the genome that can be used as markers called SNP markers. Continuous efforts have been made to generate large number of genetic markers covering the entire genome, while at the same time trying to make their scoring easier and cheaper. Nevertheless, the available genetic markers still represent a small proportion of the entire genetic code of an organism, forcing researchers to make inferences about properties of a whole genome based on the markers that are basically a limited number of sample points selected from the genome. However, thanks to the recent development of WGS technology, this situation is beginning to change, and would hopefully free researchers from their dependence on classical genetic markers.

## Rice Genetic Resources at IBRC

In order to fully exploit genomics approaches for efficient crop improvement, the availability of suitable biological materials is essential. Genetic materials representing a wide genetic variation of the species under study should be carefully generated, maintained, and scored for their phenotypes. At the Iwate Biotechnology Research Center (IBRC), we have been generating and maintaining two sets of rice genetic resources to accelerate improvement of elite rice cultivars. The first set of materials include ethylmethanesulfonate (EMS)-induced mutant lines (Rakshit et al. 2010). We treated immature embryo of flowers of a Northern Japan elite rice (*Oryza sativa* spp. *japonica*) cultivar "Hitomebore" with 0.75 % EMS. The matured seeds were planted to generate M1 individuals, which were self-fertilized to obtain M2 seeds. The M2 plants were further self-fertilized to obtain M3 and subsequent generations, and we are currently maintaining seeds of over 12,000 mutant lines at M3–M5 generations. These mutant lines show a wide range of phenotypic diversity, particularly of traits of agronomic importance. The second set of materials represent recombinant inbred lines (RILs) obtained by crossing of "Hitomebore" to 22 rice cultivars representing a wide genetic variation of *O. sativa*. We currently have a total of 3,172 RILs at the F5 to F7 generations. These resources are being used for isolation of important genes and QTL, as well as to develop WGS-based methods for accelerating crop breeding.

## MutMap

Using the EMS-mutant lines of "Hitomebore" rice cultivar, we set out to rapidly identify the causal mutation responsible for a given mutant phenotypic trait of agronomic importance. For this purpose, we developed the MutMap method (Abe et al. 2012). In MutMap, a mutant of interest is crossed to the parental line used for

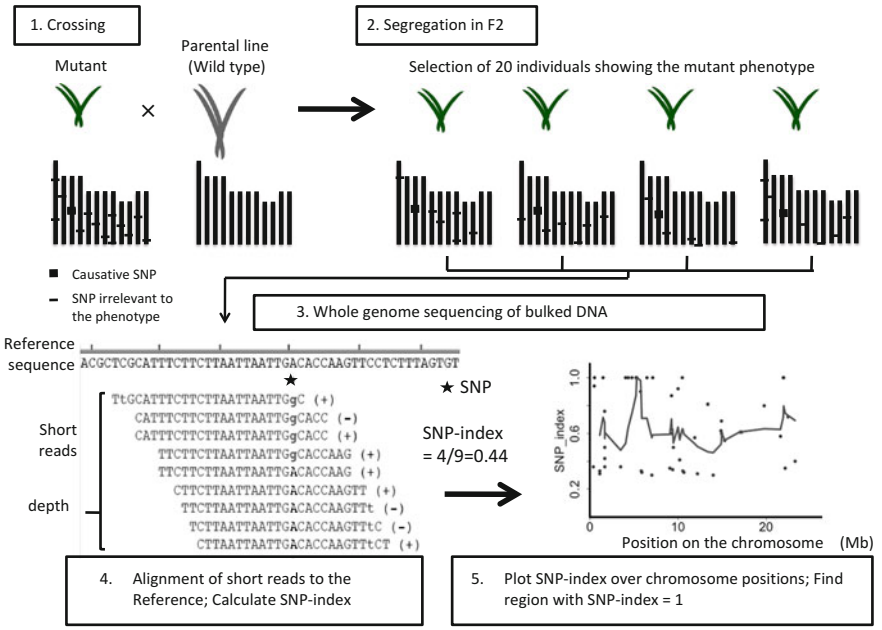


**Fig. 3.1** A simplified scheme of MutMap. *1*. A mutant showing a phenotype of interest is crossed to the parental line used for mutagenesis to obtain F2. Here we take a dwarf mutant caused by recessive mutation as an example. The mutant chromosomes with the mutations incorporated by mutagenesis and wild type individual chromosomes are shown. *2*. Among the F2 progeny segregating for wild type and mutant phenotypes, we focus on the mutant F2 progeny. Mutant F2 individuals inherit the causative mutation in homozygous state whereas mutations that are irrelevant to the phenotype are inherited in 1:1 ratio. *3*. DNA of >20 mutant F2 progeny are bulked and sequenced by Illumina sequencer. *4*. The resulting short reads are aligned to the reference sequence of wild type parental line, and SNP-index is calculated. *5*. SNP-index plots are generated to visualize the relationship between SNP-index and chromosome position. Sliding window analysis is applied to see the average value of SNP-index in a given genome interval. A peak of SNP-index suggests the position of causative mutation responsible for the mutant phenotype

mutagenesis (Fig. 3.1). If the phenotype is caused by a single recessive mutation, the F2 progeny segregates 3:1 (wild type to mutant phenotypes). We extract DNA from >20 mutant F2 individuals, mix the DNA in equal proportion to make a DNA-bulk, which is subsequently sequenced by Illumina DNA sequencer to a depth of at least 10× coverage of the genome. Since the rice genome size is about 380 Mb, we usually generate about 5 Gb short reads for each DNA-bulk. The resulting short reads of 76–100 bp in size are aligned to the reference sequence of the parental line "Hitomebore."

To facilitate the identification of the causative mutation, we introduced the concept of "SNP-index," which is the ratio of short reads containing SNPs to the total reads covering a particular position of the genome. If all short reads covering a particular genomic position have an identical sequence to the reference, the SNP-index is 0.

By contrast, if all the short reads have an SNP different from the reference sequence, the SNP-index is 1. Since the causative SNP responsible for the mutant phenotype should be inherited by all the F2 mutant progeny in homozygous state, short reads of bulked DNA corresponding to such an SNP should have SNP-index = 1, whereas SNPs not relevant to the phenotype should segregate 1:1 among the progeny, resulting in SNP-index = 0.5. The genomic region(s) tightly linked to the causative SNP are dragged with the causative mutation, thus the SNPs residing in such region should have SNP-index >0.5. Thus, if we graphically plot the relationships between SNP-index and chromosomal positions, we would observe a peak of SNP-index that is clustered around the position of the causative SNP. After locating the SNPs with SNP-index = 1, we scrutinize the genes harboring the SNPs, and identify the most likely candidate. Since WGS allows us to follow the inheritance of all the SNPs incorporated by mutagenesis, MutMap analysis does not require any genetic markers. Linkage of SNPs will allow us to visually identify the SNP-index peaks on a graph.

Practically MutMap requires only a total of around 100 F2 progeny to score the phenotype and a single WGS for identification of the causative mutation. Therefore, MutMap circumvents the time-consuming and laborious steps of conventional marker-based linkage analysis. Furthermore, MutMap is applied to F2 derived from the cross of a mutant of interest to the parental line used for mutagenesis. This procedure enables the identification of mutations that cause subtle quantitative changes of phenotype relevant to agronomic traits. This feature makes MutMap a more efficient method to isolate causative mutations with quantitative effects than SHOREmap (Shneeberger et al. 2009), another bulked DNA sequencing method, in which mutant lines are crossed to a distantly related line.

## MutMap+

As discussed above, MutMap application requires crossing to the wild type parental line. To address early death or sterile mutants that don't allow crossing, we recently developed MutMap+ (Fekih et al. 2013) as an extension of MutMap. MutMap+ involves identification of a causal mutation by comparing SNP-index plots of two DNA-bulks, mutant bulk and wild type bulk, obtained from a segregating M3 progeny that is derived from a self-fertilized heterozygous M2 individual. MutMap+ analysis starts by the identification of a phenotype of interest in a small number of segregating M2 individuals (about 10) obtained by selfing of each M1 line. Then, the wild type siblings of the identified mutants are left to self-fertilize and grow to maturity to generate M3 seeds. If the mutation is caused by a single recessive gene, two-third of the wild type M2 individuals are expected to harbor the causal mutation in heterozygous state. The M3 progeny established from seeds of these heterozygous individuals segregate 3:1 for wild type and mutant phenotypes. Accordingly, about 100 M3 individuals are grown separately for each M2 line, and for those that segregate for the phenotype of interest, we make two sets of DNA-bulks: one from ~20 mutant M3 progeny and the other from ~20 wild type M3 progeny, both of

which are derived from a single heterozygous M2 plant. The two DNA-bulks are separately sequenced, aligned to the reference, and SNP-index graphs are plotted as in MutMap analysis and compared to each other. A genomic region showing different patterns of SNP-index plots between the two bulks points to the location of the causal mutation differentiating the mutant from the wild type.

For each M2 individual, half of the mutations incorporated by mutagenesis are randomly fixed to homozygote state. Therefore, selfing of an M2 individual results in a large chromosome regions exhibiting SNP-index = 1 in M3 generation, making it difficult to locate the position of causative mutation. However, genomic regions exhibiting SNP-index = 1 by random fixation should be shared by all the M3 progeny, whereas the region showing SNP-index = 1 caused by bulking of mutant progeny is specific to the mutant DNA-bulk. We can thus identify location of the causative mutation by comparing SNP-index plots of the mutant and wild type bulks of M3 progeny. MutMap+ does not require crossing, so it is particularly useful for identifying mutations that cause early stage lethality and infertility thereby hampering crossing. This method is also applicable to crops that are not amenable to artificial crossing. MutMap+ is also potentially useful to isolate dominant mutations. If the mutation is dominant, the expected SNP-index value for mutant type M3 is 0.66 whereas that for the wild type is 0.

## MutMap-Gap

To apply MutMap and MutMap+ for the identification of candidate genes, we need a reference sequence of the parental line. In most cases, the cultivars used for mutagenesis are not the same as the ones for which an accurate reference genome is publicly available. In rice, a highly accurate genome sequence is available for a cultivar "Nipponbare" (International Rice Genome Sequencing Project 2005), but not for the cultivar "Hitomebore" used in our studies. Therefore, we generated a pseudo-reference sequence of the cultivar "Hitomebore," by first obtaining short sequence reads of a "Hitomebore" wild type plant and aligning them to the reference genome of "Nipponbare." After identifying all the SNPs between "Hitomebore" and "Nipponbare," the nucleotides of "Nipponbare" were replaced by those of "Hitomebore" at all the sites of SNPs to make the "Hitomebore" reference sequence. Consequently, this "Hitomebore" reference sequence is useful to identify mutations residing in the genomic regions that are conserved between "Hitomebore" and "Nipponbare." However, if the mutation of interest resides in the genomic region present in "Hitomebore" but absent from "Nipponbare," we cannot identify such mutations by using this pseudo-reference sequence. To solve this problem, we developed a method we named MutMap-Gap, which combines MutMap and local de novo assembly (Takagi et al. 2013a).

For MutMap-Gap analysis, we first apply MutMap to a "Hitomebore" mutant of interest, obtain an SNP-index plots, and identify a peak of SNP-index corresponding to the possible genomic region of the causative mutation. If after scrutiny of

candidate SNPs around the SNP-index peak region we cannot identify any promising SNPs in the "Hitomebore" pseudo-reference, this should prompt us to suspect that the causative mutation may reside in "Hitomebore"-specific region linked to the region identified by MutMap. To explore this possibility, we can apply local de novo assembly by using (1) short reads of wild type "Hitomebore" mapped to the genomic region corresponding to the SNP-index peak as well as (2) short reads of "Hitomebore" that could not be mapped (unmapped) to the "Nipponbare" reference. These unmapped reads are likely to be derived from "Hitomobore"-specific genomic region absent from "Nipponbare." After contigs are generated by the local de novo assembly, short reads of bulked DNA of mutant plants are aligned against the newly prepared reference sequence comprising "Hitomebore" pseudo-reference plus the newly generated contigs. This MutMap-Gap analysis may locate a contig(s) not present in "Nipponbare" that harbors an SNP(s) with SNP-index = 1. By applying MutMap-Gap to "Hitomebore," we successfully identified an SNP in the resistance (R-) gene *Pii* that confers resistance against rice blast fungus with *AVR-Pii*. The complete Hitomebore *Pii* gene region was not represented in "Nipponbare" genome, and it was recovered only by using MutMap-Gap. MutMap-Gap is particularly useful to identify mutations in highly variable genomic regions like *R*-gene clusters that are known to be rapidly evolving.

## QTL-Seq

The majority of agronomically important traits are controlled by multiple genes called quantitative trait loci (QTL) (Falconer and Mackay 1996) each with a relatively minor effect. Identification of QTL is an important task in plant breeding, and has been carried out mainly by linkage analysis. Following a cross of two distantly related varieties, F2 or RILs are generated, and their phenotype scored. Using genetic markers, association between trait values and marker genotypes are studied. Due to the necessity of large number of genetic markers, the two mapping parents are usually selected from genetically distantly related lines. However, such parents tend to have differences in multiple QTL, making isolation of individual QTL difficult.

QTL-seq (Takagi et al. 2013b) is a WGS-based method of QTL identification based on bulked-segregant analysis (Giovannoni et al. 1991; Michelmore et al. 1991; Mansur et al. 1993; Darvasi and Soller 1994). First, we cross two cultivars with different trait values and obtain the progeny of F2 generation or RILs (Fig. 3.2). Trait values are measured in the progeny. If the trait is controlled by multiple QTL, frequency distribution of trait values will be close to Normal (Gaussian) distribution. Here we focus on the individual lines that belong to the upper and lower tails of the distribution. We then bulk the DNA of the individuals belonging to the upper tail to make High bulk DNA. Similarly we bulk the DNA of the lower tail individuals to make Low bulk DNA. DNA extracted from High and Low bulks are separately subjected to WGS, and the resulting short reads are aligned to the reference sequence of either of the parental lines (e.g., Cultivar A). The SNP-index as defined

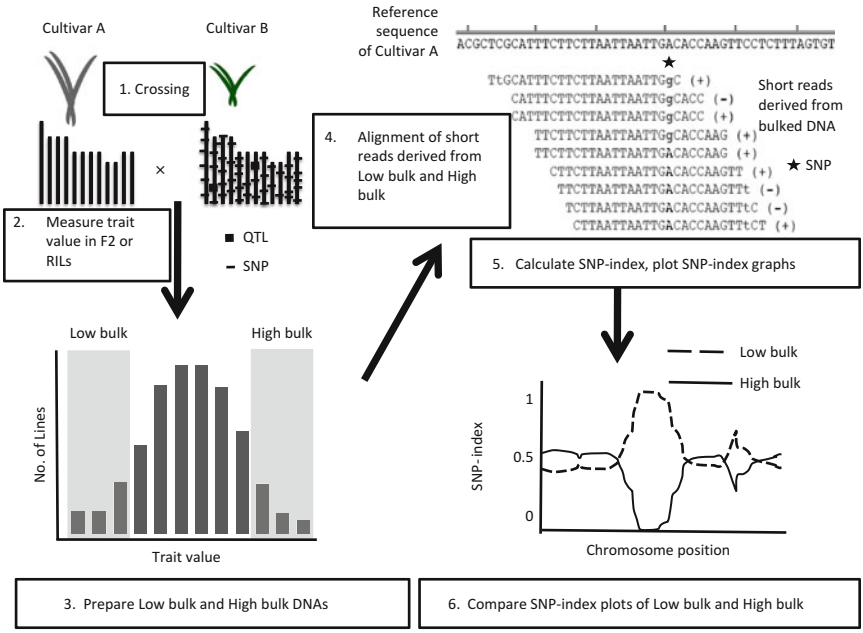


**Fig. 3.2** A simplified scheme of QTL-seq. *1*. A cross is made between Cultivars A and B that show contrasting difference for the trait of interest. The chromosomes of the two cultivars are shown, and Cultivar A is used as reference. *2*. We score the phenotype of the progeny (F2 or RILs) derived from the cross between Cultivar A and B. *3*. If the trait is controlled by multiple QTL, the frequency distribution of trait value follows a Normal (Gaussian) distribution. We focus on the upper and lower tails of the distribution, and make two DNA-bulks, one from the upper tail (High bulk) and the other from the lower tail (Low bulk). *4*. High bulk and Low bulk DNAs are separately sequenced, and aligned to the reference sequence of either of the parent used for the cross (in this case Cultivar A). *5*. SNP-index is calculated and the relationships between SNP-index and chromosome position is depicted separately for High bulk and Low bulk. *6*. SNP-index graphs are compared between High and Low bulks. Genome positions with contrasting patterns of SNP-index plots between High and Low bulks indicate the localization of QTL differentiating the two bulks

in MutMap is calculated for all the SNPs identified between the each DNA-bulk and the reference, and the relationships between SNP-index and chromosome position (SNP-plots) are depicted separately for the High and Low bulks. For most of the genomic regions, genomes of the two parents are inherited to the progeny in equal probability; therefore, SNP-index should be around 0.5. However, genomic regions harboring QTL responsible for the differentiation of trait values among the progeny should exhibit contrasting patterns of SNP-index plots between High and Low bulks, which is easily identified by comparing SNP-index plots of the two bulks. We applied QTL-seq to F2 and RILs of rice, and successfully identified positions of QTL for partial resistance to blast fungus and seedling vigor. QTL-seq rapidly allows identification of QTL by two whole genome sequencing. Since all the SNPs in the genome are used as "genetic markers," the method is also applicable to progeny derived from crosses between closely related cultivars.

## SNP-Index

In all the WGS-based methods described above, we used SNP-index for locating the candidate gene or genomic region. From the viewpoint of population genetics, SNP-index can be interpreted as a measure of nucleotide diversity of the population under study for a given genomic position. In the case of MutMap, the expected value of SNP-index in F2 population is 0.5, except for the genomic region harboring the causative mutation where SNP-index = 1. In bi-allelic situation, allele frequency of 0.5 corresponds to the highest genetic diversity, which is always reduced by deviating to 1 or 0. The SNP-index peak (SNP-index = 1) in MutMap can be viewed as a signature of selective sweep with reduced genetic diversity caused by selection of mutant type individuals in F2 population. Similarly, deviation of SNP-index values from 0.5 in QTL-seq is caused by selective sweep derived from phenotypic selection on the trait values (High and Low bulks). Therefore, we can reinterpret MutMap and QTL-seq as special applications of general methodology whereby SNP-index is used to identify genomic regions that underwent artificial selective sweeps. Note that MutMap and QTL are categorically "linkage studies" since we use progeny populations of F2 or RILs derived from known crosses. We expect that application of SNP-index-based method to "association study" will provide fruitful results not only in human but also in crop species in future.

## Summary

Here we introduce a suite of WGS-based methods of gene/QTL identification in plants. We demonstrated that these methods can be applied to crop plants by applying them to rice with a focus on the improvement of an elite cultivar of Northern Japan. We expect that the MutMap and QTL-seq methods to have wide applicability in plant breeding along with related methods of mapping-by-sequencing that have been primarily developed in *Arabidopsis* (James et al. 2013).

The analysis pipelines are publicly available in the URL links below:
http://genome-e.ibrc.or.jp/home/bioinformatics-team/mutmap

**Acknowledgements** This study was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences, the Ministry of Education, Cultures, Sports and Technology, Japan to HK and RT (Grant-in-Aid for Scientific Research on Innovative Areas 23113009) and JSPS KAKENHI to RT (Grant No. 24248004). We thank Shigeru Kuroda for general supports.

## References

Abe A, Kosugi S, Yoshida K, Natsume S, Takagi H, Kanzaki H, Matsumura H, Yoshida K, Mitsuoka C, Tamiru M, Innan H, Cano L, Kamoun S, Terauchi R (2012) Genome sequencing reveals agronomically-important loci in rice from mutant populations. Nat Biotechnol 30:174–178