nature biotechnology

Genome sequencing reveals agronomically important loci in rice using MutMap

Akira Abe[1,2,7], Shunichi Kosugi[3,7], Kentaro Yoshida[3], Satoshi Natsume[3], Hiroki Takagi[2,3], Hiroyuki Kanzaki[3], Hideo Matsumura[3,4], Kakoto Yoshida[3], Chikako Mitsuoka[3], Muluneh Tamiru[3], Hideki Innan[5], Liliana Cano[6], Sophien Kamoun[6] & Ryohei Terauchi[3]

The majority of agronomic traits are controlled by multiple genes that cause minor phenotypic effects, making the identification of these genes difficult. Here we introduce MutMap, a method based on whole-genome resequencing of pooled DNA from a segregating population of plants that show a useful phenotype. In MutMap, a mutant is crossed directly to the original wild-type line and then selfed, allowing unequivocal segregation in second filial generation (F_2) progeny of subtle phenotypic differences. This approach is particularly amenable to crop species because it minimizes the number of genetic crosses ($n = 1$ or 0) and mutant F_2 progeny that are required. We applied MutMap to seven mutants of a Japanese elite rice cultivar and identified the unique genomic positions most probable to harbor mutations causing pale green leaves and semidwarfism, an agronomically relevant trait. These results show that MutMap can accelerate the genetic improvement of rice and other crop plants.

The world population is predicted to reach 9 billion within the next 40 years, requiring a 70–100% increase in food production relative to current levels[1]. It is a major challenge to ensure sustainable food production without further expanding farmland and damaging the environment, in the midst of adverse conditions such as rapid climatic changes. Crop breeding is important for improving yield and tolerance to existing and emerging biotic and abiotic stresses[2]. However, current breeding approaches are mostly inefficient, and have not incorporated the findings of the genomics revolution[3].

Most crop traits relevant to agronomic improvement are controlled by several loci, including quantitative trait loci (QTL), that when disrupted lead to minor phenotypic effects[4]. To enable plant breeding by marker-assisted selection, it is important to identify the locus or chromosome region harboring each gene contributing to an improved trait. However, compared with genes with major effects that determine discrete characteristics, allelic substitutions at agronomic trait loci lead to only subtle changes in phenotype. Consequently, cloning genes that control agronomic traits is not straightforward.

Here we describe MutMap, a method of rapid gene isolation using a cross of the mutant to wild-type parental line, and apply it to a large population of mutant lines of an elite Japanese rice cultivar. We used MutMap to localize genomic positions of rice genes controlling agronomically important traits including semidwarfism. As mutant plants and associated molecular markers can be made available to plant breeders, this approach could markedly accelerate crop breeding and genetics.

RESULTS

MutMap method

We explain the principle of MutMap (**Fig. 1**) using the example of rice. We first use a mutagen (for example, ethyl methanesulfonate) to mutagenize a rice cultivar (X) that has a reference genome sequence. Mutagenized plants of this first mutant generation (M_1) are self-pollinated and brought to the second (M_2) or more advanced generations to make the mutated gene homozygous. Through observation of phenotypes in the M_2 lines or later generations, we identify recessive mutants with altered agronomically important traits such as plant height, tiller number and grain number per spike. Once the mutant is identified, it is crossed with the wild-type plant of cultivar X, the same cultivar used for mutagenesis. The resulting first filial generation (F_1) plant is self-pollinated, and the second generation (F_2) progeny (>100) are grown in the field for scoring the phenotype. Because these F_2 progeny are derived from a cross between the mutant and its parental wild-type plant, the number of segregating loci responsible for the phenotypic change is minimal, in most cases one, and thus segregation of phenotypes can be unequivocally observed even if the phenotypic difference is small. All the nucleotide changes incorporated into the mutant by mutagenesis are detected as single-nucleotide polymorphisms (SNPs) and insertion-deletions (indels) between mutant and wild type. Among the F_2 progeny, the majority of SNPs will segregate in a 1:1 mutant/wild type ratio. However, the SNP responsible for the change of phenotype is homozygous in the progeny showing the mutant phenotype. If we collect DNA samples from recessive mutant F_2 progeny and bulk sequence them with substantial genomic coverage

[1]Iwate Agricultural Research Center, Kitakami, Japan. [2]United Graduate School of Agricultural Sciences, Iwate University, Morioka, Japan. [3]Iwate Biotechnology Research Center, Kitakami, Japan. [4]Gene Research Center, Shinshu University, Ueda, Japan. [5]Graduate University for Advanced Studies, Hayama, Japan. [6]The Sainsbury Laboratory, Norwich Research Park, Norwich, UK. [7]These authors contributed equally to this work. Correspondence should be addressed to R.T. (terauchi@ibrc.or.jp).

Received 28 July 2011; accepted 14 December 2011; published online 22 January 2012; doi:10.1038/nbt.2095

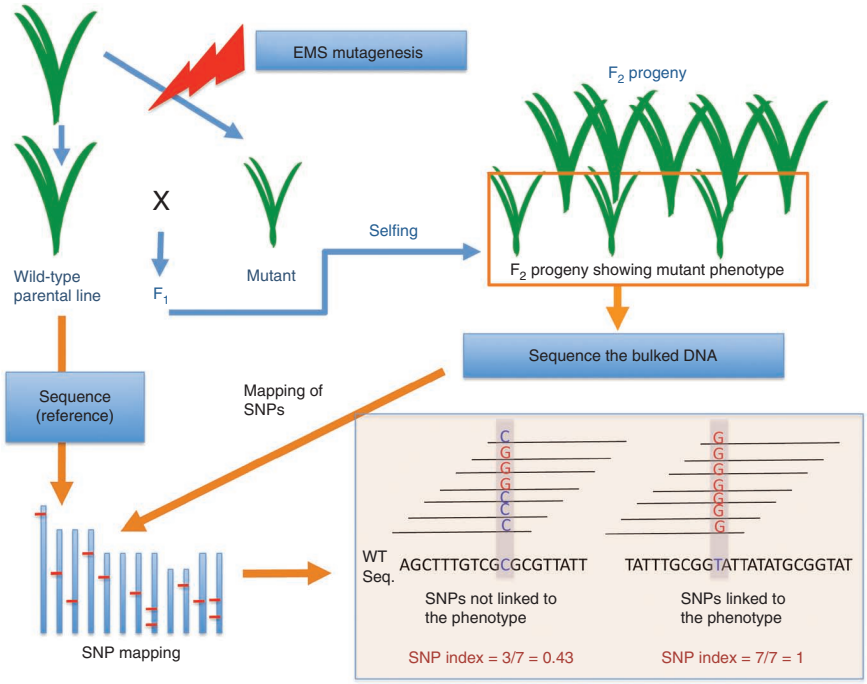

Figure 1 Simplified scheme for application of MutMap to rice. A rice cultivar with a reference genome sequence is mutagenized by ethyl methanesulfonate (EMS). The mutant generated, in this case a semidwarf phenotype, is crossed to the wild-type plant of the same cultivar used for the mutagenesis. The resulting F_1 is self-pollinated to obtain F_2 progeny segregating for the mutant and wild-type phenotypes. Crossing of the mutant to the wild-type parental line ensures detection of phenotypic differences at the F_2 generation between the mutant and wild type. DNA of F_2 displaying the mutant phenotype are bulked and subjected to whole-genome sequencing followed by alignment to the reference sequence. SNPs with sequence reads composed only of mutant sequences (SNP index of 1; see text) are closely linked to the causal SNP for the mutant phenotype.

(>10× coverage), we expect to have 50% mutant and 50% wild-type sequence reads for SNPs that are unlinked to the SNP responsible for the mutant phenotype. However, the causal SNP and closely linked SNPs should show 100% mutant and 0% wild-type reads. SNPs loosely linked to the causal mutation should have >50% mutant and <50% wild-type reads. If we define the SNP index as the ratio between the number of reads of a mutant SNP and the total number of reads corresponding to the SNP, we expect that this index would equal 1 near the causal gene and 0.5 for the unlinked loci. SNP indices can be scanned across the genome to find the region with a SNP index of 1, harboring the gene responsible for the mutant phenotype.

The rate of false positives can be assessed because allelic segregation follows a binomial distribution with a probability parameter of 0.5 (probability of mutant SNP equals 0.5) at a SNP with no linkage to the causal SNP. If the sample size (read depth of the site) is 10, the probability of having a SNP index of 1 is $P = (0.5)^{10} = 10^{-3}$. Therefore, in a data set with a known number of genotyped SNPs (L), the expected number of clusters of SNPs with SNP index of 1 ($\geq k$) would be approximately $p^k L = 10^{-3k} L$. In our case, the maximum estimate of L is 2,225 (**Supplementary Table 1**), and the probability of observing a cluster of more than four consecutive SNPs with SNP index of 1 would be $\leq 2.3 \times 10^{-9}$. Statistical considerations of how the number of F_2 progeny to be bulked and the average coverage (depth) of genome sequencing affect the false-positive rate, and how misclassification of phenotypes between mutant and wild type affects true positives, are in **Supplementary Data**.

MutMap applied to pale green leaf mutants

We have maintained 12,000 ethyl methanesulfonate–mutagenized rice lines of third and fourth mutant generations (M_3–M_4) with a background of Hitomebore, an elite cultivar of Northern Japan[5]. Whole-genome resequencing of five independent mutants indicated that each line harbors 1,499 ± 469 (mean ± s.d.; range 960–2,225) SNPs that are different from wild type (**Supplementary Table 1**). Using our mutant stock, we set out to isolate genes of agronomic importance. As a proof-of-principle experiment, we applied MutMap to two mutants showing pale-green leaf phenotypes with slightly lower chlorophyll concentrations compared with wild type (Hit1917-pl1 and Hit0813-pl2; **Fig. 2a**).

We crossed these mutants to the Hitomebore wild type in 2009, and obtained F_1 progeny. F_1 plants were self-pollinated, and >200 F_2 progeny were obtained for each cross. For both mutants we observed segregation between wild-type and mutant phenotypes in field-grown

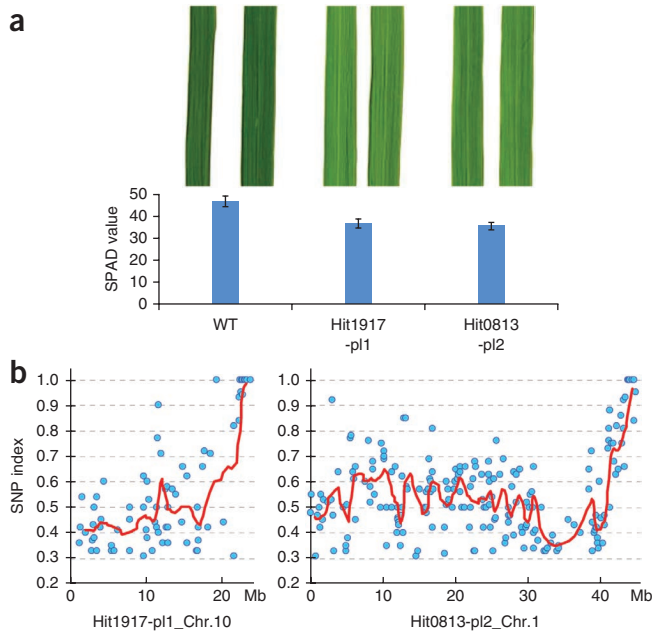

Figure 2 Identification of genomic regions harboring causal mutations for two pale green leaf mutants, Hit1917-pl1 and Hit0813-pl2, using MutMap. (**a**) Leaf color and SPAD (stability of soil plant analytical development) values (an estimate of chlorophyll content) of wild-type (WT) Hitomebore and two mutants. Error bars, s.d. (**b**) SNP index plots for two leaf color mutants (Hit1917-pl1 and Hit0813-pl2) showing chromosomes 10 and 1, respectively. Red regression lines were obtained by averaging SNP indices from a moving window of five consecutive SNPs and shifting the window one SNP at a time. The *x*-axis value of each averaged SNP index was set at a midpoint between the first and fifth SNP.

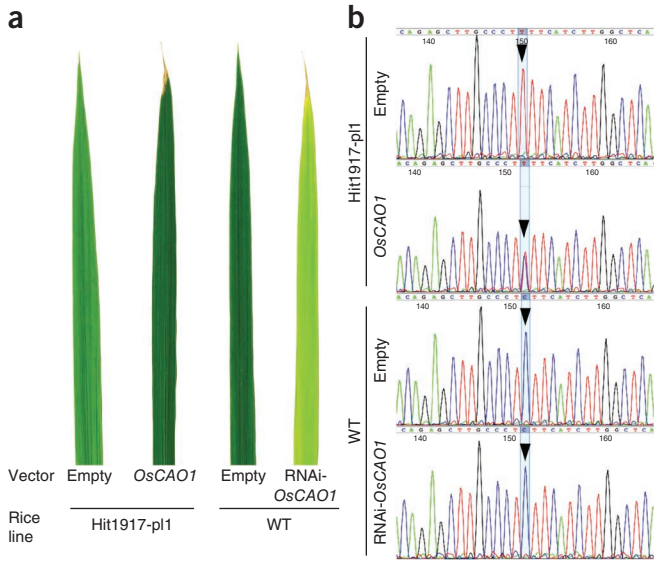

Figure 3 Genetic complementation of Hit1917-pl1 pale green leaf mutant with *OsCAO1*. (**a**) Rice lines and plasmid vectors used for transformation studies (bottom); phenotypes of four transgenic rice lines (top). (**b**). DNA sequencing peak chromatograms of *OsCAO1* cDNA close to site SNP-22981826 (arrowhead) obtained from four individuals in **a**.

F_2 progeny, with a wild type/mutant ratio of 3:1 (**Supplementary Table 2**), suggesting that each mutant phenotype was caused by a recessive mutation in a single locus. For each cross, we isolated DNA of 20 F_2 progeny showing the mutant phenotype, and bulked the samples in an equal ratio. This bulked DNA was subjected to whole-genome sequencing using an Illumina GAIIx sequencer. We obtained 70 million and 133 million sequence reads (75 bp) for Hit1917-pl1 and Hit0813-pl2, respectively, corresponding to >5 Gb of total read length with >12× coverage of the rice genome (370 Mb; **Supplementary Table 3**). These reads were aligned to a reference sequence of Hitomebore using MAQ software[6]. Aligned data were passed through a filter to reduce spurious SNP calls caused by sequencing and alignment errors (Online Methods). As a result, we found that mutants harbor 1,001 (Hit1917-pl1) and 1,339 (Hit0813-pl2) transition-type (G→A and C→T) SNPs with high-quality scores (**Supplementary Table 4**), presumably caused by ethyl methanesulfonate mutagenesis[7].

For each identified SNP, we obtained the SNP index, and we plotted SNP indices for all 12 chromosomes of rice (**Supplementary Fig. 1**). As we expected, SNP indices were distributed randomly around 0.5 for most parts of the genome for the two mutant lines. We were not surprised by the noise we observed in the dispersion of SNP indices given the many SNPs and the stochastic nature of allelic segregation at each individual SNP, which follows a binomial distribution with a probability parameter of 0.5 (**Supplementary Data**). For each mutant we identified a single unique genomic region harboring a cluster of SNPs with SNP index of 1 (**Fig. 2b**): the Hit1917-pl1 mutant showed a cluster of seven SNPs with SNP index of 1 on chromosome 10, whereas Hit0813-pl2 had a cluster of five SNPs with SNP index of 1 on chromosome 1 (**Fig. 2b**, **Supplementary Table 4** and **Supplementary Fig. 1**). These results show that MutMap allows rapid identification of the putative position of a causal mutation responsible for a mutant phenotype.

Identification of the causal SNP of a pale green leaf mutant

For Hit1917-pl1, a pale green leaf mutant, we examined SNPs with SNP index of 1 in detail. Of seven SNPs with SNP index of 1, two corresponded to exons of protein-coding genes. SNP-22981826 was localized to the gene encoding chlorophyllide *a* oxygenase (OsCAO1) leading to a L253F mutation (codon CTT→TTT). SNP-23202531 corresponded to a zinc finger–RING/FYVE/PHD-type domain protein, leading to an A106V mutation (GCA→GTA; **Supplementary Table 5**). In a study of a T-DNA insertion knockout of the *OsCAO1* gene[8], the *OsCAO1* mutant has lower chlorophyll than wild type, similar to our

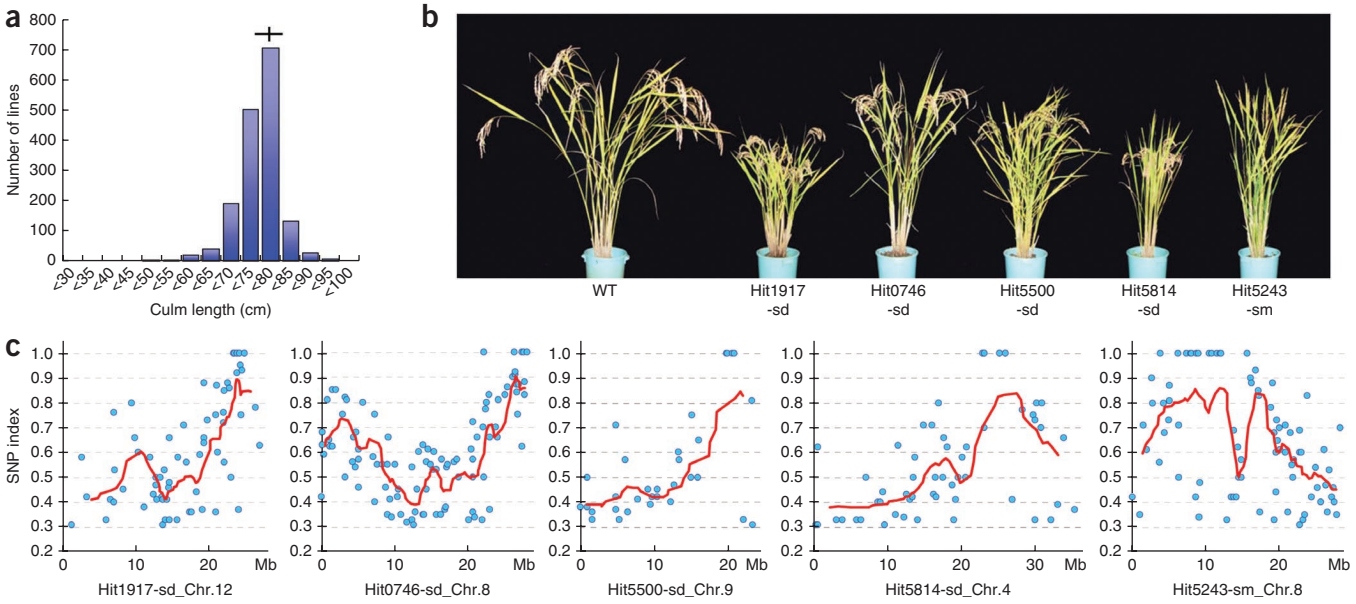

Figure 4 Identification of genomic regions possibly harboring causal mutations for five agronomically useful rice mutants using MutMap. (**a**) Distribution of culm length of 1,634 Hitomebore mutant lines. Vertical and horizontal lines above bars at ~80 cm indicate mean and s.d., respectively, of wild-type Hitomebore plants ($n = 29$). (**b**) Gross morphology of wild-type Hitomebore rice and five mutants (Hit1917-sd, Hit0746-sd, Hit5500-sd, Hit5814-sd and Hit5243-sm). (**c**) SNP index plots for five mutants.

Hit1917-pl1 mutant. Therefore, we hypothesized that Hit1917-pl1 is caused by a nonsynonymous substitution in the *OsCAO1* gene.

To verify this hypothesis, we carried out a complementation study by transforming the Hit1917-pl1 mutant with the wild-type *OsCAO1* gene driven by the native promoter. We also made a knockdown mutant of *OsCAO1* by transforming wild-type plants with an RNA interference (RNAi) construct targeting the *OsCAO1* gene (**Fig. 3**). The Hit1917-pl1 mutant transformed with wild-type *OsCAO1* expressed both mutant and wild-type alleles of *OsCAO1* (**Fig. 3b**), and its phenotype was restored to wild type (**Fig. 3a** and **Supplementary Fig. 2**). As we expected, the wild-type plant transformed with the RNAi construct for *OsCAO1* showed lower *OsCAO1* transcripts (**Supplementary Fig. 2**) and a paler green phenotype than Hit1917-pl1 (**Fig. 3a**). These data demonstrate that the Hit1917-pl1 phenotype is caused by the mutation SNP-22981826 identified by MutMap.

Application of MutMap to agronomically important traits

Our rice mutant lines showed wide variation in quantitative traits (**Supplementary Table 6**). In a subset of 1,634 M_4 lines, we measured seven traits of agronomic importance (culm length, leaf length, number of panicles per plant, panicle length, husk length, husk width and spikelet numbers per panicle). Of the seven traits, four (culm length, panicle length, husk length and husk width) had significantly greater variance (*F*-test, $P < 0.05$) in mutant lines compared with the wild-type parental Hitomebore line, indicating that our mutant lines contain greater genetic variation and are a good resource for isolating genes controlling agronomic traits (**Fig. 4a**).

In our breeding program, we are focusing on plant height (culm length) because a semidwarf phenotype leads to greater yield[9]. Therefore, we applied MutMap to four semidwarf mutants (Hit1917-sd, Hit0746-sd, Hit5500-sd and Hit5814-sd) to identify the genomic regions responsible for this phenotype. Additionally, a male sterility mutant (Hit5243-sm) was included in the analysis (**Fig. 4b**). Phenotypes of pale green leaf in Hit1917 (Hit1917-pl1) and semidwarfism in the same line (Hit1917-sd) are caused by two independent mutations in unlinked loci. All these phenotypes, except that of Hit5243-sm, are subtle, quantitative and difficult to score in F_2 when the mutants are crossed to an unrelated cultivar. With our MutMap approach, we observed F_2 segregation (**Supplementary Fig. 3**), and the segregation ratio suggested that a single recessive gene is involved in governing all the phenotypes (**Supplementary Table 2**). We bulk-sequenced 20 F_2 progeny showing mutant phenotypes, and scored SNP indices (**Supplementary Tables 3** and **4**; SNP-index plots, **Fig. 4c** and **Supplementary Fig. 1**). In all cases, only a single genomic region contained a cluster of SNPs with SNP index of 1. Sometimes multiple peaks (Hit0746-sd) or a broader peak (Hit5243-sm) of plots were observed. The former may be caused by an additional mutation affecting viability of F2 and the latter may be attributed to mutation position in recombination-deficient chromosome regions (that is, close to the centromere), but in most cases these complications did not prevent identification of putative regions harboring causal mutations. The average interval of SNPs within these regions with a SNP index ≥0.9 was 2.1 Mb, and we found at most four SNPs that could have caused nonsynonymous changes of protein-coding genes (**Supplementary Table 4**). We therefore conclude that these regions correspond to locations of the causal mutations responsible for the observed phenotypes (**Supplementary Table 5**).

DISCUSSION

We found that MutMap applied to rice can rapidly identify genomic regions harboring a causal mutation for a given phenotype. MutMap only requires crossing a mutant to the wild-type line used for mutagenesis followed by one subsequent selfing. As the mutant has been crossed back to its progenitor wild type, the F_2 progeny show unequivocal segregation between the mutant and wild-type phenotypes (**Supplementary Fig. 3**). This contrasts with conventional crossing schemes for gene isolation that involve crosses between genetically distant lines. In distant crosses, the parent lines differ in many genes so that segregation of particular phenotypes in F_2 follows a Gaussian and not a discrete distribution[4] (**Supplementary Fig. 4**). In addition, half of all the loci in F_2 are heterozygous and therefore heterosis strongly affects the phenotype. Because of these problems, isolation of genes with minor effects and QTL have been carried out using recombinant inbred lines (RILs) of later generations, in which the contribution of an individual gene is separately addressed and the effect of heterosis can be minimized. However, generation of RILs requires time and effort, and even advanced generations of RILs can be difficult to phenotype.

MutMap is technically similar to SHOREmap[10] and other related methods[11,12] such as NGM[11], which are based on bulked-segregant analysis of F_2 progeny[13,14]. In these methods, a cross is made between a mutant of one ecotype (for example, *Arabidopsis thaliana* Col) and a wild-type individual of a distantly related ecotoype (for example, *A. thaliana* Ler), and F_2 plants derived from such crosses are bulked and sequenced. Sequencing reads are aligned to the genome sequence of the wild-type parental line (Col) to search for regions with a high frequency of Col-type SNPs. That is, the SNPs that vary between distantly related ecotypes[10] (~1 SNP kb^{-1}) are used as DNA markers to locate the region harboring the causal mutation. In contrast, MutMap uses SNPs incorporated by mutagenesis as markers to look for the region harboring the mutation responsible for a given phenotype. Fewer SNPs need to be considered in MutMap (<2,300) than in the SHOREmap-NGM scheme (that is, 305,002 SNPs between Col and Ler; ftp://ftp.arabidopsis.org/Polymorphisms/Ecker_ler.homozygous_snp.txt), leading to reliable alignment between genome sequences and lower noise in SNP calling. An advantage of having many SNPs, as in SHOREmap-NGM, is that a clear cluster of SNPs with a high SNP index would be expected around the causal SNP. However, it would be difficult to pinpoint the causal SNP because of the many SNPs in the cluster. In MutMap, the causal SNP can readily be identified if the region has sufficient sequence coverage (**Supplementary Data**). In SHOREmap and NGM, only phenotypes with discrete characteristics can be unequivocally scored among the F_2 plants derived from distant crosses, and subtle alterations of quantitative traits cannot be adequately measured. Moreover, few F_2 progeny (for example, 20) are used for bulking in MutMap, whereas application of SHOREmap and NGM to *A. thaliana* has required more F_2 progeny (500 and >50, respectively[10,11]). In crop plants, growing so many F_2 progeny in the field can be impractical. The major differences between MutMap and SHOREmap-NGM are summarized in **Supplementary Table 7**.

MutMap also differs from the method of bulk sequencing of progeny derived from sequential backcrosses (more than four) of mutants to wild type, which has been applied to *Caenorhabditis elegans*[15] and *A. thaliana*[16]. In the latter method, backcrosses remove SNPs from the progeny, and only the regions harboring the causal mutation retain SNPs. Regions harboring the causal mutation for the phenotype are identified by scanning the genome for a region with high SNP density. MutMap is more practical for crops with long generation times as consecutive backcrosses are not required.

It is possible to further simplify MutMap for traits that are easily quantified in the field. Briefly, M_1 lines are generated by mutagenesis of seeds or embryos of an inbred line. Self-pollination of each M_1 plant leads to an F_2 generation that segregates for mutant and wild-type

phenotypes. If the M_1 line contains a recessive mutation (designated *a*) in the heterozygous state (*Aa*), there will be a high chance ($1 - [0.75]^{20} = 0.997$) of observing the mutant phenotype (*aa*) among 20 F_2 individuals. Thus, by growing and screening >20 F_2 individuals per line, one could obtain individuals showing the desired mutant phenotypes (*aa*) and their heterozygous siblings (*Aa*) with the wild-type phenotype. Bulked DNA of mutant progeny derived from selfing of many of the heterozygous siblings could serve the same function as that of mutant F_2 plants derived from the cross between a mutant and its wild type in the MutMap scheme. This method does not require any crossing, making it applicable to crops that are difficult to artificially cross.

In crop plants, identification of QTL has been conventionally addressed by mining natural genetic variation. However, the use of mutants for isolating QTL has been proposed[17]. We think the application of MutMap to a large mutant collection may yield much information relevant to QTL isolation. One could, for example, first identify a QTL likelihood interval by conventional QTL mapping, then search the MutMap database for mutants that have a related phenotype and that map within the QTL likelihood interval to identify candidate DNA changes that warrant further testing.

We anticipate that MutMap will facilitate gene isolation and breeding of crops by reducing the time and labor required for identifying agronomically important genes. As DNA sequencing is becoming easier and cheaper, the cost of identifying such genes could be markedly reduced. However, it is not necessary to identify a causal mutation to exploit MutMap in crop breeding. If a causal SNP cannot be identified, the SNPs flanking the regions harboring causal mutations for the desired phenotypes (those with a SNP index of 1) can be used as DNA markers for marker-assisted selection by crossing the mutant to the wild type. Moreover, if mutagenesis is done in an elite crop cultivar, as in the case of cultivar Hitomebore, then mutants and associated SNP markers can be made available to breeders to generate new varieties. As MutMap requires relatively deep sequencing coverage of the genome (>10×), its application to crops with very large genomes, such as maize, sorghum, soybean, barley and wheat, requires further investigation.

We hope to further demonstrate the power of MutMap soon. To this end, we have initiated a screen of cultivar Hitomebore mutants that can withstand high salinity. Once genes contributing to salt tolerance are identified, they will be used for developing rice cultivars suitable for cultivation in the ~20,000 ha of paddy fields of the Northern Japan coast that were flooded by the tsunami of 11 March 2011.

METHODS

Methods and any associated references are available in the online version of the paper at http://www.nature.com/naturebiotechnology/.

Accession code. DDBJ Sequence Read Archive: DRA000499. http://trace.ddbj.nig.ac.jp/DRASearch/submission?acc=DRA000499.

Note: Supplementary information is available on the Nature Biotechnology website.

ACKNOWLEDGMENTS

This study was supported by the Program for Promotion of Basic Research Activities for Innovative Biosciences, to R.T., H.I. and A.A. Ministry of Agriculture, Forestry and Fisheries of Japan (Genomics for Agricultural Innovation PMI-0010), Ministry of Education, Culture, Sports, Science and Technology of Japan (Grant-in-Aid for Scientific Research on Innovative Areas 23113009) to R.T. and a Daiwa Adrian Prize to S. Kamoun and R.T. L.C. and S. Kamoun were supported by the Gatsby Charitable Foundation. We thank S. Kuroda for general support and M.J. Terry, B. Wulff and K. Tsunewaki for suggestions to improve the paper.

AUTHOR CONTRIBUTIONS

A.A. conceived the idea and carried out rice crossing and phenotyping; S. Kosugi developed a bioinformatics pipeline to perform MutMap; K.Y., S.N. and L.C. carried out genome analysis; H.T. and K.Y. carried out rice transformation; H.K., H.M. and M.T. performed mutagenesis; C.M. carried out sequencing; H.I. performed theoretical analysis; S. Kamoun and R.T. conceived the idea, supervised the work and wrote the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/naturebiotechnology/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Godfray, H.C.J. *et al.* Food security: The challenge of feeding 9 billion people. *Science* **327**, 812–818 (2010).
2. Tester, M. & Langridge, P. Breeding technologies to increase crop production in a changing world. *Science* **327**, 818–822 (2010).
3. Baulcombe, D. Reaping benefits of crop research. *Science* **327**, 761 (2010).
4. Falconer, D.S. & Mackay, T.F.C. *Introduction to Quantitative Genetics* (Pearson/Prentice Hall, 1996).
5. Rakshit, S. *et al.* Use of TILLING for reverse and forward gentics of rice. in *The Handbook of Plant Mutation Screening: Mining of Natural and Induced Alleles.* (eds. Meksem, K. & Kahl, G.) 187–198 (Wiley-VCH, 2010).
6. Li, H., Ruan, J. & Durbin, R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.* **18**, 1851–1858 (2008).
7. Pienkowska, M., Glickman, B.W., Ferreira, A., Anderson, M. & Zielenska, M. Large-scale mutational analysis of EMS-induced mutation in the lacI gene of *Escherichia coli*. *Mutat. Res.* **288**, 123–131 (1993).
8. Lee, S. *et al.* Differential regulation of chlorophyll a oxygenase genes in rice. *Plant Mol. Biol.* **57**, 805–818 (2005).
9. Sasaki, A. *et al.* Green revolution: a mutant gibberellin-synthesis gene in rice. *Nature* **416**, 701–702 (2002).
10. Schneeberger, K. *et al.* SHOREmap: simultaneous mapping and mutation identification by deep sequencing. *Nat. Methods* **6**, 550–551 (2009).
11. Austin, R.S. *et al.* Next-generation mapping of *Arabidopsis* genes. *Plant J.* **67**, 715–725 (2011).
12. Uchida, N., Sakamoto, T., Kurata, T. & Tasaka, M. Identification of EMS-induced causal mutations in a non-reference *Arabidopsis thaliana* accession by whole genome sequencing. *Plant Cell Physiol.* **52**, 716–722 (2011).
13. Michelmore, R.W., Paran, I. & Kesseli, R.V. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA* **88**, 9828–9832 (1991).
14. Giovannoni, J.J. *et al.* Isolation of molecular markers from specific chromosomal intervals using DNA pools from existing mapping populations. *Nucleic Acids Res.* **19**, 6553–6558 (1991).
15. Zuryn, S., Le Gras, S., Jamet, K. & Jarriault, S. A strategy for direct mapping and identification of mutations by whole-genome sequencing. *Genetics* **186**, 427–430 (2010).
16. Ashelford, K. *et al.* Full genome re-sequencing reveals a novel circadian clock mutation in *Arabidopsis*. *Genome Biol.* **12**, R28 (2011).
17. Robertson, D.S. A possible technique for isolating genic DNA for quantitative traits in plants. *J. Theor. Biol.* **117**, 1–10 (1985).

ONLINE METHODS

Rice mutant lines. Immature embryos of cultivar Hitomebore were mutagenized by immersing panicles in 0.015% ethyl methanesulfonate solution (vol/vol) overnight (see ref. 5 for details). The resulting M_1 plants were self-pollinated, and M_2 seeds were obtained. M_2 plants were further self-pollinated to obtain M_3 progeny, leading to 12,000 M_3–M_4 lines.

Whole-genome sequencing of bulked DNA. DNA was extracted from 100 mg fresh rice leaves using the DNeasy Plant Mini Kit (QIAGEN Sciences). DNA was quantified using Quant-iT PicoGreen dsDNA reagent and kits (Invitrogen). To make bulked DNA from F_2 progeny, DNA from F_2 individuals was mixed in an equal ratio. Mixed DNA (5 µg) was used for preparation of libraries for Illumina sequencing according to the protocol for the Paired-End DNA Sample Prep kit (Illumina). The libraries were used for cluster generation on a flow cell and sequenced for 76 cycles on an Illumina Genome Analyzer IIx. Base calling and filtering of low-quality bases were done using sequence control software real-time analysis, Base calling (BCL) converter and the GERALD module (Illumina).

Alignment of short reads to reference sequences and SNP calling. To identify mutations incorporated by ethyl methanesulfonate, we generated a reference sequence of the Hitomebore wild-type genome on the basis of the publicly available Nipponbare rice genome sequence[18]. First, we obtained 1,083 million paired-end short reads from Hitomebore wild type and 11 mutant lines. These short reads were pooled and aligned with MAQ[6] to the Nipponbare reference sequence. Alignment files were converted to SAM or BAM files using SAMtools[19], and applied to a filter pipeline (S. Kosugi *et al.*, unpublished data) for identification of reliable SNPs. This filter pipeline was developed to maximize true SNP detection and minimize false SNP calling by (i) removal of paired-end reads of insert size >325 bp, (ii) calling SNPs only for genomic regions covered by a minimum of three reads for homozygous SNPs and five reads for heterozygous SNPs and a maximum of three-fold of average read depth over the genome, (iii) calling SNPs only on sites with an averaged Illumina phred-like quality score ≥20. This was further optimized by test data: short reads of Nipponbare experimentally obtained by Illumina GAIIx were aligned to the Nipponbare reference sequence containing 859,555 artificial nucleotide substitutions at known positions, leading to successful calling of 82% of true SNPs and false SNP calling of 0.1%. Using this pipeline, we identified 100,819 reliable SNPs between Hitomebore reads and the Nipponbare reference sequence. On the basis of this result, we generated a Hitomebore reference sequence (DDBJ Project ID67163) by replacing Nipponbare nucleotides with those of Hitomebore at 100,819 sites. To remove the effect of SNPs irrelevant to the mutant screen, we generated and used a reference sequence of the same wild-type Hitomebore line that was used for mutagenesis. We further refined this reference sequence by taking a consensus of cumulative genome sequences of the mutants.

Paired-end sequence reads of bulked DNA of mutant F_2 progeny were aligned to the Hitomebore reference sequence, and SNPs were scored. We divided SNPs into two categories: homozygous SNPs and heterozygous SNPs. Homozygous SNPs were defined as SNPs with SNP index ≥0.9 and a minimum coverage of the sites of three reads. Heterozygous SNPs were defined as SNPs with SNP index ≥0.3 and <0.9 with a coverage of the position of more than four reads. We further filtered SNPs with two more steps: (i) removal of common SNPs shared by at least two mutant lines and (ii) extraction of SNPs that exhibit G→A or C→T transitions, which are the most frequent changes caused by ethyl methanesulfonate mutagenesis. After identifying the genomic regions harboring a cluster of SNPs with SNP index of 1, we relaxed the condition of the filter to consider all SNPs (caused by all the transition and transversion) in the region as candidate SNPs for the causal mutation.

SNP index plot regression lines were obtained by averaging SNP indices from a moving window of five consecutive SNPs and shifting the window one SNP at a time. The *x*-axis value of each averaged SNP index was set at a midpoint between the first and fifth SNP.

Genetic complementation of Hit1917-pl1 mutant. For the complementation test, a 6,696-bp genomic fragment containing the *OsCAO1* gene 2 kb upstream and 1 kb downstream of the transcribed region was amplified from wild-type Hitomebore genomic DNA by PCR and subcloned into binary vector pGWB1 (ref. 20) to yield pGWB1-OsCAO1. pCAMBIA-empty was used as control in this study. For RNAi of the *OsCAO1* gene, a 337-bp *OsCAO1* partial fragment was amplified from wild-type Hitomebore genomic DNA by PCR and subcloned into binary vector pANDA[21] to yield pANDA-CAO1. All binary vectors were introduced into *Agrobacterium tumefaciens* strain EHA105 for rice transformation. Hitomebore plants were transformed as described[22].

18. International Rice Genome Sequencing Project. The map-based sequence of the rice genome. *Nature* **436**, 793–800 (2005).
19. Li, H. *et al.* The Sequence Alignment/Map (SAM) format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
20. Nakagawa, T. *et al.* Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. *J. Biosci. Bioeng.* **104**, 34–41 (2007).
21. Miki, D., Itoh, R. & Shimamoto, K. RNA silencing of single and multiple members in a gene family of rice. *Plant Physiol.* **138**, 1903–1913 (2005).
22. Toki, S. *et al.* Early infection of scutellum tissue with *Agrobacterium* allows high-speed transformation of rice. *Plant J.* **47**, 969–976 (2006).

doi:10.1038/nbt.2095