Plant Reproduction (2018) 31:117–128
https://doi.org/10.1007/s00497-018-0331-8

ORIGINAL ARTICLE

Development of the Mitsucal computer system to identify causal mutation with a high-throughput sequencer

Takamasa Suzuki[1,2,3] · Tsutae Kawai[3] · Shunsuke Takemura[4] · Marie Nishiwaki[4] · Toshiya Suzuki[4] · Kenzo Nakamura[3] · Sumie Ishiguro[4] · Tetsuya Higashiyama[1,2,5]

Received: 27 January 2018 / Accepted: 16 February 2018 / Published online: 1 March 2018

Key message **Development of Mitsucal.**

Abstract Recent advances in DNA sequencing technology have facilitated whole-genome sequencing of mutants and variants. However, the analyses of large sequence datasets using a computer remain more difficult than operating a sequencer. Forward genetic approach is powerful even in sexual reproduction to identify key genes. Therefore, we developed the Mitsucal computer system for identifying causal genes of mutants, using whole-genome sequence data. Mitsucal includes a user-friendly web interface to configure analysis variables, such as background and crossed strains. Other than configuration, users are only required to upload short reads. All results are presented through a web interface where users can easily obtain a short list of candidate mutations. In the present study, we present three examples of *Arabidopsis* mutants defective in sexual reproduction in which Mitsucal is used to identify causal mutation. One mutant was screened from seeds of a transgenic line with a reporter gene to elucidate the mechanisms involved in the regulation of seed oil storage. The identified gene codes for a protein may be involved in mRNA splicing. Other two mutants had defects in the surface walls on pollen termed exine. Both causal genes were identified, and mutants were found to be allele of known mutants. These results show that Mitsucal could facilitate identification of causal genes.

Keywords Whole-genome sequence · *Arabidopsis* mutant · Bioinformatics · Bulk segregant analysis

Introduction

Recent advances in DNA sequencing technology have sharply increased the outputs of single sequencing runs, while the cost has dramatically decreased (Faino and Thomma 2014). The first fully sequenced genome of plants was that of *Arabidopsis thaliana*, which was achieved during a project launched in 1996 and resulted in the elucidation of a sequence 120 Mb in size in 2000 (The Arabidopsis

Communicated by Mengxiang Sun.

A contribution to the special issue 'Plant Reproduction Research in Asia'.

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s00497-018-0331-8) contains supplementary material, which is available to authorized users.

✉ Takamasa Suzuki
takamasa@thaliana.myhome.cx

[1] Division of Biological Sciences, Graduate School of Science, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan

[2] Exploratory Research for Advanced Technology (ERATO), Higashiyama Live-Holonics Project, Furo-cho, Chikusa-ku, Nagoya 464-8602, Japan

[3] Department of Biological Chemistry, College of Bioscience and Biotechnology, Chubu University, 1200 Matsumoto-cho, Kasugai, Aichi 487-8501, Japan

[4] Laboratory of Biochemistry, Graduate School of Bio-agricultural Sciences, Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

[5] Institute of Transformative Bio-Molecules (WPI-ITbM), Nagoya University, Furo-cho, Chikusa-ku, Nagoya 464-8601, Japan

Genome Initiative 2000). Although even the smallest plant genome required several years and international collaboration to sequence, the recent single sequencer elucidates an entire genome in only a few days. The strategy adopted to sequence an entire genome was changed dramatically from combining bacterial artificial chromosome (BAC) clones covering the whole genome to the whole-genome shotgun method (Tabata et al. 2000; Faino and Thomma 2014). Since the completion of the reference sequence made it possible to sequence similar strains without de novo assembly (resequence), many strains were re-sequenced and comparative genomics was performed (Ossowski et al. 2008, Schneeberger et al. 2011).

The high performance of the recent sequencer is useful not only for genomics, but also for application to gene expression analysis (RNA sequencing; RNA-Seq) and the analysis of DNA–protein interaction by chromatin immunoprecipitation (ChIP)-Seq. (Robertson et al. 2007; Nagalakshmi et al. 2008). Research into genetics has also benefited from a high-throughput sequencer to facilitate the generation of a genetic map by restriction site-associated DNA (RAD)-Seq. (Davey et al. 2011; Kanamori et al. 2016) and the identification of the corresponding mutation for phenotypes of interest. The sequencing of the whole genome of multicellular organisms has become popular, even in a small laboratory or by a single researcher; however, the analysis of a large sequence dataset using a computer remains more difficult than the process of sequencing itself.

Not only reverse genetic approach for redundant gene families but forward genetic approach is powerful even in sexual plant reproduction to identify critical genes. For example, genes responsible for pollen tube guidance (e.g. Palanivelu et al. 2003; Chen et al. 2007; Shimizu et al. 2008; Dai et al. 2014), pollen tube reception (e.g. Escobar-Restrepo et al. 2007; Kessler et al. 2010) and gamete fusion (e.g. von Besser et al. 2006; Mori et al. 2014) were identified by forward genetic approach using mutants defective in plant reproduction. Identification of candidate responsible genes by a large sequence dataset must accelerate researches of sexual plant reproduction.

To overcome the difficulty associated in applying bioinformatics, we developed the Mitsucal computer system which is able to identify the causal gene from raw whole-genome sequence data. Mitsucal incorporates a web interface to configure analysis variables, such as the genetic background, crossed strain and type of mutagen. Therefore, the user can configure the program without any knowledge of bioinformatics. The results of analyses, including chromosome mapping and the list of detected base substitutions, are also represented through a web browser. The re-sequencing of a mutant genome and its analysis by Mitsucal drastically reduces the effort and time required to identify a causal gene. In the present article, we describe the workflow required to identify the corresponding mutation from a collected sample so as to list candidate mutations following the proof of a causal gene.

The first example, *defective repression of Ole3p::LUC1-1* (*drol1-1*) mutant, originated from an *Arabidopsis* transgenic line carrying luciferase reporter gene. The mutant *drol1-1* was backcrossed to the parental transgenic line, and the segregating F_2 was collected and sequenced. For the second and third examples, a pair of *kaonashi* (*kns*) mutants derived from a single mutagenised population were outcrossed to the no-originating strain, and tens of F_2 progenies of each cross were collected, from which DNA was extracted in bulk. Each DNA library was constructed and sequenced by a high-throughput sequencer. Mitsucal analysed the data pair immediately and successfully narrowed down a list of candidate mutations. These results indicated that Mitsucal is a powerful computer system that is useful for the identification of causal genes.

Results

Identification of the causal gene for *drol1-1* mutant isolated from a reporter transgenic line

We first present an example in which an *Arabidopsis* mutant was screened from ethylmethanesulfonate (EMS)-treated seeds of a transgenic line with a reporter gene, following which the Mitsucal computer system was used to identify the causal gene from whole-genome sequence data. To elucidate the mechanisms involved in the regulation of seed oil storage, a transgenic line of *A. thaliana* Col-0 containing the luciferase gene (*LUC*) placed downstream of a promoter of a gene for oleosin of seed oil body (*Ole3p::LUC*) was used. Similar to the endogenous levels of *Ole3* mRNA (Miquel et al. 2014), the highest levels of bioluminescence were observed during the late stage of seed development in the *Ole3p::LUC* plants. When *Ole3p::LUC* seeds were sown on plates containing luciferin, bioluminescence began to appear within 10–18 h, with a rapid increase until 36–48 h, after which the activity dropped sharply and disappeared after 4–5 days (Fig. 1). The increase in activity during the early stage of germination is at least partly due to de novo expression of *Ole3p::LUC*, since it is greatly diminished by the presence of an inhibitor of translation (cycloheximide) or transcription (cordycepin). A dramatic change in bioluminescence after 1–2 days of sowing suggests a transition in the transcriptional status of *Ole3* during the early stage of germination. To further understand the mechanisms involved in this transition, we screened for mutants in which the expression pattern of bioluminescence during germination is significantly altered from the parental line. One of the mutants, *drol1-1* (*defective repression of Ole3p::LUC1-1*),

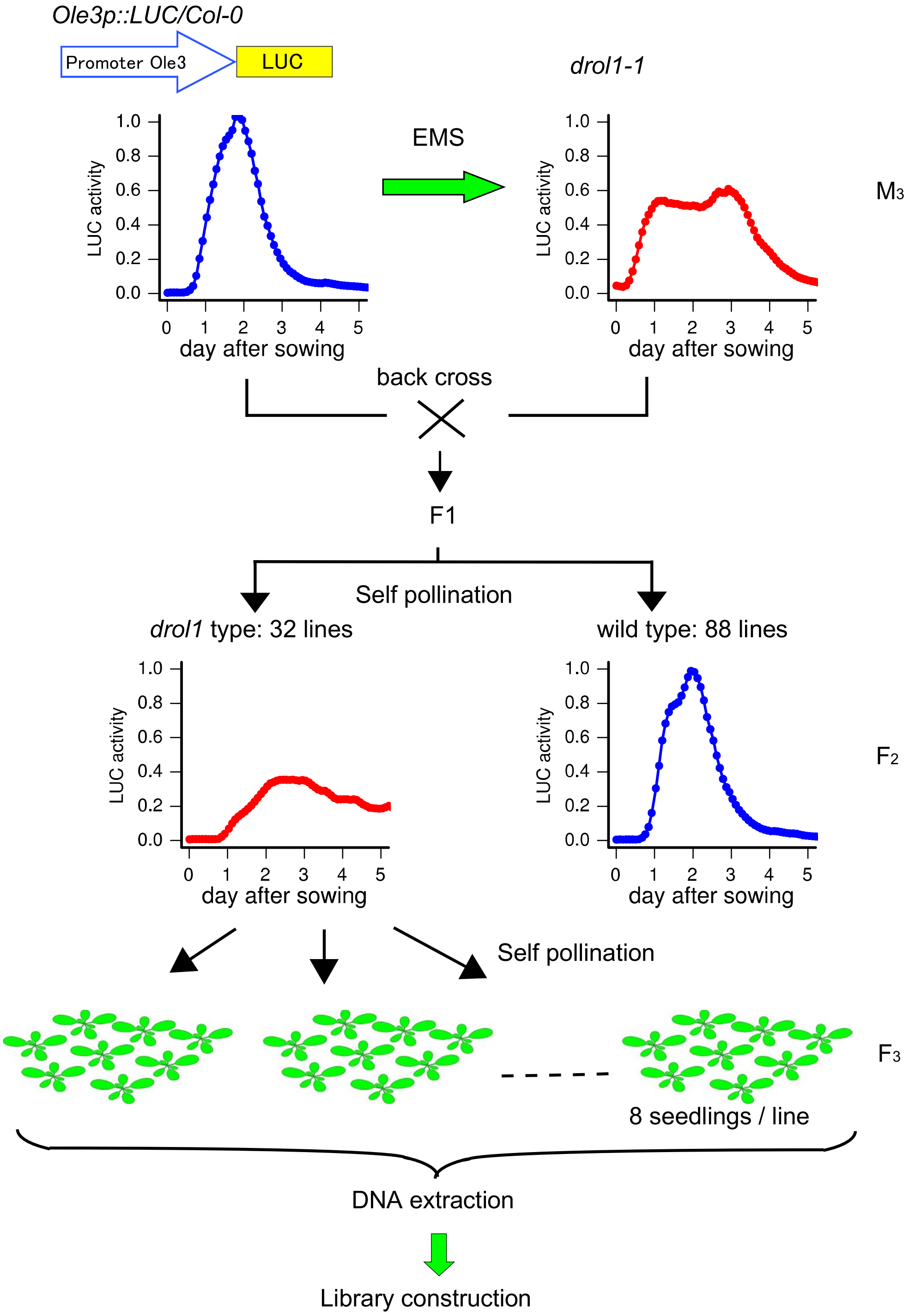


Fig. 1 Experimental scheme from *drol1-1* mutant isolation to library construction for whole-genome sequencing. The bioluminescence due to *Ole3p::LUC* in the Col-0 strain was monitored for 5 days after sowing on agar plates containing luciferin, and the relative levels of luminescence were plotted (top-left chart, $N = 72$). This parental line was mutagenised with ethylmethanesulphonate (EMS), and *drol1-1* was isolated from the M2 population. The relative levels of bioluminescence for *drol1-1* in the M3 generation were plotted in the top-right chart ($N = 96$). By crossing *drol1-1* to the parental line, F1 progenies were obtained and were self-pollinated to obtain F_2. Among the 120 F_2 progenies, 32 and 88 showed the phenotypes of *drol1-1* and the wild type, respectively. A part of their bioluminescence was plotted in the middle charts ($N = 15$ and 57 for *drol1-1* and wile type, respectively). F_2 progenies were self-pollinated to obtain F_3 seed, and eight seedlings of each were collected. The bulk DNA extracted from F_3 seedlings was used to construct the library for whole-genome sequencing

was selected for further analysis, and an additional mutant numbered #3–9 was used as a reference.

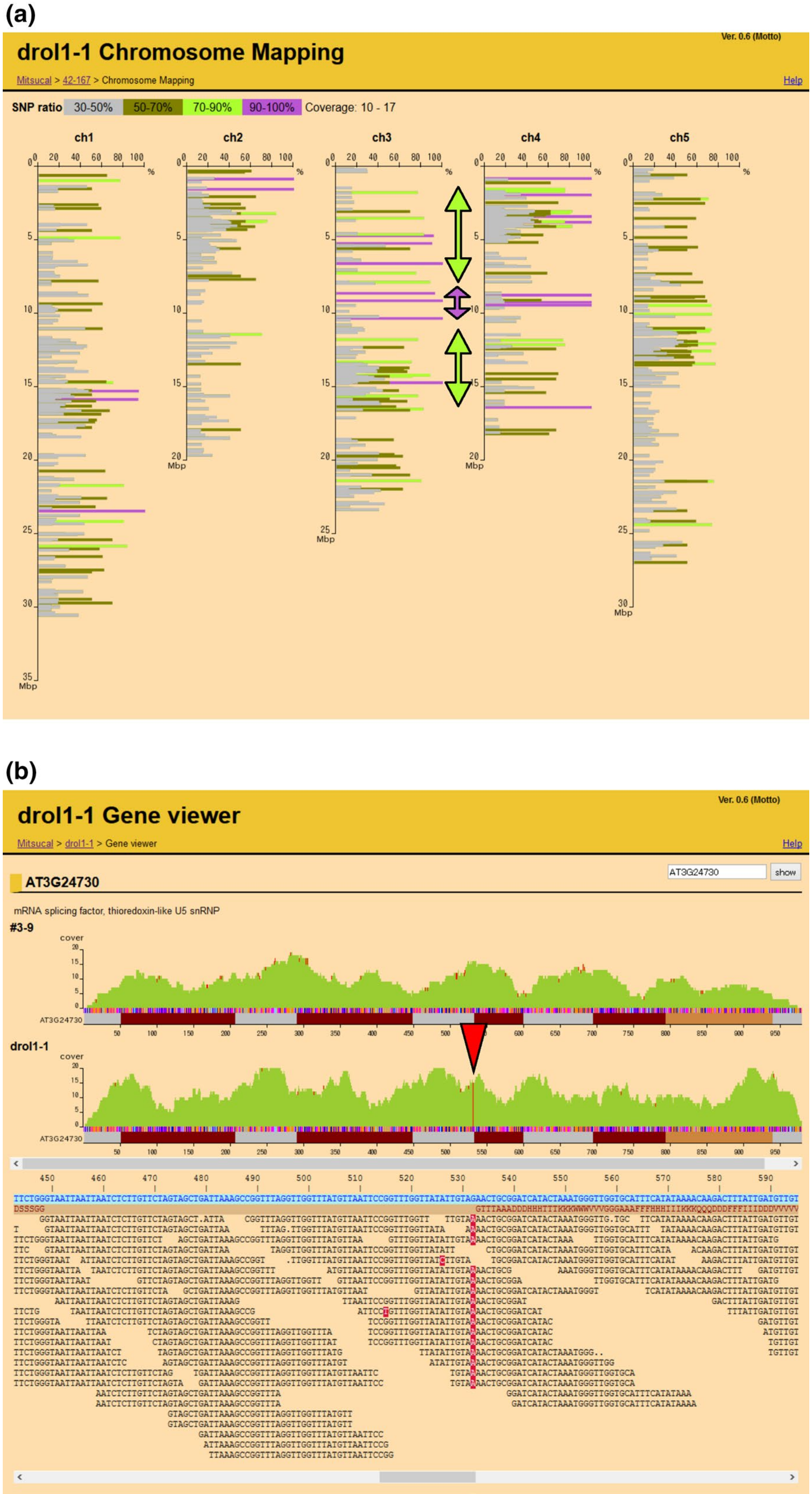
To identify the corresponding mutation on the genome responsible for the phenotypes of *drol1-1* mutant, *drol1-1* was backcrossed to the parental line *Ole3p::LUC* and the F_2 progenies were obtained. Among 120 F_2 progenies, 32 siblings showed defective repression of luciferase activity at 3 days after germination, which suggested that the phenotypes of *drol1-1* resulted from a single recessive mutation in the nuclear genome. Self-pollinated seeds harvested from each mutant F_2 plant were sown, and bulked seedlings were used for DNA extraction and subsequent library construction for sequencing. A total of 1.99 Gbp, approximately 17-fold the size of the *Arabidopsis* genome, was sequenced, and the obtained short reads were analysed by Mitsucal. The result of *drol1-1* by Mitsucal described below is presented at the following URL (http://dandelion.liveholonics.com/mitsucal/drol1/).

We have developed the Mitsucal computer system to process short reads produced by a high-throughput sequencer, namely the next-generation sequencer, to identify the corresponding mutation. Mitsucal first aligned all obtained 55 million short reads to the reference of the chromosome of the Col-0 strain, which was downloaded from the Arabidopsis Information Resource (TAIR) website (http://www.arabidopsis.org/), using bowtie (Langmead et al. 2009), and the uniquely aligned 37 million short reads (67%) were recovered. The aligned reads were compared to the reference sequence, and all mismatches were counted (Supplemental Fig. 1A). All mismatches were listed with the coverage, which was the number of short reads covering the nucleotide, and the ratio of mismatches. This list would contain true base substitutions in addition to sequence errors. To select a reliable marker (base substitution) for chromosome mapping, the mismatch list was sorted by degree of the coverage. The first 25% of mismatches from the lowest coverage were removed from the list, since their ratios of mismatches largely fluctuated by probability. The last 25% of mismatches was additionally removed from the list as they were speculated to be repetitive. In the case of *drol1-1*, mismatches with higher (> 17) and lower (< 10) coverage were removed from the list. The commonly observed mismatches between *drol1-1* and #3–9, which was an additional mutant line originating from the same parental line, were further removed from the list. The remaining mismatches were used as chromosome markers for further analysis. Mitsucal calculated the ratio of mismatch by a marker and constructed images for chromosome mapping (Fig. 2a). Each bar represented the ratio of mismatch and the marker position in the chromosome. They were coloured by the ratio of mismatch to be easily distinguished. Three purple bars indicating the highest ratio of mismatch (90–100% ratio) were observed in the 8–11-Mbp region of chromosome 3, and yellow green bars (70–90% ratio) were excluded from this region (purple double arrow in Fig. 2a). Because of the linkage among substitutions, it was speculated that the ratio of substitutions neighbouring causal mutations was higher than that of the proximal region. It was noteworthy that yellow green bars were neighbouring this region (yellow green double arrow in Fig. 2a). This result indicated that the corresponding mutation for the phenotype of *drol1-1* was located in this region (Supplemental Fig. 2A).

To identify the mutation responsible for the phenotypes of *drol1-1*, Mitsucal aligned all reads for *drol1-1* to the references of genes by bowtie with a parameter permitting multiple alignments. Using the same method as described above, all mismatches were detected. Within the 8–11-Mbp region of chromosome 3, 13 nearly homozygous mutations with a > 90% ratio of mismatch were detected. Four mutations were commonly observed in #3–9 as control, and five of the remainder had no effect on protein structure (mutation at intron or synonymous substitution). Only one mutation out of the remaining four was a substitution from G to A (or C to T), the most typical mutation caused by EMS. The mutation was located at the 3′ end of the second intron of the gene labelled AT3G24730 (Fig. 2b). To assess whether this mutation corresponded to the phenotype of *drol1-1*, *drol1-1* was crossed to a disruptant for AT3G24730 (designated *drol1-2*) obtained from the Arabidopsis Biological Resource Center (numbered SALK_087803, http://abrc.osu.edu/). The obtained F1 progenies, which had both a G to A mutation and a T-DNA insertion (hemizygous F1), displayed the extended expression of luciferase after germination (Fig. 3a). To assess whether expression of a cDNA for AT3G24730 could complement the phenotypes of *drol1-1*, we constructed a recombinant gene to express the cDNA for AT3G24730 under the control of cauliflower mosaic virus 35S promoter and introduced it into the *drol1-1* mutant. The obtained transgenic plant (35S::DROL1/*drol1-1*) showed an expression pattern of luciferase similar to the *Ole3p::LUC* parental line, but not to *drol1-1* (Fig. 3b). These results indicated that the mutation in AT3G24730 is responsible for the phenotypes of the *drol1-1* mutant. The *DROL1* codes for a protein may be involved in mRNA splicing. Further characterisation of the *drol1-1* mutant will be described elsewhere.

Identification of the corresponding mutations for two *kaonashi* mutants isolated from a mutagenised population of a Landsberg *erecta* strain

Pollen grains have light, irrefrangible and chemically stable surface walls termed 'exine'. The *Arabidopsis* exine shows a reticulate structure, most commonly observed in angiosperms (Fig. 4a, d). To determine how the reticulate structure is constructed, we attempted to identify the mutants



```
     450       460       470       480       490       500       510       520       530       540       550       560       570       580       590
      |         |         |         |         |         |         |         |         |         |         |         |         |         |         |
TTCTGGGTAATTAATTAATCTCTTGTTCTAGTAGCTGATTAAAGCCGGTTTAGGTTGGTTTATGTTAATTCCGGTTTGGTTATATTGTAGAACTGCGGATCATACTAAATGGGTTGGTGCATTTCATATAAAACAAGACTTTATTGATGTTGT
DSSSGG                                                                                    GTTTAAADDDHHHTTTKKKWWWVVVGGGAAAFFFHHHIIIKKKQQQDDDFFFIIIDDDVVVVV
     GGTAATTAATTAATCTCTTGTTCTAGTAGCT.ATTA    CGGTTTAGGTTGGTTTATGTTAATTCCGGTTTGGTT   TTGTAAAACTGCGGATCATACTAAATGGGTTG.TGC  TTCATATAAAACAAGACTTTATTGATGTTGT
T     GTAATTAATTAATCTCTTGTTCTAGTAGCTGATTAA      TTTAG.TTGGTTTATGTTAATTCCGGTTTGGTTATA    AAAACTGCGGATCATACTAAATGGGTTGGTGCATTT  TATAAAACAAGACTTTATTGATGTTGT
TTCTGGGTAATTAATTAATCTCTTGTTCT   AGCTGATTAAAGCCGGTTTAGGTTGGTTTATGTTAA     GTTTGGTTATATTGTAAAACTGCGGATCATACTAAA    TTGGTGCATTTCATATAAAACAAGACTTTATTGATG
TTC   GTAATTAATTAATCTCTTGTTCTAGTAGCTGATTAA        TAGGTTGGTTTATGTTAATTCCGGTTTGGTTATATT      CTGCGGATCATACTAAATGGGTTGGTGCATTTCATA    ACAAGACTTTATTGATGTTGT
TTCTGGGTAAT  ATTAATCTCTTGTTCTAGTAGCTGATTAAAGCCGGT    .TTGGTTTATGTTAATTCCGGTTTGGTTATCTTGTA    TGCGGATCATACTAAATGGGTTGGTGCATTTCATAT     AAGACTTTATTGATGTTGT
TTCTGGGTAATTA  TAATCTCTTGTTCTAGTAGCTGATTAAAGCCGGTTT          ATGTTAATTCCGGTTTGGTTATATTGTAAAACTGCG         AAATGGGTTGGTGCATTTCATATAAAACAAGACTTT   GATGTTGT
TTCTGGGTAATTAAT         GTTCTAGTAGCTGATTAAAGCCGGTTTAGGTTGGTT   GTTAATTCCGGTTTGGTTATATTGTAAAACTGCGGA              TTGGTGCATTTCATATAAAACAAGACTTTATTGATG
TTCTGGGTAATTAATTAATCTCTTGTTCTA   GCTGATTAAAGCCGGTTTAGGTTGGTTTATGTTAAT          GTTATATTGTAAAACTGCGGATCATACTAAATGGGT         TCATATAAAACAAGACTTTATTGATGTTGT
        AATTAATTAATCTCTTGTTCTAGTAGCTGATTAAAG                    TTAATTCCGGTTTGGTTATATTGTAAAACTGCGGAT                                    GACTTTATTGATGTTGT
TTCTG      TAATTAATCTCTTGTTCTAGTAGCTGATTAAAGCCG                    ATTCCTGTTTGGTTATATTGTAAAACTGCGGATCAT                                    TTTATTGATGTTGT
TTCTGGGTA     TTAATCTCTTGTTCTAGTAGCTGATTAAAGCCGGTT                   TCCGGTTTGGTTATATTGTAAAACTGCGGATCATAC                                        GATGTTGT
TTCTGGGTAATTAATTAA        TCTAGTAGCTGATTAAAGCCGGTTTAGGTTGGTTTA       TCCGGTTTGGTTATATTGTAAAACTGCGGATCATAC                                         ATGTTGT
TTCTGGGTAATTAATTAAT        CTAGTAGCTGATTAAAGCCGGTTTAGGTTGGTTTAT      TCCGGTTTGGTTATATTGTAAAACTGCGGATCATAC                                          TGTTGT
TTCTGGGTAATTAATTAATCT       TAGTAGCTGATTAAAGCCGGTTTAGGTTGGTTTATG               TTATATTGTAAAACTGCGGATCATACTAAATGGG..                                TGTTGT
TTCTGGGTAATTAATTAATCTC       AGTAGCTGATTAAAGCCGGTTTAGGTTGGTTTATGT                ATATTGTAAAACTGCGGATCATACTAAATGGGTTGG
TTCTGGGTAATTAATTAATCTCTTGTTCTAG    TGATTAAAGCCGGTTTAGGTTGGTTTATGTTAATTC              TGTAAAACTGCGGATCATACTAAATGGGTTGGTGCA
TTCTGGGTAATTAATTAATCTCTTGTTCTAGTA   GATTAAAGCCGGTTTAGGTTGGTTTATGTTAATTCC             TGTAAAACTGCGGATCATACTAAATGGGTTGGTGCA
                AATCTCTTGTTCTAGTAGCTGATTAAAGCCGGTTTA                                            GGATCATACTAAATGGGTTGGTGCATTTCATATAAA
                AATCTCTTGTTCTAGTAGCTGATTAAAGCCGGTTTA                                             GATCATACTAAATGGGTTGGTGCATTTCATATAAAA
                              GTAGCTGATTAAAGCCGGTTTAGGTTGGTTTATGTT
                              GTAGCTGATTAAAGCCGGTTTAGGTTGGTTTATGTT
                                    GATTAAAGCCGGTTTAGGTTGGTTTATGTTAATTCC
                                     ATTAAAGCCGGTTTAGGTTGGTTTATGTTAATTCCG
                                      TTAAAGCCGGTTTAGGTTGGTTTATGTTAATTCCGG
```

Fig. 2 Screenshots of Mitsucal for *drol1-1*. **a** Screenshot of chromosome mapping for *drol1-1*. Five images from ch1 to ch5 represented the chromosomes of *Arabidopsis*. The vertical and horizontal axes indicate the coordinate of the chromosome in Mb and the ratio of substitutions in per cent, respectively. The horizontal bars are coloured according to their substitution ratio (see also Supplemental Fig. 1). Three double arrows were added to indicate the regions mentioned in the main text. **b** Screenshot of gene viewer of Mitsucal. The images for piled reads aligned to AT3G24730 were drawn in the top (for #3-9 as the control) and middle (for *drol1-1*). The coverage and mismatches were drawn as height of the green area and red vertical line (see also Supplemental Fig. 1A). The colourful bars just below the green area represented the reference sequence, as shown in Supplemental Fig. 1A. The dark red, grey, and dark orange bars represented the intron, translated region and untranslated region, respectively. The red arrowhead was added to indicate the causal mutation for *drol1-1*. In the bottom part of the figure representing alignment of reads, the reference sequence (blue background) and amino acid sequence (brown character) of AT3G24730 were shown at the top two lines. The mismatches were indicated as white characterised in a red background

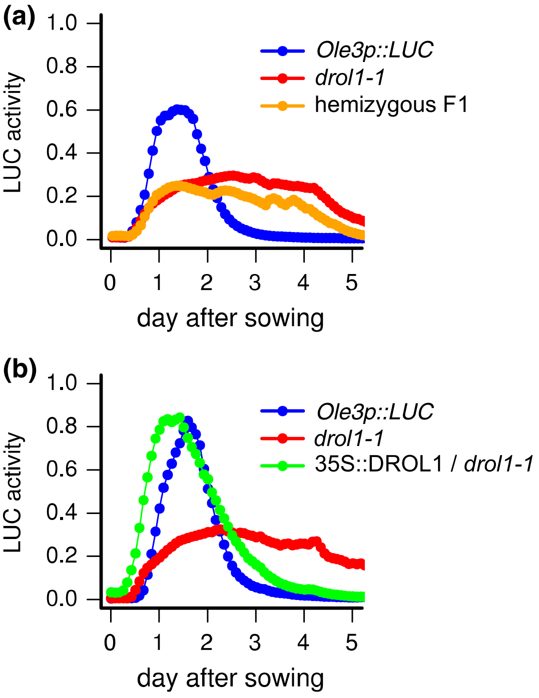


Fig. 3 **a** Allelism test between *drol1-1* and *drol1-2*. The relative expression level of *Ole3p::LUC* in the wild type (blue), *drol1-1* (red) and hemizygous F1 (orange) containing both *drol1-1* and *drol1-2* mutations was monitored for 5 days. The average intensity was plotted (*N* = 50, 50 and eight for the *Ole3p::LUC*, *drol1-1* and hemizygous, respectively). **b** Complementation test for *DROL1* gene. The cDNA for *DROL1* was designed to express under the control of 35S promoter, and the construct was introduced into *drol1-1* mutant. The bioluminescence caused by *Ole3p::LUC* was monitored for 5 days and plotted. The blue, red and green plots represent *Ole3p::LUC*, *drol1-1* and *drol1-1* complemented with 35S::DROL1, respectively (*N* = 1)

exhibiting abnormal pollen grain exine structures. By screening of an EMS mutagenised population of *Arabidopsis* Landsberg *erecta* (Ler) strain, we found many mutants named *kaonashi* (*kns*) (Suzuki et al. 2008). We chose *kns3-1* and *kns19-1* for determination of their causal genes by a high-throughput sequencer and Mitsucal computer system. A fine meshed exine is produced by *kns3-1* by increasing exine pillars (Fig. 4b, Suzuki et al. 2008), whereas *kns19-1* collapses exine structures almost completely (Supplemental Fig. 5B, Suzuki et al. 2008).

To identify the mutation responsible for the phenotypes, *kns3-1* and *kns19-1* were outcrossed to the *A. thaliana* Col-0 strain. Out of 111 F_2 progenies of a cross with *kns3-1*, 29 siblings displayed a phenotype similar to the original *kns3-1* mutant, suggesting that *kns3-1* was a single recessive mutation occurring in a nuclear genome. Self-pollinated seeds (F_3) of 24 F_2 plants were obtained, and ten F_3 seedlings of each F_2 line were harvested and combined for DNA extraction. Similarly, 22 siblings out of 139 F_2 progenies of a cross with *kns19-1* displayed a mutant phenotype, and their F3 progenies were used as a source of DNA extraction.

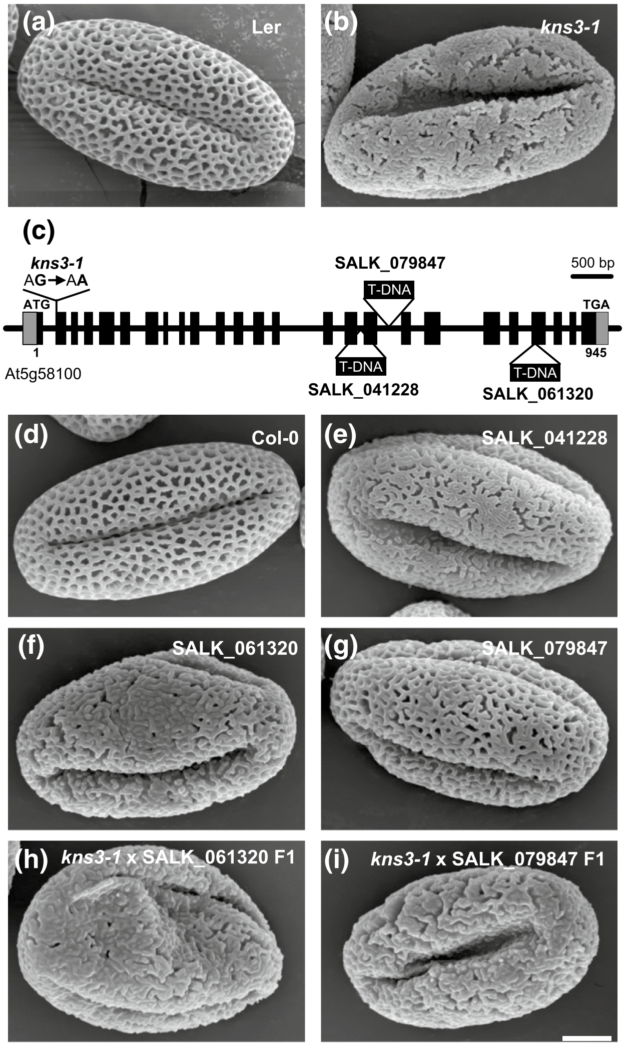


Fig. 4 Identification of *KNS3* gene. **a**, **b** Exine structures of pollen grains observed by scanning electron microscopy. **a** The WT Ler, which is the background strain of *kns3-1*. **b** *kns3-1*. **c** Structure of the *KNS3* gene. Boxes represent exons, and black and grey colours indicate coding and non-coding regions, respectively. The positions of initiation and termination codons are indicated, together with numbers of the first and the last amino acid residues. The point mutation in *kns3-1* and the T-DNA insertions in previously reported *spot1* alleles are shown. **d**–**i** Exine structures of T-DNA insertion alleles of At5g58100 gene. **d** The WT Col-0, which is the background strain of T-DNA insertion alleles. **e**–**g** Three T-DNA insertion alleles previously reported as *spot1*. **h**, **i** Exine structures of F1 progenies of *kns3-1* x *spot1* crosses showing the identity between *KNS3* and At5g58100. Bar = 5 µm

Each bulk extracted DNA was used to construct the library for sequencing. Using these libraries, a total of 2.39 and 2.44 Gbp, approximately 20-fold the size of the *Arabidopsis* genome, were sequenced for *kns3-1* and *kns19-1*, respectively. The obtained short reads were processed in parallel by Mitsucal. The analysis of *kns3-1* and *kns19-1* by Mitsucal described below is presented at the following URL (https://dandelion.liveholonics.com/mitsucal/kns3/ and replace kns3 to kns19 for *kns19-1*).

Similar to *droI1-1*, Mitsucal first aligned all short reads to the reference sequence of the chromosome of the Col strain, which was a strain used for backcrossing using bowtie in same manner as for *droI1-1*. Only uniquely aligned short reads were used for the chromosome mapping described below. From the list of all detected mismatches, Mitsucal emitted mismatches with both lower and higher coverage. In contrast to the case of *droI1-1*, these lists for *kns3-1* and *kns19-1* were mainly occupied by numerous single-nucleotide polymorphisms (SNPs) between the Col and Ler strains (Supplemental Fig. 2B). The common SNPs were selected as markers and used for constructing images of chromosome mapping (Supplemental Fig. 3A). Mitsucal calculated the ratio of Ler SNPs at each marker aggregated at 50-kbp intervals (Supplemental Fig. 1C). In regard to *kns3-1*, the highest ratios of SNPs as indicated by purple bars were observed in the 22–26-Mbp region of chromosome 5 (Supplemental Fig. 3A). This result suggested that the corresponding mutation for the phenotype of *kns3-1* was located in this region. Similarly, *kns19-1* was mapped to the 0–10-Mbp region of chromosome 2 (Supplemental Fig. 4A).

To identify the *kns3-1* and *kns19-1* mutations responsible for the phenotypes, Mitsucal aligned all reads for *kns3-1* and *kns19-1* to the reference genes of the Ler strain by bowtie. The gene sequences of Ler were generated by the integration of Ler SNP in a public database (http://www.1001genomes.org/; Schneeberger et al. 2011) into the corresponding Col-0 gene sequences. For *kns3-1*, we searched for mutations in the 22–26-Mbp region of chromosome 5 in *kns3-1*, but not in *kns19-1*. Out of 222 nearly homozygous substitutions (90–100% substitutions ratio) in the 22–26-Mbp region of chromosome 5, only 7 were *kns3-1* specific (not observed in *kns19-1*), affected protein sequences and comprised G to A (or C to T) substitutions characteristically resulting from EMS. We noted that one of the candidates, AT5G58100, was already reported as the causal gene of the *spotty1* (*spot1*) mutant which displays phenotypes similar to *kns3-1* (Fig. 4e–g, Dobritsa et al. 2011). Using a conventional sequencing method, we verified the existence of a G to A substitution in the AT5G58100 gene in *kns3-1*. The mutation disrupted a conserved splice acceptor sequence (AG) of the first intron, which was altered to AA (Fig. 4c and Supplemental Fig. 3B). Reverse transcription-polymerase chain reaction (RT-PCR) revealed that the first intron was not removed from the transcripts of this gene in *kns3-1* (Fig. 5). It is assumed that the translation of AT5G58100 mRNA in *kns3-1* was terminated at an in-frame termination codon existing in the unspliced first intron. Thus, it appears likely that *kns3-1* is a null allele of this gene. To confirm that the *kns3-1* was allelic to *spot1*, we crossed *kns3-1* with two reported alleles of *spot1*, namely SALK_061320 and SALK_079847 (Dobritsa et al. 2011), and examined

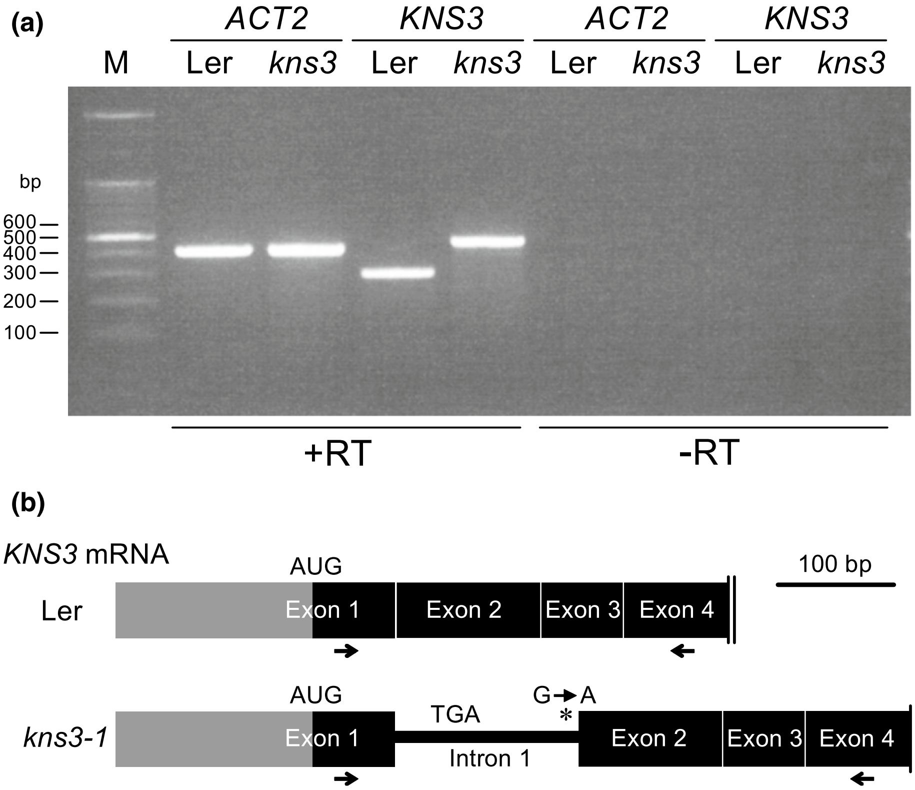


Fig. 5 Missplicing of intron 1 in *kns3-1* mutants. **a** Reverse transcription-polymerase chain reaction (RT-PCR) experiments to compare the structure of *KNS3* mRNA in *kns3-1* mutants (*kns3*) to that in the wild-type (Ler). Prior to PCR, reverse transcription was carried out with (+RT) or without (−RT) adding reverse transcriptase. The *ACT2* gene was a control. M, size marker. **b** Structures of *KNS3* mRNA in wild-type (Ler) and *kns3-1*. Black and grey boxes represent coding and non-coding regions, respectively. The positions of initiation codon (AUG), in-frame termination codon in intron 1 (TGA), and the point mutation in *kns3-1* (asterisk) are indicated. Primers used in the PCR are indicated by arrows

the pollen phenotypes of F1 progenies. The pollen grains showed obvious mutant phenotypes (Fig. 4h, i). As a consequence, we concluded that *KNS3* is identical to *SPOT1* and is AT5G58100.

We searched for the candidate mutations of *kns19-1* in the 0–10-Mbp region of chromosome 2 by Mitsucal under the same conditions as above. Sixteen G to A (or C to T) substitutions unique in *kns19-1* were identified. One of these was a missense mutation causing a substitution of glutamic acid to lysine in a gene labelled AT2G13680, known as *CALLOSE SYNTHASE 5* (*CALS5*) (Supplemental Figs. 4B and S5, Hong et al. 2001). The substitution was present in the conserved glucan synthase domain at the C-terminal of this protein. It was reported that the mutants for *CALS5* produced crumpled pollen grains with a spotted exine structure very similar to *kns19-1* (Supplemental Fig. 5, Dong et al. 2005; Nishikawa et al. 2005). To confirm that *kns19-1* was allelic to *cals5*, we crossed *kns19-1* with two reported *cals5* alleles, namely *cals5-1* and *cals5-5* (Dong et al. 2005; Nishikawa et al. 2005). Both F1 progenies showed the collapsed exine phenotype, suggesting that AT2G13680/*CALS5* was the causal gene for *kns19-1* (Supplemental Fig. 5B–E). We changed the name of *kns19-1* to *cals5-6*. The allele is believed to be partly functional due to its relatively mild phenotype.

Theoretic approach to estimate the best practice

To estimate the optimum condition of sequence volume and number of F_2 progenies required for mapping of causal mutation, we modelled and simulated genetic recombination and random sequence using a computer (Supplemental Fig. 6). It is assumed that a causal mutation and non-causal mutations between the mutant and control line regarded as markers are distributed on a chromosome. The distance (d) (unit: cM) between the neighbouring markers could be expected to follow the exponential distribution with parameter M, where M is an average of mutations per cM. Once d has been determined, the recombination between two neighbouring markers occurs at $d/100$. To simulate recombinations, a uniformly distributed random number ranging from 0 to 1 is generated for each chromosome and the chromosomes containing a random number < $d/100$ are virtually recombined. The number of parental chromosomes is referred to as f in following simulation. The number of short reads covering a marker (c) follows a Poisson distribution with parameter C, which is the average of coverage, namely the sequence volume relative to genome size. When F represents the number of F_2 progenies composing the DNA library, in which the causal mutation is homozygous and f represents the number of recombined alleles in a certain marker, the probability of sequencing a DNA fragment covering the recombined marker was $f/2F$. Under these conditions, the probability of detecting a recombined marker by sequence follows a binominal distribution with c and $f/2F$. Using this model, we tested various M, F and C and measured the distance (D) between m_0 and the closest recombined marker (M_n) (Supplemental Fig. 6A). The sum of the two consecutively estimated D provides the mapping range (S), which is the expected size of the region where the causal mutation is mapped. The result of the simulation in the case of $M = 1$ is shown in Fig. 6a, whereas the other cases are shown in Supplemental Fig. 7. In Fig. 6, the S is plotted against C. A shortening of S is observed according to the increase in F and C. When F is increased from 10 to 20, S is effectively shortened; however, an increase in F from 20 to 35 causes less shortening. The increase in C up to 20 also has a large impact on the size of S. In Fig. 6b, S is

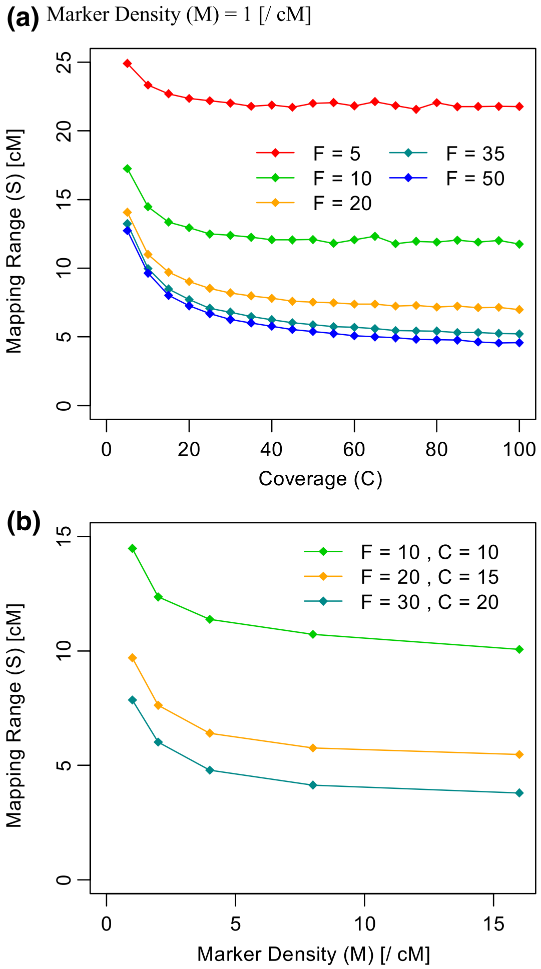


Fig. 6 Computer-simulated estimation of mapping range. **a** Estimated mapping range (S) is plotted against the coverage (C), when the different values of F_2 number (F, from 5 to 50) are given. Details of the model are explained in Supplemental Fig. 7. Plotted values are average of ten thousand trials. The density of molecular markers (M) is 1 (per cM) in this figure, and the cases when other M is given are shown in Supplemental Fig. 8. **b** Estimated mapping range (S) is plotted against the marker density (M). The parameters of F_2 number (F) and coverage (C) are indicated in the figure

plotted against M. The green ($F = 10$ and $C = 10$) and yellow lines ($F = 20$ and $C = 15$) are distant with no reference to M, whereas the yellow and blue ($F = 30$ and $C = 20$) lines are closer, confirming that the increase in F (to 20) and C (to 15) contributes largely to shorten S. M also largely affects the size of S with no reference to F and C, although it is automatically determined by the type of cross and strength of mutagenesis. On the other hand, F and C are dependent on the effort and cost expended and easily controlled by researchers. Considered together, we proposed $F > 20$ and $C > 15$ as a standard condition.

Discussion

The identification of causal mutations provides an important breakthrough for understanding the molecular mechanisms of biological phenomena including sexual plant reproduction. Although traditional methods of chromosome mapping and subsequent sequence analysis of candidate genes were remarkably time-consuming, improvement of DNA sequencing technology using high-throughput sequencers has improved this situation. As represented in the present report, we established an efficient methodology for mapping and identification of causal mutations using the computer system ‘Mitsucal’. There have been many services that provide web interfaces for various bioinformatics tools and computer resources, such as the DDBJ Annotation Pipeline (https://p.ddbj.nig.ac.jp/pipeline/) and Galaxy (https://usegalaxy.org/). As these resources are designed to be suitable for multiple purposes, they require users to set parameters to achieve optimal outputs. Mitsucal is a new tool designed only to identify corresponding mutations, and as such, all parameters have been pre-configured. Users are required only to select their plant species and the specific cross type. Mitsucal provides graphical images to select a candidate locus harbouring a corresponding mutation and uses filtering methods with biological meanings, which greatly reduce the steps required to identify mutations.

The most important tool for chromosome mapping is a set of molecular markers, namely SNPs (or mutations), of which positions are determined on the genome sequence. Since there are a remarkable number of SNPs (5.8×10^5 in a recent report; Zapata et al. 2016) between Ler and Col-0 genomes, the outcross between these strains is an effective method to map the causal mutation. In the second and third examples of this report (i.e. *kns3-1* and *kns19-1*), we used bulk segregant analysis (BSA) for sequencing of bulked F_2 DNA. The first report of BSA was published in 1991 before the emergence of the high-throughput sequencer for identification of markers linked to a specific phenotype using F_2 DNA in bulk (Michelmore et al. 1991). The concept of BSA applied to chromosome mapping by the sequences of the bulk of F_2 progenies was achieved in SHOREmap (Schneeberger et al. 2009), MutMap (Abe et al. 2012) and Mitsucal. We successfully mapped *kns3-1* in a 4-Mbp region (22–26 Mbp) of chromosome 5 showing a high (> 90%) substitution score. The identified *KNS3* is identical to At5g58100, which is located near the centre (23.5 Mbp) of the mapped region. The mapped region of *kns19-1* is wider (0–8 Mbp in chromosome 2), and the substitution score is lower (> 85%) than that of the case of *kns3-1*, suggesting a possibility that some F_2 lines used for sequencing were heterozygous for wild-type and mutant alleles. Nevertheless, Mitsucal identified 16 candidate genes, in which the true causal gene *CalS5* is included. Since laboratory strains accumulated spontaneous mutations during multiple generations in yeast (Gu et al. 2005), removing the common mutations between the mutant and its parental strain is important to reduce the candidates of causal mutation. Nevertheless, it is not necessary to sequence the parental line because Mitsucal can use a sibling mutant derived from the same parental line as a control. We used *kns19-1* for the control of *kns3-1* to restrict candidate mutations and vice versa.

Although these advantages, the outcross is sometimes difficult to apply. For example, if the suppressor and enhancer mutants in a background of other recessive mutation are outcrossed with wild-type plants of a different strain, the phenotypes in F_2 populations will be complicated. A similar difficulty will be expected for the mutants isolated from transgenic lines and recombinant inbred lines. To solve the problem, we propose utilisation of base substitutions occurred by mutagenesis. Because they are unique in the isolated mutant, they are regarded as SNPs between the mutant and parental line. Hence, if the mutations are effectively identified, they can be used as molecular markers. In the first example of this paper, *drol1-1* was backcrossed with the parental line containing *Ole3p::LUC*, and genomic DNA prepared from bulked F_2 mutants was sequenced. Although the linkage was less clear than the cases of *kns3-1* and *kns19-1*, Mitsucal successfully mapped and identified the causal mutation.

We simulated the expected mapping range (S) in relation to marker density (M), number of F_2 mutants bulked for sequencing (F) and sequencing coverage (C). Although large M contributes effectively to decrease S, particularly when C is small, it is automatically determined as a constant dependently on the type of cross (outcross or backcross) and strength of mutagenesis. Instead, F and C are easily varied to optimise the result. A large F provides the largest contribution with no relation to C and M and is recommended to be 20 (or hopefully 35). A large C up to 50 also contributes to reduce S when F is small, although it increases the cost. We propose $C = 15$ as a realistic solution.

We have reported here the identification of the corresponding genes of sporophytic mutants by Mitsucal;

however, gametophytic mutants were also important for the study of sexual reproduction. Mutations at reproductive genes involved in like a pollen–pistil interaction are appeared in gametophytes, which mean no homozygous plants are recovered and mapping method described above is not applicable. To map the gametophytic mutation on chromosomes, two populations of siblings, one is homozygous for wild-type allele and the other is heterozygous, are required to be sequenced. The mutations found in heterozygous siblings are selected as marker, and the ratio of that was expected to be 0 at the locus harbouring corresponding gene in wild-type siblings. This method is implemented to Mitsucal in near future.

We recommend EMS as a mutagen for Mitsucal analysis, because it results in a strictly limited and therefore easily identified pattern of sequence substitution in DNA, whereas the development of new software suitable for identification of gaps and insertions is currently progress. The current version of Mitsucal is capable of identifying mutations of *Arabidopsis* and *Oryza sativa* and will be expanded to *Solanum lycopersicum* and other species.

Materials and methods

Plant material and growth conditions

A. thaliana (L.) Heynh. strain Columbia (Col-0) and strain Landsberg *erecta* (Ler) were used as the wild-type plants. *kns3-1* (former name *kns3*) was described previously (Suzuki et al. 2008). *kns19-1* is a newly isolated mutant from the same parental population as *kns3-1*. T-DNA insertion alleles of *drol1* (SALK_087803), *kns3*/*spot1* (SALK_041228, SALK_061320 and SALK_079847) and *kns19*/*cals5* (*cals5-1*/SALK_009234) were obtained from the Arabidopsis Biological Resource Center (https://abrc.osu.edu/). *cals5-5*/SALK_072226 was provided by Dr. Shuh-ichi Nishikawa. Unless otherwise indicated, seeds were sterilised on 0.3% (wt/vol) gellan gum plates containing Murashige and Skoog (MS) medium (Wako, Japan) (pH 5.7) supplemented with 100 mg/L myo-inositol, 10 mg/L thiamine HCL and 1% sucrose. Plates were incubated in a growth chamber at 22 °C under continuous fluorescent light at an intensity of 65 μmol m^{-2} s^{-1}. For observation of pollen grains, the seedling were transplanted to vermiculite and grown at the same condition as above.

Generation of transgenic plants, mutagenesis and screening

The upstream 1200-bp region for Ole3 was obtained by PCR with the following primers: GGGGACAAGTTTGTACAAAAAAGCAGGCTTATGTAGAACTAAAGACTAAGGGACAGAG and GGGGACCACTTTGTACAAGAAAGCTGGGTCCGCCATTTTTTTGTTCTTGTTTACTAGAG and cloned into pDONR201. The promoter sequence was placed upstream of the luciferase in the binary vector pGWB435 (Nakagawa et al. 2008) by Gateway cloning technology (Invitrogen https://www.thermofisher.com/). Col-0 plants were transformed with *Agrobacterium tumefaciens* GV3101 harbouring the binary vector by floral dip infiltration. The transgenic line harbouring one copy of the *Ole3p::LUC* reporter was used for the mutagenesis with EMS. Mutagenised M2 seed was sterilised and sown on a well (1 seed per well) of a 96-well assay plate filled with a solidified medium containing 0.01 mM luciferin. Luciferase activity was examined in a real-time bioluminescence monitoring system (CL96, Churitsu, Japan).

For complementation, the coding sequence of *DROL1* PCR was amplified with the following primers: 5′-ggggACAAGTTTGTACAAAAAAGCAGGCTCGATGTCGTATAATTCTTACGCCGAGATGA-3′ and 5′-ggggACCACTTTGTACAAGAAAGCTGGGTACACATCCTTGTACACGAGCTG-3′. The PCR product was cloned into pDONR221 and then transferred into the pGWB502 omega by Gateway BP and LR reactions. The resulting plasmid (pGWB-35S-DROL1) was used for the Agrobacterium-mediated transformation of *drol1-1*/*Ole3p::LUC* plants. Scanning electron microscopy for observation of pollen grains was carried out as described previously (Suzuki et al. 2008).

Gene expression analysis

Total RNA was prepared by use of the RNeasy Plat Mini Kit (Qiagen, http://www.qiagen.com). Reverse transcription and subsequent PCR were conducted with PrimeScript RT reagent Kit and Ex Taq polymerase (Takara, Japan).

DNA extraction and library construction

The DNA was extracted from young seedlings of *Arabidopsis* using the DNeasy Plant Mini Kit from Qiagen. The extracted DNA was fragmented by sonication (S220 from Covaris http://covarisinc.com/) and then converted to the library using TruSeq DNA Sample Prep Kits (Illumina, http://www.illuminakk.co.jp/) according to the manufacturers protocol.

Sequence and data extraction

The libraries were sequenced by Genome Analyzer IIx (Illumina). The produced bcl files were converted to fastq files by CASAVA (Illumina).

Implementation of Mitsucal

Mitsucal was built on the cluster computer Kiku 1st, consisting of a single head node (dual quad-core CPUs and 20 GB RAM), eight computer nodes (single six-core CPU and 24 GB RAM), one file server and one database server (single six-core CPU and 64 GB RAM). All computers, except for the head node, were home-made. The computers in Kiku 1st were connected by a 1-Gbp ethernet. In Linux, installed as an OS, a Sun Grid Engine was used for the job management system. PostgreSQL was installed to the database server.

Data processing in Mitsucal

Fastq files were acceptable to Mitsucal. Csfasta files from SOLiD were also acceptable. All reads were mapped to the reference by bowtie, and the results were directly stored in PostgreSQL. After mapping by bowtie, the reads were sorted by PostgreSQL, and substitutions were detected by the originally developed scripts written in PHP. The detected substitutions were stored as a table in PostgreSQL. The web interface was also written in PHP. All scripts used in Mitsucal were submitted as supplemental information.

Author contribution statement T.S, T.K., K.N., S.I. and T.H. conceptualised the study; T.S. carried out formal analysis; T.K., S.T., M.N., T.S. and S.I. carried out investigation; K.N., S.I. and T.H. collected resources; T.S., T.K., K.N., S.I. and T.H. wrote the manuscript.

Acknowledgements The authors thank Ms. T. Shinagawa for her skilled technical assistance.

Accession number Sequence data from this article can be found in the DDBJ Sequence Read Archive under accession number DRA006126.

Funding This work was supported by grants from the Japan Science and Technology Agency (ERATO project to T.H.) and the Japan Society for the Promotion of Science (Nos. 15H05955 to T.S., 16H06465 to T.H. and 15H01229 to S.I.).

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 using MutMap. Nat Biotechnol 30:174–178

Chen YH, Li HJ, Shi DQ, Yuan L, Liu J, Sreenivasan R, Baskar R, Grossniklaus U, Yang WC (2007) The central cell plays a critical role in pollen tube guidance in Arabidopsis. Plant Cell 19:3563–3577

Dai XR, Gao XQ, Chen GH, Tang LL, Wang H, Zhang XS (2014) ABNORMAL POLLEN TUBE GUIDANCE1, an endoplasmic reticulum-localized mannosyltransferase homolog of GLYCOSYLPHOSPHATIDYLINOSITOL10 in yeast and PHOSPHATIDYLINOSITOL GLYCAN ANCHOR BIOSYNTHESIS B in HUMAN, is required for arabidopsis pollen tube micropylar guidance and embryo development. Plant Physiol 165:1544–1556

Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12:499–510

Dobritsa AA, Geanconteri A, Shrestha J, Carlson A, Kooyers N, Coerper D, Urbanczyk-Wochniak E, Bench BJ, Sumner LW, Swanson R, Preuss D (2011) A large-scale genetic screen in Arabidopsis to identify genes involved in pollen exine production. Plant Physiol 157:947–970

Dong X, Hong Z, Sivaramakrishnan M, Mahfouz M, Verma DP (2005) Callose synthase (CalS5) is required for exine formation during microgametogenesis and for pollen viability in Arabidopsis. Plant J 42:315–328

Escobar-Restrepo JM, Huck N, Kessler S, Gagliardini V, Gheyselinck J, Yang WC, Grossniklaus U (2007) The FERONIA receptor-like kinase mediates male-female interactions during pollen tube reception. Science 317:656–660

Faino L, Thomma BP (2014) Get your high-quality low-cost genome sequence. Trends Plant Sci 19:288–291

Kanamori A, Sugita Y, Yuasa Y, Suzuki T, Kawamura K, Uno Y, Kamimura K, Matsuda Y, Wilson CA, Amores A, Postlethwait JH, Suga K, Sakakura Y (2016) A genetic map for the only self-fertilizing vertebrate. A genetic map for the only self-fertilizing vertebrate (Bethesda) 6:1095–1106

Kessler SA, Shimosato-Asano H, Keinath NF, Wuest SE, Ingram G, Panstruga R, Grossniklaus U (2010) Conserved molecular components for pollen tube reception and fungal invasion. Science 330:968–971

Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25

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

Miquel M, Trigui G, d'Andréa S, Kelemen Z, Baud S, Berger A, Deruyffelaere C, Trubuil A, Lepiniec L, Dubreucq B (2014) Specialization of oleosins in oil body dynamics during seed development in Arabidopsis seeds. Plant Physiol 164:1866–1878

Mori T, Igawa T, Tamiya G, Miyagishima SY, Berger F (2014) Gamete attachment requires GEX2 for successful fertilization in Arabidopsis. Curr Biol 24:170–175

Nagalakshmi U, Wang Z, Waern K, Shou C, Raha D, Gerstein M, Snyder M (2008) The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320:1344–1349

Nakagawa T, Nakamura S, Tanaka K, Kawamukai M, Suzuki T, Nakamura K, Kimura T, Ishiguro S (2008) Development of R4 gateway binary vectors (R4pGWB) enabling high-throughput promoter swapping for plant research. Biosci Biotechnol Biochem 72:624–629

Nishikawa S, Zinkl GM, Swanson RJ, Maruyama D, Preuss D (2005) Callose (beta-1,3 glucan) is essential for Arabidopsis pollen wall patterning, but not tube growth. BMC Plant Biol 5:22

Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D (2008) Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res 18:2024–2033

Palanivelu R, Brass L, Edlund AF, Preuss D (2003) Pollen tube growth and guidance is regulated by POP2, an Arabidopsis gene that controls GABA levels. Cell 114:47–59

Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nat Methods 4:651–657

Schneeberger K, Ossowski S, Lanz C, Juul T, Petersen AH, Nielsen KL, Jørgensen JE, Weigel D, Andersen SU (2009) SHOREmap: simultaneous mapping and mutation identification by deep sequencing. Nat Methods 6:550–551

Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N, Henz SR, Huson DH, Weigel D (2011) Reference-guided assembly of four diverse *Arabidopsis thaliana* genomes. Proc Natl Acad Sci USA 108:10249–10254

Shimizu KK, Ito T, Ishiguro S, Okada K (2008) MAA3 (MAGATAMA3) helicase gene is required for female gametophyte development and pollen tube guidance in *Arabidopsis thaliana*. Plant Cell Physiol 49:1478–1483

Suzuki T, Masaoka K, Nishi M, Nakamura K, Ishiguro S (2008) Identification of kaonashi mutants showing abnormal pollen exine structure in *Arabidopsis thaliana*. Plant Cell Physiol 49:1465–1477

Tabata S, Kaneko T, Nakamura Y, Kotani H, Kato T, Asamizu E, Miyajima N, Sasamoto S, Kimura T, Hosouchi T, Kawashima K, Kohara M, Matsumoto M, Matsuno A, Muraki A, Nakayama S, Nakazaki N, Naruo K, Okumura S, Shinpo S, Takeuchi C, Wada T, Watanabe A, Yamada M, Yasuda M, Sato S, de la Bastide M, Huang E, Spiegel L, Gnoj L, O'Shaughnessy A, Preston R, Habermann K, Murray J, Johnson D, Rohlfing T, Nelson J, Stoneking T, Pepin K, Spieth J, Sekhon M, Armstrong J, Becker M, Belter E, Cordum H, Cordes M, Courtney L, Courtney W, Dante M, Du H, Edwards J, Fryman J, Haakensen B, Lamar E, Latreille P, Leonard S, Meyer R, Mulvaney E, Ozersky P, Riley A, Strowmatt C, Wagner-McPherson C, Wollam A, Yoakum M, Bell M, Dedhia N, Parnell L, Shah R, Rodriguez M, See LH, Vil D, Baker J, Kirchoff K, Toth K, King L, Bahret A, Miller B, Marra M, Martienssen R, McCombie WR, Wilson RK, Murphy G, Bancroft I, Volckaert G, Wambutt R, Düsterhöft A, Stiekema W, Pohl T, Entian KD, Terryn N, Hartley N, Bent E, Johnson S, Langham SA, McCullagh B, Robben J, Grymonprez B, Zimmermann W, Ramsperger U, Wedler H, Balke K, Wedler E, Peters S, van Staveren M, Dirkse W, Mooijman P, Lankhorst RK, Weitzenegger T, Bothe G, Rose M, Hauf J, Berneiser S, Hempel S, Feldpausch M, Lamberth S, Villarroel R, Gielen J, Ardiles W, Bents O, Lemcke K, Kolesov G, Mayer K, Rudd S, Schoof H, Schueller C, Zaccaria P, Mewes HW, Bevan M, Fransz P (2000) Sequence and analysis of chromosome 5 of the plant *Arabidopsis thaliana*. Nature 408:823–826

The Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature 408:796–815

von Besser K, Frank AC, Johnson MA, Preuss D (2006) Arabidopsis HAP2 (GCS1) is a sperm-specific gene required for pollen tube guidance and fertilization. Development 133:4761–4769

Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S, Schneeberger K (2016) Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci USA 113:E4052–E4060