bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CRISPR-finder: A high throughput and cost effective method for identifying successfully edited *A. thaliana* individuals

Efthymia Symeonidi[1†], Julian Regalado[1†], Rebecca Schwab[1], Detlef Weigel[1#]

Affiliations:

[1]Max Planck Institute for Developmental Biology, Department of Molecular Biology, Max-Planck-Ring 5, 72076 Tübingen, Germany.

[†]Current address: Center for Plant Molecular Biology (ZMBP), University of Tübingen, Auf der Morgenstelle 32, 72076, Tübingen, German (E.S.); The Globe Institute, Faculty of Health and Medical Sciences, University of Copenhagen, Øster Voldgade 5 - 7, 1350 Copenhagen, Denmark (J.R.)

[#]Correspondence to: weigel@weigelworld.org

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Abstract

Background

Genome editing with the CRISPR/Cas9 system allows the user to mutate a targeted region of the genome using an endonuclease (Cas9) and an artificial single-guide RNA (sgRNA). Both because of variable efficiency with which such mutations arise and because the repair process produces a spectrum of mutations, one needs to ascertain the genome sequence at the targeted locus for many individuals that have been subjected to CRISPR/Cas9 mutagenesis. This process can be laborious, expensive and inefficient with conventional methods such as the T7E1 assay or Sanger sequencing. An alternative comprises methods for amplicon sequencing, but most available protocols do not include a facile way for high throughput generation of the samples for sequencing.

Results

In this study we provide a full pipeline based on amplicon sequencing, CRISPR-finder. We provide a complete protocol for the generation of amplicons up until the identification of the exact mutations in the targeted region. CRISPR-finder can be used to process thousands of individuals in a single sequencing run. For example, we were able to analyze in one sequencing reaction over 900 *Arabidopsis thaliana* individuals whose genomes had been targeted with the CRISPR/Cas9 system.

Conclusions

In order to validate the potential of CRISPR-finder, we targeted the *ISOCHORISMATE SYNTHASE 1* gene in *A. thaliana* using the CRISPR/Cas9 system. We successfully identified a mutant line in which the production of salicylic acid was impaired compared to the wild type, as expected. These features establish CRISPR-finder as a high-throughput, cost-effective and -efficient genotyping method of individuals whose genomes have been targeted using the CRISPR/Cas9 system.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Background

Genome editing has become a routine approach to investigate gene function *in vivo*. The recent development of CRISPR/Cas9-based systems has opened new doors for genome editing by simplifying the requirements for genome targeting, particularly in comparison to zinc finger nucleases and TALENs [1]. The system requires a nuclease (Cas9), an artificial single-guide RNA (sgRNA), and a short sequence upstream of the sgRNA binding site called a Protospacer Adjacent Motif (PAM), which has the sequence 5'-NGG-3' [2,3]. Part of the sgRNA is complementary to 20 nucleotides in the targeted region of the genome, and the rest is responsible for the stabilisation of the Cas9/sgRNA complex.

Interaction of the Cas9/sgRNA complex with the target site enables Cas9's endonuclease domain to generate a double-stranded break (DSB). Such breaks can be repaired through either the non-homologous end joining (NHEJ) or the homology directed repair (HDR) pathway. NHEJ is error-prone, and can introduce small insertions or deletions that can lead to the disruption of open reading frames [4,5]. In the case of HDR, a donor template complementary to the target needs to be present to introduce a specific region to the genome of interest [6,7]. The CRISPR/Cas9 and related systems have been used to generate knock-outs [8,9], knock-ins [10,11] and to delete entire genes [12] in several species including the plant *Arabidopsis thaliana* [13–16].

While the generation of mutants using CRISPR/Cas9 is relatively easy, identification of desired mutations often requires screening many events. Two common approaches to screen for induced mutations are Sanger sequencing [14,17] or the T7 Endonuclease I (T7E1) assay [18,19] applied to individual PCR products. Unfortunately, neither method provides immediately a precise identification of mutations in the desired region. For example, in the case of Sanger sequencing, the final readout merges the most abundant products in the template into one chromatogram [20,21]. This can lead to secondary peaks and sometimes a mixed signal due to other amplified molecules in the mixture, and can make it very hard to detect desired but rare events that might have occurred during editing. Confirmation of successful editing through

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

subsequent cloning of a mixed PCR product followed by retrieval of bacterial colonies that carry the rare variant is time-consuming and expensive. Use of T7E1 can also yield inconclusive results due to its reliance on the T7 Endonuclease I to digest only fragments carrying mismatches [22], which would miss homozygous mutants, as there are no mismatched fragments available for digestion. In addition, both techniques can be expensive for screening a large number of samples (>100).

These limitations led us to develop a robust and cost-efficient way of efficiently screening large numbers of samples. Here we introduce a high-throughput screening approach for identifying mutations using Illumina sequencing, called CRISPR-finder. We describe both the library preparation of the samples and the analysis pipeline for identifying editing events. The method is compatible with sequencing on different Illumina instruments, and the adapter sequences could be modified for use on other platforms.

Our approach is an adaptation of an amplicon sequencing method previously developed for pooling samples for the analysis of microbiomes [23]. In our approach, the amplicon libraries are generated through a two-step PCR amplification. During the PCRs, frameshifting nucleotides and one of 96 unique indices are added. Based on the unique combination of the frameshifting nucleotides and the barcode we were able to sequence hundreds of samples, for example >900 samples, in a single MiSeq run. To illustrate the accuracy and the precision of the method we describe how we identified and characterised a Cas9-free *ISOCHORISMATE SYNTHASE 1* (*ICS1*) mutant.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Target site identification

The aim of this study was to improve the speed of mutant identification with the CRISPR/Cas9 system. To demonstrate the efficacy of this new approach, the *ISOCHORISMATE SYNTHASE I* (*ICS1*) gene was targeted in different *A. thaliana* accessions (Supplementary Table 1). *ICS1* encodes an enzyme involved in salicylic acid biosynthesis [24].

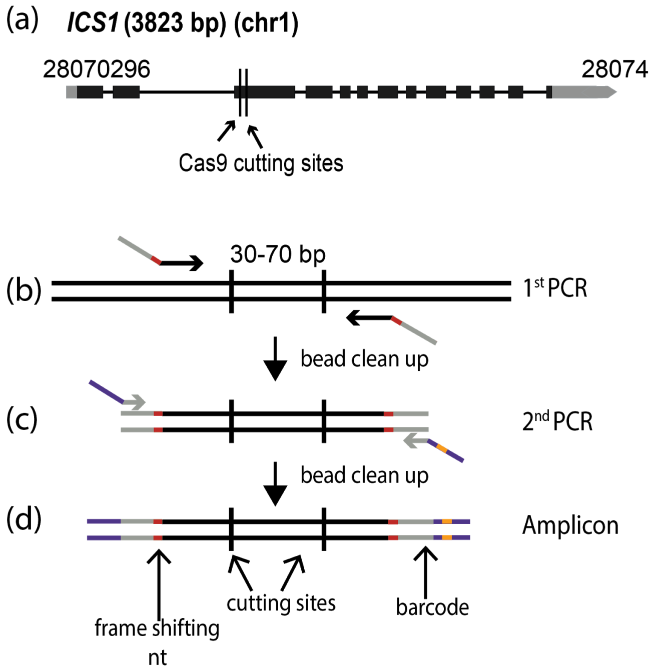

Figure 1: Amplicon preparation. (a) Diagram of the targeted gene, *ISOCHORISMATE SYNTHASE I* (*ICS1*). Black boxes indicate exons, and grey boxes untranslated regions. The arrow shows the direction of transcription. **(b)-(d)** Amplicon preparation. **(b)** The first PCR step to amplify a specific region of the genome. The oligonucleotide primers in this step fuse the first part of the TruSeq adapters (grey) and the frame shifting nucleotides (red). **(c)** The second PCR amplification adds the last part of the TruSeq adapters (purple) and one of the 96 barcodes (orange). **(d)** The final amplicon with frameshifting base pairs(s) (red), TruSeq adapters (grey and purple) and barcode (orange).

The accessions of *A. thaliana* used in this study are from the first phase of the 1001 Genomes Project [25]. The polymorph tool (http://polymorph.weigelworld.org) was used to align sequences of *ICS1* from the different accessions. Target sites without sequence variation among the accessions were identified to select the guide RNAs (Figure 1a).

Plants were transformed separately with the *ICS1* targeting construct (Supplementary Table 2). The primary transformants were found to have somatic editing events by using the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

CRISPR-finder genotyping pipeline. Two versions of Cas9 were used, either plant-codon-optimised (pcoCas9) [26] or Arabidopsis-codon-optimized (AthCas9) [17]. The selection of the transgene was based on glufosinate or the seed-specific expression of mCherry [27,28].

Generation and sequencing of amplicons spanning CRISPR/Cas9 target sites

In order to quickly and unambiguously identify CRISPR/Cas9-induced mutations in a large number of plants, the targeted regions were amplified by PCR, attaching different barcodes for different individual plants, and then pools of barcoded PCR products were sequenced on an Illumina MiSeq (or HiSeq) instrument. The *ICS1* locus was targeted in different accessions to determine the efficacy of the method at different genetic backgrounds. Two sites were targeted in the gene, 72 bp apart for *ICS1*. The amplified regions were 211 bp long.

The amplicons were prepared based on a two-step PCR amplification protocol (Figure 1b-d). During the first round of amplification, the specific region of interest was amplified, and frameshifting nucleotides and part of the Illumina TruSeq adapters were added (Figure 1c) (Supplementary Table 3). The cleaned PCR product was used as a template for the second round of amplification, where the remainder of the TruSeq adapters and one of 96 barcodes were added [23] (Figure 1d) (Supplementary Table 3). Each PCR amplification step was carried out for 15 cycles.

The PCR products were quantified using the Quant-iT^{TM} PicoGreen® dsDNA assay, normalized (described in Methods), and pooled. For the sequencing on the MiSeq platform, the MiSeq reagent kit v2 (300-cycles) (MS-102-2002) was used. The adapters were designed and chosen in order to be compatible with both MiSeq and HiSeq3000 platforms (Illumina, USA); successful runs were carried out on both platforms.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Demultiplexing process

After sequencing, the pooled reads were demultiplexed in a two-step process. 96 batches of combined samples were first identified via the indices that were located at the TruSeq adapters incorporated in the 2nd PCR amplification. This process was carried out with bcl2fastq (1.8.4) software, provided by Illumina, which also trims the sequence of the barcodes (https://my.illumina.com) (Figure 2a).

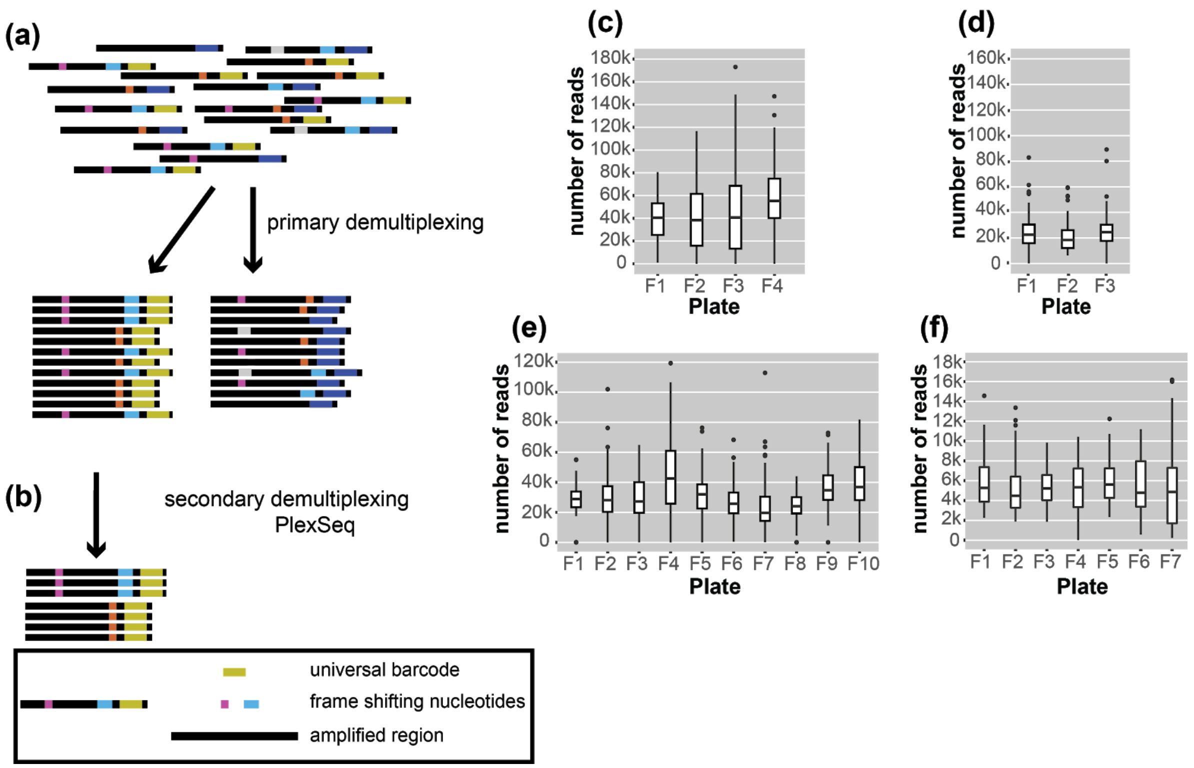

Figure 2: Diagrams of the demultiplexing procedure and graphs representing the average number of reads/plate/run. (a) The primary demultiplexing step is carried out by the Illumina software and separates the samples based on the indices that are located within the adapter region into 96 pools. **(b)** The secondary demultiplexing script (PlexSeq) then assigns the reads to individual plants based on the frameshifting nucleotides. **(c)-(f)** Each graph shows the average number of reads per plate (≤ 96 samples) in each run. For different runs, different numbers of plates were sequenced depending on the number of samples. **(c)** MiSeq run 010, **(d)** MiSeq run 024, **(e)** MiSeq run 046, and **(f)** MiSeq run 083.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Subsequently, sequencing reads from different samples were mixed under the same barcode. In order to assign each read to the individual from which it came, we took advantage of the frameshifting nucleotides incorporated during the first step of the two-step PCR amplification. The first nine nucleotides from each read were used as "secondary" barcodes to determine from which sample each read in the sequencing run originated; 9 bases are sufficient to capture the unique frameshifting nucleotides used during the amplicon generation (Figure 2b).

For binning of reads, the PlexSeq Python script (https://github.com/7PintsOfCherryGarcia/plexseq) was developed, which successfully demultiplexes >98% of reads in each dataset (Figure 2b). Since PlexSeq was run without allowing any mismatches of the "secondary" barcodes, around 2% of the data could not be separated because of errors in PCR primers or errors introduced during the sequencing process; a loss of 2% of reads was deemed acceptable (FIgure 2c-f). These unassigned reads are ignored in downstream analyses. A file with the expected "secondary" barcodes needs to be provided in order for the script to successfully proceed with demultiplexing (Supplementary Figure 1).

Analysis pipeline

After the demultiplexing process, each sample was genotyped in order to detect single nucleotide polymorphisms (SNPs), as well as small insertions and deletions in the region of interest.

For each sample, reads were mapped back to the reference sequence for the gene of interest (Gene ID: 843810) using the MEM algorithm of the BWA read mapping tool [29] with standard parameters (Figure 3a). The resulting alignment files were genotyped with freebayes using standard parameters (Figure 3a) [30]. The resulting VCF file was then filtered with vcftools [31] to only keep samples in which high quality variants were detected at regions of interest.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

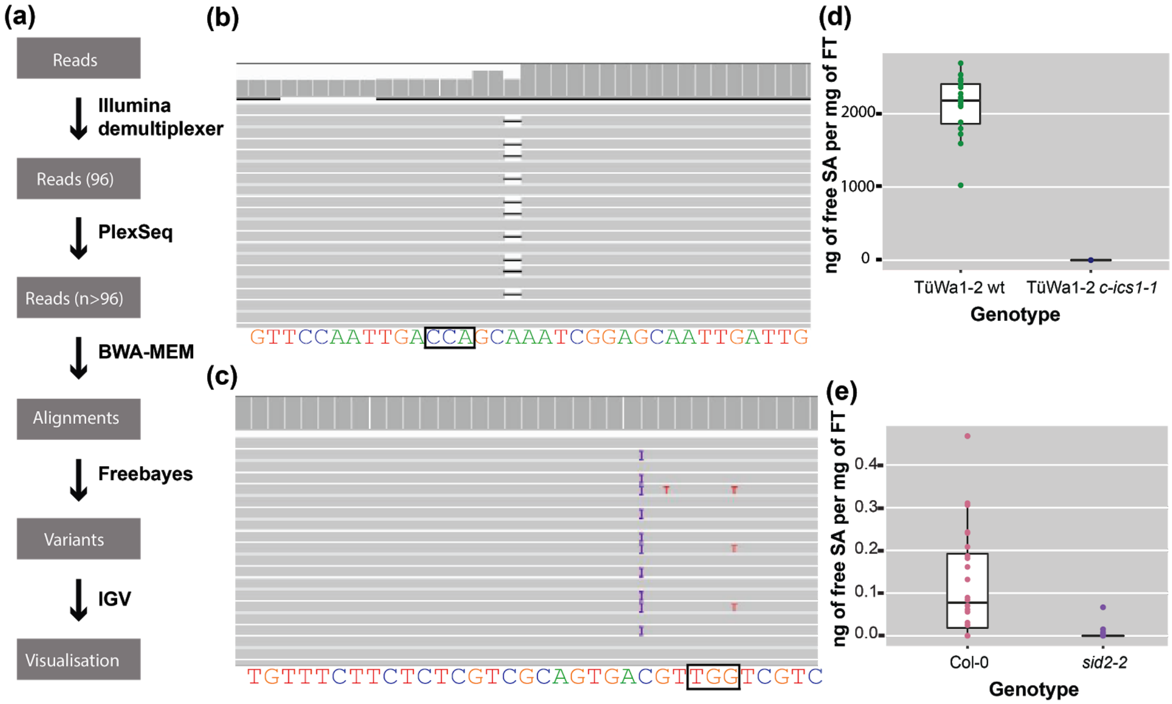

Figure 3: The analysis pipeline and visualised alignments and SA levels of different genotypes. (a) Diagram of the analysis pipeline. BWA-MEM is used for the alignment and the Freebayes algorithm for variant calling. Finally, IGV is used for visualizing the alignments or the vcf files. **(b)** Alignment that shows a deletion visualised in IGV. On the top track in the coverage panel it is apparent how coverage is decreased at the location of the deletion. The black box indicates the location of the PAM site. **(c)** Alignment that shows a 1-bp insertion (purple 'I') in IGV. The black box indicates the location of the PAM site. **(d)** SA content of TüWa1-2 wild-type and the derivative TüWa1-2 *c-ics1-1* mutant. **(e)** SA content of Col-0 reference wild type and derivative *sid2-2* mutant for comparison. *SID2* is a synonym for *ICS1*. Note the very different scale from (d). The measurements were obtained using plants that were growing in 23°C short-day conditions (8 h light/16 h dark) for 43 days.

At the end, IGV [32,33] was used for visual inspection of read mapping and variant calls (Figure 3b,c). All software was used with standard parameters unless otherwise noted. The required memory for the analysis can be 5-20 Gb depending on the output of the run.

Identifying mutations

Using CRISPR-finder, plants either heterozygous or homozygous for targeting events were identified. As a proof of concept for our approach, we targeted the *ICS1* gene in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

TüWa1-2 background. The TüWa1-2 *c-ics1-1* mutant was identified after screening more than one hundred individuals. The parental genotype was originally collected in Germany and its phenotype shows extensive necrotic lesions on the leaves (Supplementary Figure 2), which can be attributed to extensive cell death. It was hypothesized that this is caused by elevated levels of SA. Using a biosensor assay, the SA content in plants was quantified [34,35]. As expected, the levels of free SA in the TüWa1-2 *c-ics1-1* mutant were significantly lower than in the wild-type parental lines (Figure 3d,e). These results demonstrate that our approach of screening can easily and rapidly identify individuals with targeted mutations that have the desired effect.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Discussion

We describe a high-throughput screening approach, called CRISPR-finder, that increases the accuracy and reduces the time and cost required for identifying CRISPR/Cas induced mutations (Figure 4). We generate barcoded amplicons of the targeted region through a two-step PCR amplification (Figure 1b-d). For each individual a unique combination of frameshifting nucleotides and index sequence is used, which greatly increases the number of barcodes. An important consideration is that pooling of amplicons for sequencing can lead to unbalanced representation of samples. However, if we aim for average coverage of 1,000x, and assume that 10% of individuals provide 10x as many reads as aimed for, and 10% of individuals provide only one tenth of the reads aimed for, a single MiSeq run (~15 to 20 million reads) would still provide sufficient coverage to analyze over 10,000 samples in a single run. Of course, the coverage can be adjusted to the needs of different experimental set ups.

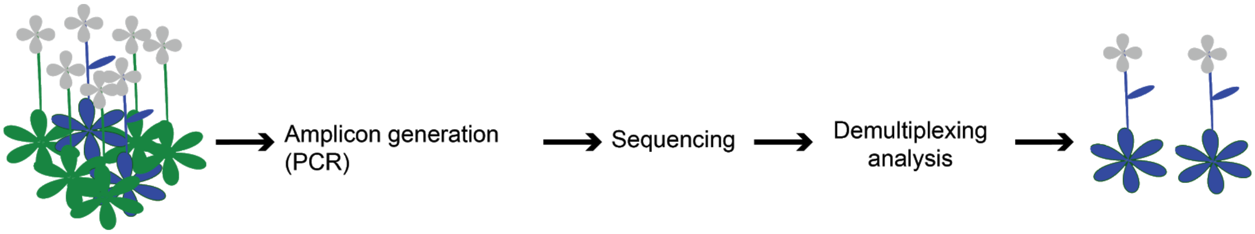

Figure 4: Schematic representation of the screening pipeline. Starting from hundreds of samples the amplicon generation takes place by preparing the individuals for sequencing. By the end of the sequencing run the demultiplexing and analysis can take place that can lead to the identification of the desired edited individuals.

For processing large numbers of samples, CRISPR-finder is a particularly cost-effective method. While with conventional assays such as Sanger sequencing and the the T7E1 assay, costs scale linearly with the number of samples to screen, for CRISPR-finder in one sequencing pool, thousands of samples can be sequenced with high resolution and the cost per sample decreases as more samples are added to the pool. Additionally, "spiking in" samples into another sequencing run to use only part of a flowcell's capacity is possible, further increasing flexibility and reducing costs.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

There have been attempts over the last few years to address the difficulties posed by available screening methods. There are available packages for the downstream analysis of demultiplexed reads, or packages that demultiplex reads that originated from different regions of the genome [36,37]. There is also an available R package (CrispRVariants) that one can use to summarize variant features such as variant type, location and coverage [38]. The input for this package can be either Sanger or NGS data, but a method to process numerous individuals is not described in the pipeline [38]. The advantage of CRISPR-finder is that we introduce not only the preparation method for the amplicons, but also methods to multiplex numerous amplicons from hundreds of samples with high resolution. Our method uses frameshifting nucleotides for higher sequence quality without the necessity of PhiX controls for Illumina sequencing, and it has particularly high multiplexing capability in comparison with other methods [39].

Finally, the first oligonucleotide set does not need HPLC purification, limiting the cost of the multiplexing procedure. This is because oligonucleotide primers are synthesized from 3' to 5', and truncations or errors will therefore be concentrated towards the ends of the amplicon during the first round of amplification. These ends serve as the binding sites for the oligonucleotides that are used for the second round of amplification, which will anneal despite minor errors and result in products with the correct adapter sequence.

While our method was developed for screening *Arabidopsis thaliana* CRISPR/Cas9 mutagenized individuals, it can be easily adopted for any organism that has been genome edited using the CRISPR/Cas9 or related systems. Note, however, that large deletions induced by CRISPR/Cas9 editing [40] would escape detection with our pipeline.

In conclusion, a full pipeline from DNA extraction to identification of individuals carrying mutations generated with the CRISPR/Cas9 system is described in detail – CRISPR-finder. Compared to more conventional methods (Sanger and T7E1 assay), large-scale amplicon sequencing is more robust and less expensive.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Material and Methods

Plant growth

Arabidopsis thaliana seeds were kept at -80°C overnight and then surface-sterilised with 70% ethanol and 0.05% (v/v) Triton X-100 for 5 minutes, followed by 100% ethanol for 5 minutes. Seeds were air-dried in a sterile hood until all residual ethanol had evaporated. Seeds were stratified in 0.1% (w/v) agar-agar for 7 days in the dark at 4°C prior to sowing on soil. Vernalization-requiring seedlings (highlighted with blue in Supplementary Table 1) were placed for seven weeks in 4°C short-day conditions (8 h light/16 h dark) and then transferred to 23°C long-day conditions (16 h light/8 h dark). For SA assays, plants were grown in 23°C short-day conditions (8 h light/16 h dark).

Plasmid generation

Constructs for plant transformation were generated using the GreenGate cloning system [41]. The five different constructs used are described in Supplementary Table 2. Two versions of Cas9 were used: the plant codon optimised (pcoCas9) [26] and the Arabidopsis codon optimised (AthCas9) [17]. The promoters used were CaMV35S, *ICU2* and EC1.1 (courtesy of Dr. Martin Bayer) [41–43]. The sgRNA constructs were generated as described in [44], pEF016 (5'-AATCAATTGCTCCGATTTGC-3') and pEF017 (5'-TTCTCTCGTCGCAGTGACGT-3').

Plant transformation

Plants were transformed using the flora dip method as described by [45].

Selection of Cas9-free plants

Two selection markers were used, resistance to glufosinate ammonium (BASTA SL, Bayer Crop Science, Leverkusen, Germany) and AT2S3::mCherry [28]. To select transgene-free plants that no longer carried BASTA resistance, leaves were brushed with a solution, diluted from the

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

original stock (200 g/l) BASTA (1:1,000 or 1:2,000) (Bayer Crop Science, Leverkusen, Germany). The treatment caused leaves from plants without the transgene to become wrinkled and yellowish.

Seeds from plants that were carrying the AT2S3::mCherry [27] cassette were screened for fluorescence or absence thereof under a LEICA MZFLIII Fluorescence stereoscope (Wetzlar, Germany) with a SOLA 365 SM Light Engine© lamp (Lumencot, Beaverton, OR, United States).

DNA isolation

Genomic DNA was extracted following a published protocol [46], with an additional ethanol wash. DNA was resuspended in 100 µL of ddH_2O.

Salicylic acid quantification

The protocol was adapted from [47]. Fresh tissue was collected and frozen at -80°C overnight. For every 175 mg of fresh tissue, 250 µL of 0.1 M pH 5.5 sodium acetate was added post grinding for further vortexing. *Acinetobacter* sp. ADPWH_lux strain was used [34] for the quantification of salicylic acid. Overnight culture of *Acinetobacter* sp. ADPWH_lux at 37°C was diluted (1:20) and grown at 37°C while shaking at 200 rpm until it reached OD_{600} of 0.4. For measuring free and 2-O-β-D-glucoside (SAG) SA, plant crude extract from the samples was incubated at 37°C for 1.5 hours with 0.4 U/µL of β-glucosidase prior to measurement.

Black Optiplates (96 wells, Greiner Bio-one, ref:655906) (Kremsmuenster, Austria) were used for the measurements. They were loaded with 50 µL of LB, 60 µL of the cell culture and 30 µL of the plant extract. Standards were prepared with 50 µL of LB, 60 µL of the cell culture, 10 µL of known SA concentrations and 20 µL of plant extract from *35S::NahG* plants as control (Col-0 background) (prepared the same way as the samples). The plates were incubated at 37°C for 2 hours without shaking and the luminescence was measured using the TECAN infinite F200 instrument (Maennedorf, Switzerland) and the i-control 1.12 software.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Amplicon library preparation

The amplicon libraries were generated with a two-step PCR protocol. The first reaction consisted of 1 µL of genomic DNA as template, 0.5 µM forward oligonucleotide (G-40604/ G-40605/ G-40606/ G-40606/ G-42015), 0.5 µM reverse oligonucleotide (G-40607/ G-40608/ G-40609/ G-42016), 1x Phusion HF buffer (1.5 mM $MgCl_2$) (Thermo FIsher scientific, Waltham, MA, United States), 0.2 mM dNTPs (Thermo Fisher Scientific, #R0182, Waltham, MA, United States) and 0.02 U/µL Phusion High-Fidelity DNA polymerase (Thermo Fisher scientific, #F530, Waltham, MA, United States) to a final volume of 25 µL.

The second PCR amplification consisted of 2.5 µL of the cleaned PCR product of the previous reaction, 0.5 µM forward oligonucleotide (G-40610), 0.25 µM reverse oligonucleotide that had one of the 96 indices (Lundberg et al. 2013), 1x Phusion HF buffer (1.5 mM $MgCl_2$) (Thermo Fisher Scientific, Waltham, MA, United States), 0.2 mM dNTPs (Thermo Fisher Scientific, #R0182, Waltham, MA, United States) and 0.02 U/µL Phusion High-Fidelity DNA polymerase (Thermo Fisher Scientific, #F530, Waltham, MA, United States) to a final volume of 25 µL.

Sequencing libraries were prepared using Q5® High-Fidelity DNA polymerase (New England BioLabs, #M0491, Ipswich, MA, United States) in a final concentration of 0.02 U/µL along with 1x Q5 reaction buffer (2 mM $MgCl_2$). The rest of the reaction components (DNA template, dNTPs) remained the same.

The MJ Research PTC225 Peltier (marshall scientific Hampton, NH, USA) or the BIO-RAD C1000 Touch (Hercules, CA, United States) thermal cyclers were used. The PCR programs had 15 cycles in which the denaturing temperature was 94°C for 30 s, followed by annealing at 60°C for 30 s, and extension at 72°C for 10 s for program 1, and 15 s for program 2. A final extension step was at 72°C for 2 minutes.

Bead clean up

For the generation of the amplicon libraries, two bead-based clean-up steps were carried out using SPRI beads (Magnetic SpeedBeads™, GE Healthcare No.:65152105050250, Chicago, IL,

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

USA). The first PCR product was cleaned using a ratio of 1:0.9 and resuspended in 17 µL of ddH2O. The second PCR product was cleaned using the same ratio of beads and resuspended in 27 µL. The ratios of clean ups were chosen after optimisation.

Quant-iT™ PicoGreen® dsDNA assay

Amplicons were quantified using the Quant-iT™ PicoGreen (Invitrogen, Carlsbad, CA, USA) dsDNA assay. One µL of each amplicon was used according to the manufacturer's instructions for the quantification. The samples were prepared in black 96 well, F-bottom, non-binding microplates (96 wells, Greiner Bio-one, ref:655906, Kremsmuenster, Austria), and the TECAN Infinite M200 PRO plate reader was used for all the measurements using the Magellan 7.2 software.

Pooling

To roughly normalize samples when pooling, the DNA concentration of all samples in each 96 well plate was first measured fluorometrically (PicoGreen assay). First, all the 96 samples from each plate were pooled, creating subpools. From samples with concentrations less than half of the mean, 6 µL were taken. From samples with concentrations more than twice the mean, 1.5 µL was taken. For all other samples falling between these extremes, 3 µL was taken. After each plate was pooled in this way, the subpools representing entire plates were again measured fluorometrically (Qubit dsDNA-HS assay) (Thermo Fisher scientific, Waltham, MA, United States) and pooled in an equimolar manner to create a final pool containing all samples. The concentration of the subpools and the final pool were evaluated using the Qubit dsDNA-HS assay. Each pool was analyzed on the Agilent 2100 Bioanalyzer (Agilent Technologies, Santa Clara, CA, USA) according to the manufacturer's instructions. DNA1000 chips were used for the amplicon libraries.

Illumina MiSeq sequencing

The libraries were diluted for Illumina sequencing following manufacturers' protocols and sequenced on the MiSeq platform using MiSeq reagent kit v2 (300-cycles) (MS-102-2002).

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Data availability

All data in this manuscript has been deposited in the European Nucleotide Archive (ENA). It can be accessed under the project number PRJEB39078. At https://www.ebi.ac.uk/ena.

Author Contributions

ES, RS and DW planned the study. RS and ES prepared initial plasmids for the CRISPR/Cas9 system. ES performed all cloning, plant transformations, DNA extractions and prepared all the libraries. JR designed and wrote PlexSeq for sample demultiplexing. ES carried out data analysis. ES conducted *Hpa* in infections. ES wrote the manuscript. ES, JR, RS and DW revised the manuscript.

Acknowledgements

We thank Derek Lundberg for discussions, suggestions regarding PCR and library preparation and comments on manuscript, Talia Karasov for comments on the manuscript, Christa Lanz and Julia Hildebrandt for technical help sequencing, Ilja Berzukov for illumina demultiplexing. Funding was provided by the Max Planck Society.

Competing interests

The authors declare no competing interests.

References

1. Gaj T, Gersbach CA, Barbas CF 3rd. ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. Elsevier; 2013;31:397–405.

2. Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science. American Association for the Advancement of Science; 2012;337:816–21.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

3. Gasiunas G, Barrangou R, Horvath P, Siksnys V. Cas9-crRNA ribonucleoprotein complex mediates specific DNA cleavage for adaptive immunity in bacteria. Proc Natl Acad Sci U S A. 2012;109:E2579–86.

4. Phillips JW, Morgan WF. Illegitimate recombination induced by DNA double-strand breaks in a mammalian chromosome. Mol Cell Biol. Am Soc Microbiol; 1994;14:5794–803.

5. Ma Y, Lu H, Tippin B, Goodman MF, Shimazaki N, Koiwai O, et al. A biochemically defined system for mammalian nonhomologous DNA end joining. Mol Cell. Elsevier; 2004;16:701–13.

6. Liang F, Han M, Romanienko PJ, Jasin M. Homology-directed repair is a major double-strand break repair pathway in mammalian cells. Proc Natl Acad Sci U S A. National Acad Sciences; 1998;95:5172–7.

7. Gratz SJ, Ukken FP, Rubinstein CD, Thiede G, Donohue LK, Cummings AM, et al. Highly specific and efficient CRISPR/Cas9-catalyzed homology-directed repair in Drosophila. Genetics. Genetics Soc America; 2014;196:961–71.

8. Chang N, Sun C, Gao L, Zhu D, Xu X, Zhu X, et al. Genome editing with RNA-guided Cas9 nuclease in zebrafish embryos. Cell Res. nature.com; 2013;23:465–72.

9. Li D, Qiu Z, Shao Y, Chen Y, Guan Y, Liu M, et al. Heritable gene targeting in the mouse and rat using a CRISPR-Cas system. Nat Biotechnol. nature.com; 2013;31:681–3.

10. Auer TO, Duroure K, De Cian A, Concordet J-P, Del Bene F. Highly efficient CRISPR/Cas9-mediated knock-in in zebrafish by homology-independent DNA repair. Genome Res. genome.cshlp.org; 2014;24:142–53.

11. Platt RJ, Chen S, Zhou Y, Yim MJ, Swiech L, Kempton HR, et al. CRISPR-Cas9 knockin mice for genome editing and cancer modeling. Cell. Elsevier; 2014;159:440–55.

12. Canver MC, Bauer DE, Dass A, Yien YY, Chung J, Masuda T, et al. Characterization of genomic deletion efficiency mediated by clustered regularly interspaced palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. J Biol Chem. ASBMB; 2014;289:21312–24.

13. Feng Z, Zhang B, Ding W, Liu X, Yang D-L, Wei P, et al. Efficient genome editing in plants using a CRISPR/Cas system. Cell Res. search.proquest.com; 2013;23:1229–32.

14. Feng Z, Mao Y, Xu N, Zhang B, Wei P, Yang D-L, et al. Multigeneration analysis reveals the inheritance, specificity, and patterns of CRISPR/Cas-induced gene modifications in Arabidopsis. Proc Natl Acad Sci U S A. National Acad Sciences; 2014;111:4632–7.

15. Hyun Y, Kim J, Cho SW, Choi Y, Kim J-S, Coupland G. Site-directed mutagenesis in Arabidopsis thaliana using dividing tissue-targeted RGEN of the CRISPR/Cas system to generate heritable null alleles. Planta. Springer; 2015;241:271–84.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

16. Peterson BA, Haak DC, Nishimura MT, Teixeira PJPL, James SR, Dangl JL, et al. Genome-Wide Assessment of Efficiency and Specificity in CRISPR/Cas9 Mediated Multiple Site Targeting in Arabidopsis. PLoS One. Public Library of Science; 2016;11:e0162169.

17. Fauser F, Schiml S, Puchta H. Both CRISPR/Cas-based nucleases and nickases can be used efficiently for genome engineering in Arabidopsis thaliana. Plant J. Wiley Online Library; 2014;79:348–59.

18. Xie K, Yang Y. RNA-Guided Genome Editing in Plants Using a CRISPR–Cas System. Mol Plant. Elsevier; 2013;6:1975–83.

19. Ablain J, Durand EM, Yang S, Zhou Y, Zon LI. A CRISPR/Cas9 vector system for tissue-specific gene disruption in zebrafish. Dev Cell. Elsevier; 2015;32:756–64.

20. Sanger F, Coulson AR. A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. J Mol Biol. Elsevier; 1975;94:441–8.

21. Strauss EC, Kobori JA, Siu G, Hood LE. Specific-primer-directed DNA sequencing. Anal Biochem. Elsevier; 1986;154:353–60.

22. Mashal RD, Koontz J, Sklar J. Detection of mutations by cleavage of DNA heteroduplexes with bacteriophage resolvases. Nat Genet. 1995;9:177–83.

23. Lundberg DS, Yourstone S, Mieczkowski P, Jones CD, Dangl JL. Practical innovations for high-throughput amplicon sequencing. Nat Methods. 2013;10:999–1002.

24. Wildermuth MC, Dewdney J, Wu G, Ausubel FM. Isochorismate synthase is required to synthesize salicylic acid for plant defence. Nature. 2001;414:562–5.

25. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43:956–63.

26. Li J-F, Norville JE, Aach J, McCormack M, Zhang D, Bush J, et al. Multiplex and homologous recombination-mediated genome editing in Arabidopsis and Nicotiana benthamiana using guide RNA and Cas9. Nat Biotechnol. nature.com; 2013;31:688–91.

27. Kroj T, Savino G, Valon C, Giraudat J, Parcy F. Regulation of storage protein gene expression in Arabidopsis. Development. 2003;130:6065–73.

28. Gao X, Chen J, Dai X, Zhang D, Zhao Y. An Effective Strategy for Reliably Isolating Heritable and Cas9-Free Arabidopsis Mutants Generated by CRISPR/Cas9-Mediated Genome Editing. Plant Physiol. Am Soc Plant Biol; 2016;171:1794–800.

29. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25:2078–9.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

30. Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing [Internet]. arXiv [q-bio.GN]. 2012. Available from: http://arxiv.org/abs/1207.3907

31. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. Oxford Univ Press; 2011;27:2156–8.

32. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29:24–6.

33. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. Oxford Univ Press; 2013;14:178–92.

34. Huang WE, Huang L, Preston GM, Naylor M, Carr JP, Li Y, et al. Quantitative in situ assay of salicylic acid in tobacco leaves using a genetically modified biosensor strain of Acinetobacter sp. ADP1. Plant J. Wiley Online Library; 2006;46:1073–83.

35. Defraia CT, Schmelz EA, Mou Z. A rapid biosensor-based method for quantification of free and glucose-conjugated salicylic acid. Plant Methods. 2008;4:28.

36. Boel A, Steyaert W, De Rocker N, Menten B, Callewaert B, De Paepe A, et al. BATCH-GE: Batch analysis of Next-Generation Sequencing data for genome editing assessment. Sci Rep. 2016;6:30330.

37. Pinello L, Canver MC, Hoban MD, Orkin SH, Kohn DB, Bauer DE, et al. CRISPResso: sequencing analysis toolbox for CRISPR genome editing [Internet]. bioRxiv. 2016 [cited 2017 Jan 20]. p. 031203. Available from: http://www.biorxiv.org/content/early/2016/02/07/031203.abstract

38. Lindsay H, Burger A, Biyong B, Felker A, Hess C, Zaugg J, et al. CrispRVariants charts the mutation spectrum of genome engineering experiments. Nat Biotechnol. 2016;34:701–2.

39. Brocal I, White RJ, Dooley CM, Carruthers SN, Clark R, Hall A, et al. Efficient identification of CRISPR/Cas9-induced insertions/deletions by direct germline screening in zebrafish. BMC Genomics. 2016;17:259.

40. Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol [Internet]. 2018; Available from: http://dx.doi.org/10.1038/nbt.4192

41. Lampropoulos A, Sutikovic Z, Wenzl C, Maegele I, Lohmann JU, Forner J. GreenGate - A Novel, Versatile, and Efficient Cloning System for Plant Transgenesis. PLoS One. Public Library of Science; 2013;8:e83043.

bioRxiv preprint doi: https://doi.org/10.1101/2020.06.25.171538; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

42. Sprunck S, Rademacher S, Vogler F, Gheyselinck J, Grossniklaus U, Dresselhaus T. Egg Cell–Secreted EC1 Triggers Sperm Cell Activation During Double Fertilization. Science. American Association for the Advancement of Science; 2012;338:1093–7.

43. Hyun Y, Yun H, Park K, Ohr H, Lee O, Kim D-H, et al. The catalytic subunit of Arabidopsis DNA polymerase α ensures stable maintenance of histone modification. Development. dev.biologists.org; 2013;140:156–66.

44. Wu R, Lucke M, Jang Y-T, Zhu W, Symeonidi E, Wang C, et al. An efficient CRISPR vector toolbox for engineering large deletions in Arabidopsis thaliana. Plant Methods. plantmethods.biomedcentral.com; 2018;14:65.

45. Clough SJ, Bent AF. Floral dip: a simplified method forAgrobacterium-mediated transformation ofArabidopsis thaliana. Plant J. Wiley Online Library; 1998;16:735–43.

46. Edwards K, Johnstone C, Thompson C. A simple and rapid method for the preparation of plant genomic DNA for PCR analysis. Nucleic Acids Res. 1991;19:1349.

47. Marek G, Carver R, Ding Y, Sathyanarayan D, Zhang X, Mou Z. A high-throughput method for isolation of salicylic acid metabolic mutants. Plant Methods. 2010;6:21.