Title

DAJIN-assisted multiplex genotyping to validate the outcomes of CRISPR-Cas-based genome editing

Author information

Akihiro Kuno[1,2†*], Yoshihisa Ikeda[3,4†], Shinya Ayabe[5†], Kanako Kato[4], Kotaro Sakamoto[2,6], Sayaka Suzuki[2,7], Kento Morimoto[8], Arata Wakimoto[1,2], Natsuki Mikami[9], Miyuki Ishida[4], Natsumi Iki[4], Yuko Hamada[4], Megumi Takemura[4], Yoko Daitoku[4], Yoko Tanimoto[4], Tra Thi Huong Dinh[4], Kazuya Murata[2,4], Michito Hamada[1,4], Atsushi Yoshiki[5], Fumihiro Sugiyama[4], Satoru Takahashi[1,4], Seiya Mizuno[4*]

* Correspondence: akuno@md.tsukuba.ac.jp; konezumi@md.tsukuba.ac.jp

† Akihiro Kuno, Yoshihisa Ikeda, and Shinya Ayabe contributed equally to this work.

Affiliations

[1]Department of Anatomy and Embryology, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[2]Ph.D Program in Human Biology, School of Integrative and Global Majors, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[3]Doctoral program in Biomedical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[4]Laboratory Animal Resource Center, Transborder Medical Research Center, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[5]Experimental Animal Division, RIKEN BioResource Research Center, 3-1-1 Koyadai, Tsukuba, Ibaraki 305-0074, Japan

[6]Department of Computer Science, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[7]Bioinformatics Laboratory, Faculty of Medicine, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[8]Doctoral program in Medical Sciences, Graduate School of Comprehensive Human Sciences, University of Tsukuba, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

[9]School of Medical Sciences, University of Tsukuba, Tsukuba, Japan, 1-1-1 Tennodai, Tsukuba, Ibaraki 305-8575, Japan

Abstract

Genome editing induces various on-target mutations. Accurate identification of mutations in founder mice and cell clones is essential to perform reliable genome editing experiments. However, no genotyping method allows the comprehensive analysis of diverse mutations. We developed a genotyping method with an on-target site analysis software named Determine Allele mutations and Judge Intended genotype by Nanopore sequencer (DAJIN) that can automatically identify and classify diverse mutations, including point mutations, deletions, inversions, and knock-in. Our genotyping method with DAJIN can handle approximately 100 samples within a day and may become a new standard for validating genome editing outcomes.

Keywords

Long-read sequencing, Machine learning, Bioinformatic software, Complex genome-edited mutations, Nanopore signal annotation, Allele validation, CRISPR-Cas

Background

The development of new technologies such as CRISPR-Cas has facilitated genome editing of any species or cell type. Nucleases such as Cas9 and FokI and deaminase fused with Cas9 have been used for introducing DNA double-strand breaks and performing base editing, respectively [1-3]. However, as double-strand break repair pathways are regulated by host cells [4], verifying the result and selecting desired mutated alleles for precise genome editing are essential. Thus, studies have been focusing on the prediction of the outcomes of genome editing, optimization of editing strategy, or engineering of novel variants with high specificity [5].

Recent work has uncovered that genome editing can induce various on-target events such as inversions, deletions, and endogenous and exogenous DNA insertions as well as indels and substitutions at, and in some cases, away from the target site [6-9]. Short-range assessments with targeted PCR amplification and Sanger sequencing likely to miss long-range mutation events, which may result in pathogenic phenotypes through unintended changes in gene expression [10]. Furthermore, there is a possibility of gene conversion between homologous regions following genomic DNA cleavage [11-13]. Although targeted long-read sequencing allows the detection of complex on-target mutations over several kilobases [9, 14], this method has instrumental limitations such as error rates that need to be overcome to validate the target locus to a single-base level [15].

Multiple alleles exist in a population of cells or single animals that have undergone genome editing. In most cases, animals born following editing events at early embryonic stages are mosaic [16]. Heterogeneous cell populations can be obtained by genome editing of cultured cells or delivery of genome editing tools to somatic cells [17, 18]. Besides the unexpected editing results described above, demultiplexing of highly homologous mutated alleles is required to separate the signals of each allele from genetically engineered samples. Cell populations with incorrectly edited alleles need to be detected and excluded for therapeutic uses to ensure precise genome editing [19]. In some cases, unintended alleles with similar genetic impact may be tolerated depending on the purpose of genome editing, e.g. generation of null alleles through the deletion of critical exon(s) by using multiple guide RNAs, resulting in multiple patterns in the total deleted length [20]. The assessment of on-target editing outcomes and selection of correct, precisely edited alleles lead to not only the efficient production and breeding of founder animals and their offspring but also efficient *in vivo* and *ex vivo* engineering. However, the subcloning of amplified products is laborious, and tracking of indels by decomposition analysis of Sanger sequencing data does not support large-scale mutations [21]. Moreover, short-range PCR analysis cannot identify multiple intended or unwanted mutations in *cis* or in *trans* [22, 23]. Thus, more accessible and high-throughput methods of validation of genome editing outcomes are needed to detect multiple, unpredictable editing events.

Herein, we describe a novel method for analysing genome editing outcomes, in which long-chain PCR products with barcodes obtained using two-step long-range PCR were used as samples, and allele validation was performed using our original software named Determine Allele mutations and Judge Intended genotype by Nanopore sequencer (DAJIN) that enables the comprehensive analysis of long reads

generated using the nanopore long-read sequencing technology. DAJIN identifies and quantifies allele numbers along with their mutation patterns and reports consensus sequences to visualise mutations in alleles at single-nucleotide resolutions. Moreover, it allows multiple sample processing, and approximately 100 samples can be processed within a day. Because of these features, our strategy with DAJIN can validate the quality of genome-edited samples to efficiently select animals or clones with intended results and has the potential to contribute to more precise genome editing.

Results

A brief summary of DAJIN

DAJIN allows to capture diverse mutations at single-nucleotide resolution in multiple genome-edited samples. The output of DAJIN includes the percentage of predicted alleles in each sample and the consensus sequences of each allele with mutant bases. The overall workflow of DAJIN is presented in Fig. 1a.

DAJIN requires (1) a multi-FASTA formatted file describing possible alleles; (2) FASTQ files of nanopore sequencing reads; (3) gRNA sequences, including the protospacer adjacent motif (PAM); and (4) a reference genome name such as hg38 and mm10. The FASTA file needs to include the sequences of wild-type (WT) and the desired allele for genome editing (Target). The FASTQ files need to include WT control to generate simulation reads and fix nanopore sequencing errors. After the data are input, DAJIN generates simulation reads by using NanoSim [24] according to the sequences in the multi-FASTA file. Next, DAJIN pre-processes both the simulated and real reads to convert nucleotides to strings that can be used to train a model. Subsequently, the simulated reads are used to train a deep neural network (DNN)

model to detect abnormal reads and classify allele types. DAJIN defines abnormal alleles as sequences that are different from those of the user-inputted FASTA file by using the unsupervised anomaly detection method. It defines allele types by using the sequence labels following the header ('>') of the FASTA file. Thus, each read is annotated as at least 'Target', 'WT', and 'Abnormal'. Next, DAJIN conducts clustering to estimate the alleles in each allele type according to the mutation patterns at single-nucleotide resolution. Finally, it reports the consensus sequence to visualise the mutations in each allele and labels the alleles. The details are described in Methods and Additional file 1: Figs. S1–5.

The outputs of DAJIN are shown in Fig. 1b. DAJIN reports the number and type of each allele and the percentage of reads corresponding to that allele in the comma-separated value (CSV) and Portable Document Format (PDF) format. The consensus sequences for each allele are output in HTML and FASTA format. This facilitates the understanding of the mutation site and the type of mutation (insertion, deletion, and replacement). All the nanopore sequence reads with each allele are outputted as BAM files. This allows the identification of other genomic loci that have unexpected mutations.

CRISPR-Cas genome editing has been reported to induce unexpectedly large indels, which might be overlooked by conventional short PCR-based genotyping methods. Conversely, a nanopore long-read sequencer can capture large indels within its amplicon size, which allows the detection of abnormal alleles. Therefore, we incorporated an anomaly detection method into DAJIN and determined whether unexpected abnormal alleles could be identified (Additional file 1: Fig. S6a). We simulated nanopore sequencing reads with a deletion in the range of 10 bp to 200 bp based on the *Cables2* flox knock-in design (Additional file 1: Fig. S6a). To evaluate the

effect of MIDS conversion on abnormal allele detection, we pre-processed the simulated reads with or without MIDS conversion and analysed the data using Uniform Manifold Approximation and Projection (UMAP) [25] and Local Outlier Factor (LOF) [26] by using outputs from the last fully connected (FC) layer (Additional file 1: Fig. S6b). The results showed that UMAP visualisation revealed a cluster of 50 bp deletion with MIDS conversion, which was unclear without MIDS conversion (Additional file 1: Fig. S6c). Next, we performed binary classification of normal and abnormal alleles and investigated the accuracy of abnormal allele detection. The results showed that DAJIN labelled more than 50 bp deletion as 'Abnormal' only with MIDS conversion, but not without, indicating that the MIDS conversion improved the sensitivity of abnormal allele detection (Additional file 1: Fig. S6d). Next, we assessed the accuracy of the classification of simulation reads by DAJIN. Although DAJIN uses a lightweight model that consists of fewer layers than the typical deep learning models, it could accurately classify alleles in all genome editing designs conducted in this study (Additional file 2: Table. S1).

DAJIN identifies target alleles with point mutation

We tested the point mutation (PM) design to evaluate whether DAJIN can capture single nucleotide substitutions induced by CRISPR-Cas9 genome editing. We induced *Tyr* c.140G>C PM by using C57BL/6J mouse fertilized eggs and analysed 13 founder mice (barcode (BC) 01–13) and used a WT mouse as a control (BC32). In addition, a C57BL/6J-*Tyr*^{em2Utr} mouse was added to BC31 as a control for 'Mutated WT'. This mutant mouse carries the *Tyr* c.230G>T PM [16]. Next, we amplified a 2845 bp sequence that included target nucleotides and analysed nanopore sequencing by using DAJIN (Fig. 2a). Because the PM genome-editing design may generate WT and PM alleles, DAJIN reported 'WT', 'PM', and 'Abnormal' allele types. 'WT' and 'PM' alleles

were labelled as ‘Intact WT’ and ‘Intended PM’, respectively, when the DAJIN consensus sequence perfectly matched the sequences of ‘WT’ and ‘PM’ described in the user-inputted FASTA file, whereas their unexpected mutations were labelled as ‘Mutated WT’ and ‘Mutated PM’. DAJIN also reported ‘Abnormal’ alleles with large indels (approximately more than 50 bp). It reported the percentages of the predicted allele types and identified the two mice (BC08, BC12) with the desired PM (‘Intended PM’; Fig. 2b). Read visualisation using IGV software showed that DAJIN accurately captured the c.140G>C target PM in BC08 and BC12. In addition, DAJIN detected an unexpected 2 bp insertion in BC08 at 23 bp downstream from the target PM and labelled these reads as ‘Mutated PM’, suggesting an allele with unexpected mutations in an ideal PM allele (Fig. 2c). To evaluate the mutation loci and types within the entire amplicon sequence, we, next, examined the DAJIN consensus sequence of the dominant alleles of BC08 and BC12, which revealed the ‘Intended PM’ allele in BC12 (Fig. 2d). Sanger sequencing of BC12 at the PM locus supported the induction of the target PM (Fig. 2e). Conversely, BC08 included an unexpected 2 bp (CT) insertion as well as the target PM (Fig. 2f). The same CT insertion was validated by Sanger sequencing (Fig. 2e). Sanger sequencing showed a waveform intensity ratio of about 5:1 between the target PM with CT insertion allele and that without the insertion allele. The same ratio was observed with DAJIN. This consistency indicated that DAJIN correctly quantifies the percentage of alleles.

To further validate the analytical capability of DAJIN, we generated two more PM mice, *Tyr* c.316G>C and *Tyr* c.308G>C by using early embryo genome editing (Additional file 1: Fig. S7a). For *Tyr* c.316 and c.308 projects, DAJIN reported that 1 of 6 mice (BC18) and 8 of 11 mice (BC21, BC22, BC23, BC24, BC26, BC29, and BC30) had 'Intended PM' (Additional file 1: Fig. S7b). Tracking of Insertions, DEletions, and

Recombination event (TIDER) analysis [27] of the Sanger sequencing data from the PCR products of these nine mice showed that seven of them had the target PM (Additional file 1: Fig. S7c). Next, because Cas9 may induce unexpected large indels, we evaluated whether DAJIN correctly captured abnormal alleles. We conducted short and long PCRs for detecting small and large indel mutations (Additional file 1: Fig. S7a). The PCR revealed 17 samples with aberrant PCR bands, which was consistent with the result of DAJIN (Additional file 1: Fig. S7b,c,d).

We further analysed BC02 and BC10, the alleles for which were reported as ‘Abnormal’ by DAJIN. Visualisation of the reads revealed that BC02 and BC10 had approximately 50 bp and 40 bp insertions, respectively (Additional file 1: Fig. S8a). To confirm the insertion alleles, we conducted PCR and found that BC02 and BC10 had 50 bp and 40 bp larger bands, respectively, than that in WT control (Additional file 1: Fig. S8b,c). This result indicated that DAJIN can correctly annotate alleles with 40–50 bp insertion as ‘Abnormal’ alleles. Furthermore, DAJIN annotated one WT and two abnormal alleles in BC16 (Additional file 1: Fig. S7b). Visualisation of the reads of BC16 clearly showed that it had three alleles with different deletion sizes (Additional file 1: Fig. S9), which was also confirmed by PCR (Additional file 1: Fig. S7d). This result suggested that DAJIN allowed allele clustering for large mutations, whereas conventional short PCR-based genotyping could not.

Next, we used TIDER to verify whether the 19 mice (generated in the above three *Tyr* projects) reported by DAJIN as having no 'Intended PM' actually contained the target PM. In this TIDER analysis, no target PM was detected in any mouse except for one (BC20). As with BC12, DAJIN revealed that almost all (93.6%) the nanopore sequencing reads of BC21 are ‘Intended PM’. The DAJIN consensus sequence reported the c.308G>C PM, and we detected a single waveform of the target PM by

using Sanger sequence analysis (Additional file 1: Fig. S10). Moreover, DAJIN correctly identified *Tyr* c.230G>T PM in BC31, which was used as the positive control (Additional file 1: Fig. S11). These results suggest that DAJIN can distinguish alleles from single nucleotide variants to large indels as well as identify the target PM.

DAJIN identifies target knock-out alleles at single-nucleotide resolution

Next, we designed a deletion mutation at the *Prdm14* gene locus. We intended to excise exon 6 of *Prdm14* by two-cut strategy with CRISPR-Cas9 system [28] (Fig. 3a). The predicted deletion size was 1043 bp length, may yielding an inverted allele as a byproduct. Thus, DAJIN reported 'WT', 'Deletion (Del)', 'Inversion', and 'Abnormal' allele types. 'Del' and 'WT' alleles were labelled as 'Intended Del' and 'Intact WT' when the DAJIN consensus sequence perfectly matched the sequences of 'Del' and 'WT' in the user-inputted FASTA, whereas their unexpected mutations were labelled as 'Mutated Del' and 'Mutated WT'. Moreover, DAJIN reported 'Abnormal' alleles with large indels (approximately more than 50 bp). We generated 10 *Prdm14* deletion founder mice (BC16–25) and analysed them using DAJIN with a WT mouse as a control (BC26); of the 10 mice, 5 (BC16, BC18, BC20, BC23, and BC24) contained 'Mutated Del' allele (Fig. 3b). Next, we evaluated BC18 and BC23 as DAJIN predicted that they contained 'Mutated Del' allele. Read visualisation showed that DAJIN discriminated 'Abnormal' alleles with 100–200 bp larger deletion than the target deletion (Fig. 3c). Furthermore, the DAJIN consensus of 'Mutated Del' alleles showed that BC18 had a 1-bp deletion and BC23 included 7-bp insertion and 1-bp substitution, respectively, at the joint site. The same mutations were validated using Sanger sequencing (Fig. 3d, e).

We then evaluated the phenotypes of BC18 and BC23 mice. The deletion of *Prdm14* inhibits primordial germ cell differentiation and causes the complete depletion of germ cells in adult female and male mice [29]. To verify this phenomenon in BC18 and BC23 male mice with targeted deletion features identified by DAJIN, we used immunostaining to evaluate whether PLZF1-positive spermatogonia were present. In these mice, many Vimentin-positive somatic Sertoli cells were identified; however, as expected, no spermatogonia were detected (Fig. 3f).

To confirm whether DAJIN can be used for analysing the mutations induced during genome editing by using the CRISPR-Cas12a system [30], we generated *Prdm14* KO mice by using Cas12a (Additional file 1: Fig. S12a). In all, 15 founder mice were obtained, and DAJIN analysis revealed that four of them (BC10, BC11, BC12, and BC13) had ‘Mutated Del’, suggesting an allele with unexpected mutations in an ideal deletion allele (Additional file 1: Fig. S12b). We validated these mice contained the target deletion allele by using conventional PCR analysis. In addition, in the Cas9 group, PCR results showed that DAJIN correctly predicted that five mice (BC16, BC18, BC20, BC23, and BC24) had the ‘Mutated Del’ alleles (Additional file 1: Fig. S12c,d).

To evaluate the versatility of DAJIN at other genomic loci and at different cleavage widths, we established knock-out (KO) mice for *Ddx4* gene by using the two-cut strategy with Cas9 and Cas12a system. *Ddx4* KO was designed to cleave 3377 bp, including exons 11–15. We obtained 21 founder mice and analysed the 5221 bp PCR amplicon, including the target region, by using DAJIN (Additional file 1: Fig. S13a). DAJIN reported that one mouse (BC27) subjected to Cas12a-based genome editing and four mice (BC36, BC39, BC44, and BC46) subjected to Cas9-based genome editing carried the ‘Mutated Del’ allele (Additional file 1: Fig. S13b). The presence of these alleles was also confirmed by the electrophoresis of PCR products (Additional file

1: Fig. S13c and d). Furthermore, we generated KO mice for the *Stx2* gene by using the two-cut strategy to cleave 727 bp, including exon 5 of *Stx2* [31]. DAJIN reported that 13 of 29 founder mice (BC01, BC03, BC04, BC05, BC07, BC09, BC14, BC15, BC20, BC21, BC22, BC23, and BC24) had 'Mutated Del' allele (Additional file 1: Fig. S14a, b), and this DAJIN report was consistent with the electrophoresis result of the PCR products (Additional file 1: Fig. S14c,d). In the *Stx2* analysis, DAJIN detected the 'Inversion' allele in three mice (BC08, BC16, and BC17). To verify the presence of inversion allele, we performed PCR for amplifying the genome region containing the inversion junction sites (Additional file 1: Fig. S14e). The inversion band was found in all the three mice (Additional file 1: Fig. S14f). The 1 bp (A) insertion at the inversion junction site was found in the DAJIN consensus sequence of BC17. This insertion was also confirmed by Sanger sequencing (Additional file 1: Fig. S14g). These results suggested that DAJIN can accurately identify single-nucleotide variants even in inversion alleles.

DAJIN identifies target flox knock-in alleles at single-nucleotide resolution

Cre-LoxP-based conditional KO experiments are mostly performed to analyse gene function under specific conditions. Genome editing for generating floxed alleles requires *cis* knock-in at two loci simultaneously, which lowers the generation efficiency. Moreover, genotyping of the *cis* knock-in is difficult and error-prone owing to the need to identify *cis* mutations at several kilobases of the DNA region. Moreover, the generation of floxed alleles by using single-strand oligodeoxynucleotides (ssODNs) as donor of the knock-in sequence occasionally leads to the introduction of unintended mutations in a critical LoxP sequence because of the error in the synthesis process and its secondary structure [32]. Because of these difficulties, no genotyping method is

currently available to evaluate comprehensively and accurately flox mutations induced by genome editing in one step. Therefore, we determined whether DAJIN can genotype floxed alleles at single-nucleotide resolution.

Before evaluating the actual genome-edited outcomes, we performed validation experiments by using plasmid vectors that have completely defined sequences. We generated six types of plasmids with LoxP sequences of (1) 'Intended flox', (2) 1-bp insertion in left LoxP, (3) 1-bp deletion in left LoxP, (4) 1-bp substitution in left LoxP, (5) 1-bp substitution in right LoxP, and (6) 1-bp substitution in both LoxPs (Fig. 4a). To simulate the experimental design by using mouse genomic DNA as a PCR template, we mixed the WT genomic DNA with each plasmid and mimicked the heterozygous genotype. These mixed DNA samples were used as a PCR template. The PCR products were used to perform nanopore sequencing and were analysed by DAJIN (Fig. 4b). DAJIN reported the allele percentage of 13 samples, including the two replicates of 'Intended flox', 1-bp insertion, 1-bp deletion, 1-bp left LoxP substitution, 1-bp right LoxP substitution, 1-bp left and right LoxP substitution, and a WT control (BC42). The results showed that DAJIN correctly discriminated between ‘Intended flox’ and ‘Mutated flox’. For ‘Mutated flox’, DAJIN’s DNN model annotated the allele as flox, but the consensus sequences of DAJIN show small indel mutations in the knocked-in sequence. In addition, DAJIN reported the existence of intact WT alleles, and the proportion of WT and LoxP alleles reflected the designed allele frequency (Fig. 4c). The consensus sequences of DAJIN correctly discovered all types of single nucleotide variants that we induced in the knock-in sequence. (Fig. 4d). These results indicated DAJIN can identify mutations at single-nucleotide resolution in the knock-in sequence that is not present in the reference genome.

Next, we assessed the ability of DAJIN to genotype the flox mutation induced by genome editing in actual founder mice. To induce a mutation that floxed exon 5 of *Cables2* gene, we simultaneously cut introns 5 and 6 of *Cables2* and knocked in the two LoxPs to their cut sites via homology-directed repair by using a single plasmid DNA donor. In this flox knock-in design, genome editing potentially generates predictable six types of alleles, i.e. WT, flox, Left LoxP, Right LoxP, Inversion, and Deletion (Fig. 5a). DAJIN reported these alleles as 'Intact WT', 'Intended flox', 'Left LoxP', 'Right LoxP', 'Inversion', and 'Deletion', respectively. In the case of 'WT' and 'flox', 'Intact WT' and 'Intended flox' were annotated when the DAJIN consensus sequence perfectly matched the sequences of 'WT' and 'flox' in the user-inputted FASTA, whereas unexpected mutations were labelled as 'Mutated WT' and 'Mutated flox'. Moreover, DAJIN reported 'Abnormal' alleles as those with large indels (approximately more than 50 bp). According to the design, we obtained 20 founder mice (BC01–BC20) and analysed them using DAJIN with WT mouse as a control (BC42). DAJIN reported that eight mice (BC06, BC10, BC11, BC12, BC14, BC17, BC18, and BC20) contained the 'Intended flox' allele, and 11 mice (BC01, BC05, BC06, BC09, BC11, BC12, BC13, BC17, BC18, BC19, and BC20) contained 'Deletion' alleles (Fig. 5b). Next, we evaluated whether DAJIN correctly captures the genotypes. Since the *Asc*I or *Eco*RV recognition site was also knocked in immediately next to the LoxP sequence, we conducted PCR–RFLP digestion of the left and right knock-in sites by using *Asc*I and *Eco*RV, respectively (Fig. 5c). The results were completely consistent with the DAJIN reports (Fig. 5d). We also evaluated the 'Deletion' alleles by using standard PCR (Fig. 5e), and the results were also consistent with the DAJIN reports (Fig. 5f). Moreover, we compared the DAJIN consensus sequence with the Sanger sequence to validate that DAJIN can truly capture the 'Intended flox' sequence. The consensus sequence of

DAJIN for BC14 showed that the entire 2724 bp was intact, including the left and right LoxP sites (Fig. 5g). Sanger sequencing revealed that both left and right knock-in sequences were intact, corresponding to the DAJIN consensus sequence (Fig. 5h). These results indicated that DAJIN correctly identified the intended floxed mice.

Next, to confirm whether the next generation inherits the allele determined using DAJIN, BC11, BC12, BC13, and BC14, which were determined to have an 'Intended flox' allele by DAJIN, were mated with WT to obtain F1 mice. (Additional file 1: Fig. S15). The results showed that the genotypes of F1 mice of BC14 were heterozygous on flox/WT, suggesting that BC14 has homozygous floxed alleles in the germline. Moreover, F1 mice from BC11, BC12, BC13, and BC18 had the 'Intended flox' allele, as was revealed by the DAJIN report. In summary, the progeny test provides evidence that DAJIN correctly captured the genotypes of founder mice.

DAJIN detected the 'Inversion' allele in five samples (BC02, BC05, BC07, BC10, and BC16; Fig. 5b). To verify the presence of inversion allele, we performed PCR to amplify the genomic region containing the inversion junction site. The results revealed the inversion band in the same five samples corresponding to those mentioned in DAJIN's report (Additional file 1: Fig. S16a, b). Furthermore, the consensus sequence of BC02 revealed a 1-bp substitution at the inversion junction site. Sanger sequencing detected the same substitution (Additional file 1: Fig. S16c). These results suggested that DAJIN enables the detection of mutations at single-nucleotide resolution in inversion alleles.

To confirm that DAJIN is also useful for flox analysis at other loci, we further generated and analysed floxed mice for two genes, *Exoc7* and *Usp46*. In the *Exoc7* project (Additional file 1: Fig. S17a), we obtained 40 founder mice and analysed them by using DAJIN. DAJIN clearly identified 11 mice with the 'Intended flox' allele; 7, with

the ‘Left LoxP’ allele; 14, with the ‘Right LoxP’ allele; and 5, with the ‘Deletion’ allele (Additional file 1: Fig. S17b). To verify the results of DAJIN, we performed PCR–RFLP and standard PCR to detect the LoxP band and deletion band, respectively. The PCR–RFLP results were in agreement with the DAJIN report except for BC13 (Additional file 1: Fig. S17c, d), which was shown to have no ‘Left LoxP’ allele by DAJIN, but PCR–RFLP detected this allele. Because 62.1 % of the reads in BC13 were annotated as ‘Deletion’ allele (Additional file 1: Fig. S17b), the deletion band might be predominantly amplified, and the number of ‘Left LoxP’ reads decreased owing to the PCR bias. When DAJIN enabled the detection of minor alleles with the ‘filter = off’ option, BC13 was shown to have 0.66 % reads of ‘Left LoxP’, which might be the PCR-detected allele (Additional file 2: Table. S2). PCR-RFLP for left LoxP detection uses a primer that anneals to the central arm region (Additional file 1: Fig. S17c); hence, PCR products derived from the deletion allele are not actually amplified. Thus, no PCR bias occurred in PCR–RFLP for left LoxP detection. The PCR results for the deletion band were in accordance with DAJIN’s report (Additional file 1: Fig. S17e, f). To confirm whether the allele determined by DAJIN was inherited through the next generation, we crossed WT with *Exoc7* BC14 that DAJIN determined to be heterozygous for ‘Intended flox’ allele (42.4 %) and ‘Right LoxP’ allele (45.5 %; Additional file 1: Fig. S17b). Of the total 11 F1 mice from this cross, 5 were flox/WT, and 6 were right LoxP/WT mice (Additional file 1: Fig. S18).

In the *Usp46* project (Additional file 1: Fig. S19a), we obtained 34 founder mice, and DAJIN reported 4 mice with the ‘Intended flox’ allele; 2, with the ‘Left LoxP’ allele; 2, with the ‘Right LoxP’ allele; and 15 with the ‘Deletion’ allele. Notably, DAJIN could identify both *trans* and *cis* knock-in in the BC04 mouse. DAJIN’s results were revalidated using PCR–RFLP that detected left and right LoxP alleles (Additional file 1:

Fig. S19c,d). Some results were inconsistent between PCR-RFLP and DAJIN. First, PCR-RFLP analysis of BC23 and BC33 showed that LoxPs were inserted on both sides, but DAJIN did not. Second, PCR-RFLP identified 'Left LoxP' alleles in BC09, BC13, and BC27, but not DAJIN. Since DAJIN reported that these five samples dominantly had the 'Deletion' alleles (Additional file 1: Fig. S19b), the reason for this mismatch could be PCR bias. By using DAJIN's 'filter-off' option described above, these alleles, except BC27 could be detected by DAJIN: in the DAJIN report generated using the 'filter-off' option, BC23 and BC33 had 0.39 % and 1.06 % of 'Intended flox' alleles, respectively, and BC09 and BC13 had 2.4 % of 'Intended flox' and 0.37 % of 'Left LoxP', respectively. However, even with the 'filter-off' option, DAJIN could not detect the 'Left LoxP' allele of BC27 (Additional file 2: Table. S2) because the PCR bands of BC27 indicated that the allele was too minor to be detected. In contrast, DAJIN detected the 'Left LoxP' allele in BC21, but PCR-RLFP did not. Thus, we re-examined the PCR results by adjusting the dilution ratio and detected the left LoxP band (Additional file 1: Fig. S19e). Nonetheless, the genotyping result of DAJIN and PCR-RFLP was consistent for other alleles such as 'Right LoxP' and 'Deletion' alleles (Additional file 1: Fig. S19d-g). Notably, the PCR band in six samples (BC07, BC12, BC17, BC23, BC30, and BC33) seemed to be a deletion band, whereas DAJIN reported them as 'abnormal' alleles. Visualisation of the reads revealed that the alleles contained an indel of about 30–200 bp larger/smaller than the 'Deletion' allele (Additional file 1: Fig. S20). This result indicated that allele annotation by DAJIN was accurate even when distinguishing allele types by PCR band size was difficult. Finally, we investigated whether the allele reported in DAJIN would be detected in the next generation. We obtained F1 progeny by crossing BC10 and BC11 with WT and found floxed and deletion alleles in the F1 mice (Additional file 1: Fig. S21). These results

provide definitive evidence that DAJIN can accurately and comprehensively detect diverse mutations that occur during the generation of floxed mice.

Discussion

Conventional approaches such as short-range PCR, Sanger sequencing, and PCR-RFLP are standard methods to detect on-target mutagenesis induced by CRISPR-Cas and other genome editing tools. Recent studies on on-target variability of edited materials clearly show that the characterisation of genome editing events and selection of animals or cultured cells with intended and unintended mutations require alternative methods with higher sensitivity and broader range to capture as many mutagenic events as possible [33, 34]. In this study, we developed a genotyping method using novel software DAJIN that can be applied for long-read sequencing to validate the quality of genome-edited organisms. Our method involving DAJIN can handle multiple samples obtained under different editing conditions and identify not only alleles with the desired mutation but also those with unexpected mutations, including large deletions, in a single run; this refined strategy can contribute to more precise genome editing.

We investigated whether DAJIN can accurately detect PMs, KO induced using the two-cut strategy, and flox mutations by comparing results with those obtained using conventional methods such as PCR, RFLP, Sanger sequencing, and TIDER. In most cases, the results obtained with DAJIN were consistent with those obtained using the conventional methods. With regard to PM verification, DAJIN reported the presence of intended PM alleles in 12/31 samples. TIDER analysis of the Sanger sequences showed that 2 (BC22 and BC26; Additional file 1: Fig. S7) of these 12 samples did not have an intended PM. TIDER also indicated the presence of a target PM in BC20 that

DAJIN determined to be without 'Intended PM'. The sensitivity of DAJIN in PM detection was 90.9 % (10/11), and the specificity was 90.0 % (18/20).

For the verification of the 2-cut KO design, DAJIN reported deletion alleles in a total of 26/75 samples in all 3 mouse strains investigated, but short-range PCR results showed deletion alleles in a total of 27/75 samples (sensitivity, 96.4%; specificity, 100%). Regarding flox design, DAJIN reported the detection of floxed alleles in a total of 24/94 samples in the 3 mouse strains, but PCR-RFLP results showed floxed alleles in a total of 27/94 samples (sensitivity, 88.9%; specificity, 100%). DAJIN has superior multi-specimen processing ability owing to its multiplexed PCR-based barcoding, which enables the pooling of multiple samples for sequencing and allows the coverage of numerous samples in a single run and obtaining sufficient coverage (Additional file 2: Table. S3). In this study, BC01–35 of *Usp46* (Additional file 1: Fig. S19) shared the same barcode as that of BC01–26 of *Prdm14* and BC27–35 of *Ddx4* (Fig. 3, Additional file 1: Figs. S12 and 13), and we could analyse 83 samples in a single run. Moreover, DAJIN supports parallel processing at every step of the analysis, suggesting its multi-core capability. We could analyse 226 samples (total 5,982,507 reads) by using DAJIN on a general-purpose desktop computer, and the entire process required about 15 h (Additional file 2: Table. S4). Thus, the DAJIN workflow is considerably shorter than that of conventional methods.

DNA double-strand break repair leads to long-range deletion, inversion, and insertion as well as small indels in zygotes and stem cells [8, 9, 35-37]. As long-read sequencing induces base-calling errors across a segment and cannot be used as is to validate the genome editing outcomes [15], novel screening techniques and tools need to be developed in order to detect diverse sequence changes in the genome. A parallel analysis of short- and long-read sequencing results confirmed that DAJIN could identify

editing outcomes, including unpredictable large-scale inversion events, in mouse zygotes (Additional file 1: Figs. S14 and 16). Short-range PCR amplification and Sanger sequencing confirmed that no additional mutation was detected in 'Intact' alleles identified by DAJIN, suggesting that DAJIN could be used to validate the consequences of genome editing at the base level (Figs. 2 and 4). DAJIN could detect additional sequence changes away from the target site with the intended mutation and identified the allele as 'Mutated' (Fig. 2). In addition, it reported separate mutations in *cis* generated by using two gRNAs positioned up to 2 kb apart on the same chromosome (Fig. 4, Additional file 1: Figs. S17 and S19). In previous knock-in experiments, exogenous repair templates and unwanted mismatches had been identified around the target region [23, 38-40]. Detection of mutations and/or integrations in *cis/trans* at a kilobase-scale distance requires a combination of assays. We propose that DAJIN is a novel validation tool for long-range sequencing at the single base level to capture genome editing events. Recently, DNA cleavage in cultured cells and zygotes has been shown to induce gene conversions mediated by homologous chromosomes or homologous sequences on the same chromosome [11, 12]. DAJIN might contribute to a better understanding of the consequences of editing events at the targeted locus.

Genotype assessment by using DAJIN might facilitate the selection of genome-edited samples with precisely targeted alleles or those with unwanted alleles. Comprehensive mutation analysis might reduce the overall cost of genome editing in not only laboratory mice but also other experimental animals or farm animals with a longer generation time. DAJIN can identify unintended alleles that meet the purposes of editing such as null allele production by using two gRNAs (Fig. 3). Phenotypic analysis confirmed that founders with 'Abnormal', but still presumed to be null alleles

(BC18 and BC23), were truly *Prdm14* null mice. Multiple alleles may be generated in the edited cell culture pools, but they cannot be segregated as in the case of founder animals. Our results indicate that genotyping with DAJIN can be also applied to the assessment of editing outcomes in cellular experiments (*e.g.* CRISPR screening) and cellular therapies. Thus, DAJIN offers a novel strategy to identify multiple genomic changes, including large sequence alterations or unexpected mutations, regardless of the species or types of the material.

The distinguishing feature of DAJIN is the utilization of a long-read sequencer as well as its ability to conduct automatic clustering and annotation of alleles. Genotyping tools similar to DAJIN are widely used, such as Cas-Analyzer [41], CRISPResso2 [42], and CLiCKER [43]; however, these tools are optimised for short-read sequencing. Because DAJIN uses a long-read sequencer, it can identify *cis*- or *trans*-heterozygosity and complex mutant alleles such as unexpected indels and structural variants. Secondly, in terms of estimating alleles, polyploid phasing, which allows the reconstruction of haplotypes of the polyploid genome, has a similar purpose as DAJIN. WHATSHAP POLYPHASE [44] and H-PoPG [45] are state-of-the-art tools for polyploid phasing, but these tools require prior knowledge of the polyploidy of the target organism. Because the number of alleles in a genome-edited organism cannot be predicted in advance, applying these for our research purposes is difficult. Recently, a new polyploid phasing technique called nPhase [46] has been reported. This method does not require the prior knowledge of multiplicity, but it does not annotate each allele type. Thus, in terms of using nanopore sequencing, automatic annotation of allele types, and identification of allele numbers, DAJIN is a genotyping tool with unique features.

DAJIN analysed PCR products in this study. Although PCR enables inexpensive and convenient barcoding and high-level enrichment of target genomic regions, it has two problems. The first is the PCR amplification bias, which lowers its efficiency to amplify long reads and GC-rich sequences, called length bias and GC bias, respectively. GC bias can be alleviated using high-grade DNA polymerase, but length bias cannot be removed and thus affects the accuracy of DAJIN's allele percentage. For example, in the case of *Usp46* flox knock-in, the percentage of the "Intended flox" allele was extremely low because the deletion allele might have been preferentially amplified (Additional file 1: Fig. S19). The second reason is that the genomic region that can be amplified using PCR was limited to about 10 kb, hindering the examination of a larger genome region, although nanopore sequencing can read a sequence more than 10 kb in length. Two recently developed methods have the potential to address these problems associated with PCR. First, IDMseq is used for labelling PCR amplicons by using unique molecular identifiers, which eliminates the length bias and allows more quantitative analysis of allele frequencies [47]. Second, nCATS enables the enrichment of the genome region without PCR, thereby allowing better quantitative analysis of longer sequences [48]. Importantly, DAJIN can be used to analyse the nanopore sequencing reads obtained using these techniques. Thus, combining these techniques with DAJIN can improve quantitative allele analysis and reveal considerably longer genomic regions.

Conclusion

In this study, DAJIN, an on-target site analysis software based on long-read sequencing, allowed the automatic detection and classification of various types of complex genome-edited mutations such as flox–*cis* double mutations, unexpected

deletions, and inversions at single-nucleotide resolution. Moreover, it allowed simultaneous processing of a large number of samples (approximately 100). With its high versatility, scalability, and convenience, DAJIN might significantly contribute to clone selection, which is the most important step in genome editing experiments.

Methods

Animals

ICR and C57BL/6J mice were purchased from Charles River Laboratories Japan, Inc. (Yokohama, Japan). C57BL/6J-*Tyr*^{em2Utr} mice were provided by RIKEN BRC (#RBRC06459). Mice were kept in plastic cages under specific pathogen-free conditions in a room maintained at 23.5°C ± 2.5°C and 52.5 ± 12.5 % relative humidity under a 14-h light:10-h dark cycle. Mice had free access to commercial chow (MF diet; Oriental Yeast Co., Ltd., Tokyo, Japan) and filtered water. All animal experiments were performed humanely with the approval from the Institutional Animal Experiment Committee of the University of Tsukuba following the Regulations for Animal Experiments of the University of Tsukuba and Fundamental Guidelines for Proper Conduct of Animal Experiments and Related Activities in Academic Research Institutions under the jurisdiction of the Ministry of Education, Culture, Sports, Science, and Technology of Japan.

Genome editing in mouse zygotes

Mice with PMs and two-cut KO were generated using the electroporation method [49]. The gRNA target sequences to induce each mutation are listed in Additional file 2: Table. S5. The gRNAs were synthesised and purified using a GeneArt™ Precision gRNA Synthesis Kit (Thermo Fisher Scientific, MA, USA) and dissolved in Opti-MEM

(Thermo Fisher Scientific, MA, USA). In addition, we designed three ssODN donors for inducing PMs in *Tyr* (Additional file 2: Table. S5). These ssODN donors were ordered as Ultramer DNA oligos from Integrated DNA Technologies (IA, USA) and dissolved in Opti-MEM. The mixtures of gRNA (5 ng/μL) and ssODNs (100 ng/μL) or mixtures of two gRNAs (25 ng/μL, each) were used to generate point mutant mice or two-cut KO mice, respectively. GeneArt Platinum Cas9 Nuclease (100 ng/μL; Thermo Fisher Scientific, Waltham, Massachusetts) was added to these mixtures. Pregnant mare serum gonadotropin (5 units) and human chorionic gonadotropin (5 units) were intraperitoneally injected into female C57BL/6J mice (Charles River Laboratories) with a 48 h interval. Next, unfertilised oocytes were collected from their oviducts. Next, according to standard protocols, we performed *in vitro* fertilisation with these oocytes and sperm from male C57BL/6J mice (Charles River Laboratories). After 5 h, the above-mentioned gRNA/ssODN/Cas9 or two gRNAs/Cas9 mixtures were electroplated to the mouse zygotes by using an NEPA 21 electroplater (NEPAGNENE; Chiba, Japan), under the same condition that we reported previously [50]. The electroporated embryos that developed into the two-cell stage were transferred to oviducts of pseudopregnant ICR female mice. The floxed mice were generated using the microinjection method [28]. Each gRNA target sequence (Additional file 2: Table. S5) was inserted into the entry site of pX330-mC carrying both the gRNA and Cas9 expression units. These pX330-mC plasmid DNAs and donor DNA plasmid were isolated using FastGene Plasmid Mini kit (Nippon Genetics, Tokyo, Japan) and filtered using MILLEX-GV® 0.22 μm filter unit (Merck Millipore, Darmstadt, Germany) for microinjection. Next, C57BL/6J female mice superovulated using the method described above were naturally mated with male C57BL/6J mice, and zygotes were collected from the oviducts of the mated female mice. For each gene, a mixture of two pX330-

mC (circular, 5 ng/μL each) and a donor (circular, 10 ng/μL) was microinjected into the zygote. The survived zygotes were then transferred into the oviducts of pseudopregnant ICR female mice. When the newborns were around 3 weeks of age (Additional file 2: Table. S6), the tail was sampled to obtain genomic DNA.

Library preparation and nanopore sequencing

We used PI-200 (KURABO INDUSTRIES LTD., Osaka, Japan), according to manufacturer's protocol, for the extraction and purification of genomic DNA obtained from the tail of mice. The purified genomic DNA was amplified using PCR by using KOD multiEpi (TOYOBO, Osaka, Japan) and target amplicon primers (Additional file 2: Table. S7). In the target amplicon primer, the universal sequence is located on the 5′ side, and the sequence for target gene amplification is on the 3′ side. Five-fold dilutions of the PCR products were used as templates for nested PCR performed using KOD multiEpi and barcode attachment primers (Additional file 2: Table. S8). The 5′ side of the barcode attachment primer has a barcode sequence, and the 3′ sequence is annealed to the universal sequence of the target amplicon primer. The barcoded PCR products were mixed in equal amounts and then purified using FastGene Gel/PCR Extraction Kit (Nippon Genetics, Germany). The volume of the mixed and purified PCR products adjusted to 20–30 ng/μL. The library was prepared using Ligation Sequencing 1D kit (SQK-LSK108_109; ONT, Oxford, UK) and NEBNext End repair/dA-tailing Module NEB Blunt/TA Ligase Master Mix (New England Biolabs) according to manufacturers' instructions. The prepared library was loaded onto a primed R9.4 Spot-On Flow cell (FLO-MIN106; ONT, Oxford, UK). The 24 h or 36 h run time calling sequencing protocol was selected in the MinKNOW GUI (version 4.0.20), and base calling was allowed to complete after the sequencing run was completed. After base calling, we demultiplexed the barcoding libraries by using qcat (version 1.1.0) with

default parameter settings. Total nanopore sequencing reads per sample are listed at Additional file 2: Table. S3.

Conventional genotyping analysis

To evaluate the validity of DAJIN's genotyping results, we used conventional genotyping methods, including short-amplicon PCR, PCR-RFLP, and Sanger sequencing. For TIDER analysis, we used TIDER version 1.0.2. Sanger sequence data from WT mice were used as the Control Chromatogram for all three strains. The Reference Chromatogram for *Tyr* c.140G>C and c.308G>C was composed of BC12 and BC21 data, which were confirmed to have only the target mutation, respectively, by Sanger sequence waveforms. For *Tyr* c.316G>C, the PCR product of BC18 was subcloned to obtain a plasmid vector containing the DNA sequence with the target mutation only. The PCR products obtained using this vector as a template were used as a Reference Chromatogram. For the genotyping of the two-cut KO and PM lines, genomic PCR was performed using AmpliTaq Gold 360 DNA Polymerase (ThermoFisher, Waltham, MA, USA) and the relevant primers (Additional file 2: Table. S9). Agarose gel electrophoresis was performed to confirm the size of the PCR products. In the flox knock-in design, genomic PCR was performed using KOD FX (TOYOBO) and the relevant primers (Additional file 2: Table. S9). The PCR products were digested with restriction enzymes *Asc*I (New England Biolabs) and *Eco*RV (New England Biolabs) for 2 h to check LoxP insertion on the left and right side, respectively. Agarose gel electrophoresis was performed to confirm the size of the PCR fragments. PCR products with mutant sequences were identified using Sanger sequencing by using the BigDye™ Terminator v3.1 Cycle Sequencing Kit (ThermoFisher).

Automatic assessment of genome editing design

DAJIN automatically assessed the genome editing design among PMs, knock-outs, and knock-ins based on the user-inputted FASTA file. The sequence of the 'Target' allele was aligned to the WT sequence by using minimap2 [51] with the '-ax splice' and '--cs' options. If the CS tag contained '*', '~', and '+', it was considered a PM, knock-out, and knock-in, respectively.

Nanopore-simulated reads

To prepare training data for DNN models, we generated simulation reads of the possible alleles by using NanoSim (version 2.5.0) [24]. First, we trained NanoSim on the error profile by using nanopore sequencing reads from WT controls. Next, we applied the error profiles to the sequences of each possible allele that could have been caused by genome editing and generated 10,000 simulation reads per allele (Additional file 1: Fig. S1). In the PM design, we generated simulation reads with a deletion or random nucleotide insertion of the gRNA length at the Cas-cutting site.

Pre-processing

We performed pre-processing on simulation and nanopore sequencing reads excluding those that did not contain mutant regions and MIDS conversion to improve the performance of anomaly detection (Additional file 1: Fig. S6). First, the genome-edited sequence was aligned to the user-provided WT sequence by using minimap2 with '--cs=long' option, and the position of the target mutant base was detected according to the CS-tag in the SAM file. Simulated and nanopore sequencing reads were then aligned using minimap2 to the WT sequence. Reads with lengths more than 1.1 times longer than the maximum length among possible alleles were excluded. For the remaining reads, we detected the start and end positions of each read relative to the

WT sequence based on CIGAR information and extracted the reads containing the mutant region of interest (Additional file 1: Fig. S2a).

Next, the extracted reads were subjected to MIDS conversion, during which the matched, inserted, deleted, and substituted bases between reference and query alignments were converted to M (Match), I (Insertion), D (Deletion), and S (Substitute), respectively. Next, the read lengths were trimmed or padded with '=' to equalise their length with the maximum length in the possible alleles. In the case of large deletions (two-cut deletion or flox deletion), the reads were aligned to the reference as two primary sequences, and the region of deleted bases between the two primary sequences was padded with 'D'. Moreover, in the case of inversion, a read was divided into two primary sequences and one secondary sequence, and the entire length of the secondary sequence was replaced with '='. finally, one-hot encoding was performed on the MIDS sequence (Additional file 1: Fig. S2b).

Deep learning model

To detect abnormal alleles and classify possible alleles for nanopore sequencing reads, we constructed a deep learning model. A three-layer convolutional neural network (CNN) was used to extract characteristic variations in the sequenced reads. Max pooling was used to increase the generalisation capability, and ReLU was used as the activation function. After the CNN and pooling layers were applied, the FC and Softmax layers were used to estimate the probability of the allele type. The details of the deep learning model structure are shown in (Additional file 1: Fig. S3a,b). At the training phase, the batch size was 32, the maximum number of epochs was 200, and training was stopped at the point when validation loss began to plateau during 20 epochs.

We trained the model to classify possible alleles by using NanoSim simulation reads. To detect abnormal reads, we inputted the simulated and nanopore sequencing reads

into the trained model and extracted the output vectors from the FC layer. Next, we trained the LOF [26] by using the FC layer output of the simulated reads to learn the density of vectors from possible alleles. Subsequently, the output vectors of the FC layer of the nanopore sequence reads were placed in the LOF. The LOF detected and labelled unexpected mutation reads as 'Abnormal'. After abnormal reads were detected, the nanopore sequencing reads other than abnormal reads were placed in the deep learning model, and the allele type was estimated for each nanopore sequencing read. Moreover, in the case of the PM design, the nanopore reads were annotated in four types: 'WT', 'Deletion', 'Insertion', and 'Abnormal'; the reads classified as 'Deletion' and 'Insertion' were then relabelled as 'Abnormal'.

Allele clustering

Because the deep learning model in DAJIN aimed at detecting abnormal reads with mutations of more than a few dozen bases, different alleles at the single-nucleotide resolution within the same allele type might have been omitted. In order to distinguish subgroups in each allele type precisely, DAJIN annotated the mutant bases by a compressed MIDS conversion (Additional file 1: Fig. S4) and conducted clustering. Since the MIDS-converted reads might be longer than the reference sequence owing to the insertion of bases, the base insertion information needed to be compressed to align accurately the coordinates of a mutant base. Therefore, we generated compressed MIDS conversion, which replaces successive insertions with a character corresponding to the number of insertions and then substitutes the insertion (Additional file 1: Fig. S4). A character is assigned to the number from 1 to 9 or a letter from a to z. If the number of consecutive insertions is in the range 1–9, the character is the corresponding number. If the number of consecutive insertions is in the range 10–35, the character is 'a' (=10) to 'y' (=35). If the number is greater than 35, the character is 'z' (>35). Here,

the reference sequence means the sequence of the DAJIN-predicted allele type. If this allele type was 'Abnormal', the WT sequence was used as the reference sequence.

To mitigate nanopore sequencing errors, DAJIN subtracted the relative frequencies of MIDS between WT controls and samples. First, the WT control was aligned to the WT target sequence by using minimap2, and compressed MIDS conversion was then performed. The relative frequency was calculated for the MIDS at each base position. In the case of PMs, the mutation position was set to a 100% match regardless of the sequence error in the control. In the case of insertions, the bases at the insertion were considered as a 100% match. Conversely, in the case of deletions, the corresponding nucleotide positions were deleted. In the case of inversion, the corresponding nucleotide positions were inverted. Subsequently, the reads of the predicted allele type were aligned with the sequence of their DAJIN-predicted allele type, and the relative frequency of the MIDS was calculated after compressed MIDS conversion. The relative number of MIDS frequencies was subtracted from the control MIDS score and applied to each read (Additional file 1: Fig. S5).

The MIDS score of each read was reduced into 10 dimensions by using principal component analysis, and the matrix of the number of reads × 10 was obtained by multiplying the principal component score and the contribution of a variable. Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN) [52] was used for the clustering. For parameters, we set 'min_samples', which specifies the minimum size of each cluster formed, as '1' to reduce noise points. Furthermore, we tuned 'min_cluster_size', which defines the minimum number of samples in each cluster. We set the value as 50 equal intervals between 1/5 and 2/5 of the total number of reads and then selected the 'min_cluster_size', which outputs the mode of cluster numbers.

After clustering, we performed anomaly detection by using Hotelling's T2 statistics to determine the mutations that might contribute to the clustering. Because Hotelling's T2 statistics assume a normal distribution, we performed z-score normalisation following the summation of the MIDS scores at each cluster's base position. Next, since Hotelling's T2 statistics detect anomalies by assuming few or no anomalies, data points were augmented by a factor of 100 by random sampling from a normal distribution with the mean and standard deviation calculated using normalised MIDS scores of the first 100 base positions. We then calculated Hotelling's T2 score from the sequence score as follows.

$$t^2 = \frac{(\bar{x} - \mu_0)^2}{s^2/n}$$

The 99th percentile of the Chi-squared distribution with 1 degree of freedom was used as the anomaly score threshold. The base positions with a value higher than this threshold were considered as possible mutations. We calculated the sum of MIDS scores of the abnormal base sites for each cluster to merge similar clusters, and, for merging, similar clusters were judged to have a cosine similarity score of 0.95 or more.

Filtering minor alleles

To improve the interpretability, DAJIN has a default setup to remove minor alleles. Although classification and clustering yielded predicted alleles, allowing minor alleles may result in numerous alleles in a sample, which hinders intuitive understanding. Therefore, DAJIN excluded minor alleles from the analysis by default setting. Minor alleles were defined as those in which the number of reads was 3 % or less of the total number of reads of a sample. DAJIN could report all allele information by using the 'filter=off' option.

Consensus sequence

The consensus sequence for each allele was output as a FASTA file and an HTML file with mutations appearing in coloured format. In order to identify the mutant positions, we extracted the most frequent MIDS letters at each nucleotide position from the compressed MIDS conversion. If the most frequent MIDS letters were not ‘M’, the position was identified as a mutation, and the mutation location and type (insertion, deletion, or substitution) were determined, and this information was used to identify the nucleotide at the mutation position by searching aligned reads. Mutations at the sequencing error positions were not counted as mutations, and the MIDS letters were replaced by ‘M’. In the case of ‘WT’ and ‘Target’ alleles, when the DAJIN consensus sequence perfectly matched the desired WT and Target sequences, they were labelled as ‘Intact WT’ and ‘Intended Target’, respectively. Conversely, when the DAJIN consensus sequence detected mutations that were not described in a user-inputted FASTA file, the alleles were annotated as ‘Mutated WT’ and ‘Mutated Target’, respectively.

Generation and visualisation of BAM files

DAJIN generates BAM files to visualise the DAJIN-reported alleles in a genome browser. First, DAJIN uses minimap2 to map the nanopore sequence reads to the WT sequences described in the user-inputted FASTA file and generates BAM files. Next, based on the WT sequence and reference genome information such as hg38 and mm10, the BED file, including target genome coordinates, was obtained from the UCSC Table Browser [53]. Based on the chromosome numbers in the BED file, we obtained the chromosome length from the UCSC Table Browser and converted the chromosome number and chromosome length to SN and LN headers of BAM files, respectively. We replaced the start site of each read in the BAM files with that

described in the BED file. Next, the BAM files were sorted and indexed using samtools (version 1.10) [54] and visualised using IGV [55].

Figure legends

Fig. 1 DAJIN overview

a The schema of DAJIN's workflow. DAJIN automates the procedures highlighted in grey. **b** The outputs of DAJIN. The file formats are described in parentheses.

Fig. 2 DAJIN application for point mutation design

a Genome editing design for *Tyr* c.140G>C point mutation. The scissor represents a Cas9-cutting site. The arrows represent PCR primers, including the PCR amplicon size. The boxed allele type represents the target allele, and the boxed nucleotide represents a targeted point mutation. **b** DAJIN's report of the allele percentage. The barcode numbers on the x-axis represent mouse IDs. BC32 is a WT control. The y-axis represents the percentage of DAJIN-reported alleles. The colours of the bars represent DAJIN-reported allele types. The compartments partitioned off by horizontal lines in a bar represent the DAJIN-reported alleles. **c** Visualisation of nanopore sequencing reads at *Tyr* target locus. BC12 and BC08 contain target alleles. BC32 is a WT control. The 'All alleles' track represents all reads of each sample. The 'Allele' track represents DAJIN-reported alleles. **d** Comparison between DAJIN consensus sequence and Sanger sequencing. The sequence represents the consensus sequence of a dominant allele of BC12 and BC08. The colours on the nucleotides represent mutation types, including insertion (red), deletion (sky blue), and substitution (green). The coloured boxes in the Sanger sequence represent mutated nucleotides, including insertion (red) and substitution (green).

Fig. 3 DAJIN application for knock-out design

a Genome editing design for *Prdm14* knock-out. The scissors and dotted lines represent Cas9-cutting sites. The arrows represent PCR primers, including the size of the PCR amplicon. The boxed allele type represents the target allele. The inversion allele represents a possible byproduct. The triangle on the nucleotides represents a junction site of two DNA fragments. **b** DAJIN's report of the allele percentage. The barcode numbers on the x-axis represent mouse IDs. BC26 is a WT control. The y-axis represents the percentage of DAJIN-reported alleles. The colours of the bar represent DAJIN-reported allele types. The compartments partitioned off by horizontal lines in a bar represent the DAJIN-reported alleles. **c** Visualisation of nanopore sequence reads at *Prdm14* target locus. BC18 and BC23 contain target alleles. BC26 is a WT control. The 'All alleles' track represents all reads of each sample. The 'Allele' track represents DAJIN-reported alleles. **d** DAJIN consensus sequences of the target allele. The top sequence represents the consensus sequence of target alleles of BC18 Allele 3 and BC23 Allele 3. The bottom sequences enlarge the boxed sequence of the consensus sequence. The colours on the nucleotides represent mutation types, including insertion (red), deletion (sky blue), and substitution (green). **e** Validation by Sanger sequencing. The dotted lines represent corresponding nucleotides between Sanger and DAJIN consensus sequences. **f** PLZF (red), Vimentin (green), and Hoechst (blue) staining of the testis section of WT (left), BC18 (middle), and BC23 (right). Upper panels show co-staining of PLZF (red) and Hoechst (blue). Lower panels show co-staining of Vimentin (green) and Hoechst (blue). PLZF and Vimentin are markers of undifferentiated spermatogonia and Sertori cells, respectively, along the seminiferous tubules' basal lamina. Scale bar: 100 μm

Fig. 4 Precise detection of single nucleotide variation in flox knock-in allele by using DAJIN

a Genome editing design. The red arrowheads represent LoxPs. The colours on the nucleotides represent the types of mutations, including insertion (red), deletion (sky blue), and substitution (green). 'Sub' means substitution. **b** Experimental design. **c** DAJIN's report of the allele percentage. The barcode numbers on the x-axis represent sample IDs. Barcode42 is a WT control. The y-axis represents the percentage of DAJIN-reported alleles. The colours of the bar represent DAJIN-reported allele types. The compartments partitioned off by horizontal lines in a bar represent the DAJIN-reported alleles. 'Sub' means substitution. **d** The DAJIN consensus sequences of a floxed allele in each sample. The colours on the nucleotides represent mutation types, including insertion (red), deletion (sky blue), and substitution (green). The boxed sequences in the consensus sequences are LoxP sites.

Fig. 5 DAJIN application for flox knock-in design

a Genome editing design for flox knock-in into the *Cables2* locus. The scissors represent Cas9-cutting sites. The arrows represent PCR primers, including the size of the PCR amplicon. The circular DNA represents the donor DNA. The base numbers on the donor DNA describe the left, central, and right arm sizes. The red arrowheads represent LoxPs. The boxed allele type represents the target allele. The other allele types include Left LoxP and Right LoxP. Inversion and Deletion represent possible byproducts. **b** DAJIN's report of the allele percentage. The barcode numbers on the x-axis represent mouse IDs. BC42 is a WT control. The y-axis represents the percentage of DAJIN-reported alleles. The colours of the bar represent DAJIN-reported allele types. The compartments partitioned off by horizontal lines in a bar represent the DAJIN-reported alleles. **c** Design of PCR-RFLP to validate LoxP knock-in alleles. The *Asc*I

and *Eco*RV restriction sites are adjacent to Left LoxP and Right LoxP, respectively. The arrows represent PCR primers for the digested DNA fragments, including PCR product sizes. **d** PCR results for the detection of the LoxP knock-in allele. The top and bottom panels represent DNA fragments digested with *Asc*I and *Eco*RV, respectively. The numbers on the panel mean barcode IDs. The boxed number represents the samples with LoxP alleles. **e** Design of PCR to validate deletion alleles. The arrows represent PCR primers. **f** PCR results for the detection of deletion alleles. The panels show the PCR products. The number on the panel means barcode IDs. The boxed number represents the samples with deletion alleles. **g** DAJIN consensus sequence of the floxed allele in BC14. The red and blue boxes represent left and right LoxP sites, respectively. **h** Sanger sequences for the LoxP sites.

Declarations

Author Contributions

S. A., A. Y., S. T., F. S., and S. M. designed the study with inputs from all other authors. A. K., K. S., and S. S. conducted the bioinformatics analyses. Y. I., K. K., Y. H., and T. M., conducted molecular experiments. Ke. M., A. W., N. M., T. T. H. D., Ka. M., and M. H. conducted lineage and phenotypic analysis experiments on mice. S. A., M. I., N. I., Y. D., and Y. T. conducted experiments to produce mice. A. K., Y. I., S. A., K. S., and S. M. wrote the manuscript. The authors read and approved the final manuscript.

Consent for publication

N/A

Availability of data and materials

DAJIN is accessible at https://github.com/akikuno/DAJIN under the MIT License. The version of DAJIN used in this study to reproduce the analyses can be found at https://github.com/akikuno/DAJIN/tree/manuscript-version. Raw sequencing data are available in the SRA database (accession ID, XXXXX).

Acknowledgements

We would like to thank the staff at the Laboratory Animal Resource Center University of Tsukuba for their help in generating and caring for the mice. We are grateful to Tomoyuki Fujiyama for advice on the experimental design. We also acknowledge Ozaki Haruka for the fruitful discussion.

Funding

This work was supported by Scientific Research (B) (19H03142: to S.M. and A.K.) from the Ministry of Education, Culture, Sports, Science, and Technology 542 (MEXT) and DRUG DISCOVERY & DEVELOPMENT Programs (to S.M. and S.T.) from the Japan Agency for Medical Research and Development (AMED). The funders had no role in the study design, data collection, and analysis; decision to publish; or preparation of the manuscript.

Ethics declarations

Ethics approval and consent to participate

N/A

Competing interests

The authors declare that they have no competing interests.

References

1. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, et al. Targeting DNA double-strand breaks with TAL effector nucleases. Genetics. 2010;186(2):757-61.

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. 2012;337(6096):816-21.

3. Komor AC, Kim YB, Packer MS, Zuris JA, Liu DR. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. Nature. 2016;533(7603):420-4.

4. Yeh CD, Richardson CD, Corn JE. Advances in genome editing through control of DNA repair pathways. Nat Cell Biol. 2019;21(12):1468-78.

5. Hanna RE, Doench JG. Design and analysis of CRISPR-Cas experiments. Nat Biotechnol. 2020;38(7):813-23.

6. Hendel A, Kildebeck EJ, Fine EJ, Clark J, Punjya N, Sebastiano V, et al. Quantifying genome-editing outcomes at endogenous loci with SMRT sequencing. Cell Rep. 2014;7(1):293-305.

7. Kraft K, Geuer S, Will AJ, Chan WL, Paliou C, Borschiwer M, et al. Deletions, Inversions, Duplications: Engineering of Structural Variants using CRISPR/Cas in Mice. Cell Rep. 2015;10(5):833-9.

8. Boroviak K, Fu B, Yang F, Doe B, Bradley A. Revealing hidden complexities of genomic rearrangements generated with Cas9. Sci Rep. 2017;7(1):12867.

9. Kosicki M, Tomberg K, Bradley A. Repair of double-strand breaks induced by CRISPR-Cas9 leads to large deletions and complex rearrangements. Nat Biotechnol. 2018;36(8):765-71.

10. Simeonov DR, Brandt AJ, Chan AY, Cortez JT, Li Z, Woo JM, et al. A large CRISPR-induced bystander mutation causes immune dysregulation. Commun Biol. 2019;2:70.

11. Ma H, Marti-Gutierrez N, Park SW, Wu J, Lee Y, Suzuki K, et al. Correction of a pathogenic gene mutation in human embryos. Nature. 2017;548(7668):413-9.

12. Javidi-Parsijani P, Lyu P, Makani V, Sarhan WM, Yoo KW, El-Korashi L, et al. CRISPR/Cas9 increases mitotic gene conversion in human cells. Gene Ther. 2020;27(6):281-96.

13. Liang D, Marti NG, Chen T, Lee Y, Park SW, Ma H, et al. FREQUENT GENE CONVERSION IN HUMAN EMBRYOS INDUCED BY DOUBLE STRAND BREAKS. doi: https://doi.org/10.1101/2020.06.19.162214 (2020).

14. Canaj H, Hussmann AJ, Li H, Beckman AK, Goodrich L , Cho HN, et al. Deep profiling reveals substantial heterogeneity of integration outcomes in CRISPR knock-in experiments. doi: https://doi.org/10.1101/841098 (2019).

15. McCabe VC, Codner FG, Allan JA, Caulder A, Christou S, Loeffler J, et al. Application of long-read sequencing for robust identification of correct alleles in genome edited animals. doi: https://doi.org/10.1101/838193 (2019).

16. Mizuno S, Dinh TT, Kato K, Mizuno-Iijima S, Tanimoto Y, Daitoku Y, et al. Simple generation of albino C57BL/6J mice with G291T mutation in the tyrosinase gene by the CRISPR/Cas9 system. Mamm Genome. 2014;25(7-8):327-34.

17. Yin H, Xue W, Chen S, Bogorad RL, Benedetti E, Grompe M, et al. Genome editing with Cas9 in adult mice corrects a disease mutation and phenotype. Nat Biotechnol. 2014;32(6):551-3.

18. Smith C, Gore A, Yan W, Abalde-Atristain L, Li Z, He C, et al. Whole-genome sequencing analysis reveals high specificity of CRISPR/Cas9 and TALEN-based genome editing in human iPSCs. Cell Stem Cell. 2014;15(1):12-3.

19. Teboul L, Herault Y, Wells S, Qasim W, Pavlovic G. Variability in Genome Editing Outcomes: Challenges for Research Reproducibility and Clinical Safety. Mol Ther. 2020;28(6):1422-31.

20. Canver MC, Bauer DE, Dass A, Yien YY, Chung J, Masuda T, et al. Characterization of genomic deletion efficiency mediated by clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 nuclease system in mammalian cells. J Biol Chem. 2014;289(31):21312-24.

21. Brinkman EK, Chen T, Amendola M, van Steensel B. Easy quantitative assessment of genome editing by sequence trace decomposition. Nucleic Acids Res. 2014;42(22):e168.

22. Birling MC, Schaeffer L, Andre P, Lindner L, Marechal D, Ayadi A, et al. Efficient and rapid generation of large genomic variants in rats and mice using CRISMERE. Sci Rep. 2017;7:43331.

23. Lanza DG, Gaspero A, Lorenzo I, Liao L, Zheng P, Wang Y, et al. Comparative analysis of single-stranded DNA donors to generate conditional null mouse alleles. BMC Biol. 2018;16(1):69.

24. Yang C, Chu J, Warren RL, Birol I. NanoSim: nanopore sequence read simulator based on statistical characterization. Gigascience. 2017;6(4):1-6.

25. McInnes L, Healy J, Melville J. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv:1802.03426v3 (2020).

26. Breunig MM, Kriegel HP, Ng TR, Sander J. LOF: Identifying Density-Based Local Outliers. https://doi.org/10.1145/335191.335388 (2000).

27. Brinkman EK, Kousholt AN, Harmsen T, Leemans C, Chen T, Jonkers J, et al. Easy quantification of template-directed CRISPR/Cas9 editing. Nucleic Acids Res. 2018;46(10):e58.

28. Mizuno-Iijima S, Ayabe S, Kato K, Matoba S, Ikeda Y, Dinh TTH, et al. Efficient production of large deletion and gene fragment knock-in mice mediated by genome editing with Cas9-mouse Cdt1 in mouse zygotes. Methods. 2020.

29. Yamaji M, Seki Y, Kurimoto K, Yabuta Y, Yuasa M, Shigeta M, et al. Critical function of Prdm14 for the establishment of the germ cell lineage in mice. Nat Genet. 2008;40(8):1016-22.

30. Zetsche B, Gootenberg JS, Abudayyeh OO, Slaymaker IM, Makarova KS, Essletzbichler P, et al. Cpf1 is a single RNA-guided endonuclease of a class 2 CRISPR-Cas system. Cell. 2015;163(3):759-71.

31. Osawa Y, Usui M, Kuba Y, Le TH, Mikami N, Nakagawa T, et al. EXOC1 regulates cell morphology of spermatogonia and spermatocytes in mice. doi: https://doi.org/10.1101/2020.06.07.139030 (2020).

32. Gurumurthy CB, O'Brien AR, Quadros RM, Adams J, Jr., Alcaide P, Ayabe S, et al. Reproducibility of CRISPR-Cas9 methods for generation of conditional mouse alleles: a multi-center evaluation. Genome Biol. 2019;20(1):171.

33. Mianne J, Codner GF, Caulder A, Fell R, Hutchison M, King R, et al. Analysing the outcome of CRISPR-aided genome editing in embryos: Screening, genotyping and quality control. Methods. 2017;121-122:68-76.

34. Burgio G, Teboul L. Anticipating and Identifying Collateral Damage in Genome Editing. Trends Genet. 2020;36(12):905-14.

35. Xiao A, Wang Z, Hu Y, Wu Y, Luo Z, Yang Z, et al. Chromosomal deletions and inversions mediated by TALENs and CRISPR/Cas in zebrafish. Nucleic Acids Res. 2013;41(14):e141.

36. Shin HY, Wang C, Lee HK, Yoo KH, Zeng X, Kuhns T, et al. CRISPR/Cas9 targeting events cause complex deletions and insertions at 17 sites in the mouse genome. Nat Commun. 2017;8:15464.

37. Adikusuma F, Piltz S, Corbett MA, Turvey M, McColl SR, Helbig KJ, et al. Large deletions induced by Cas9 cleavage. Nature. 2018;560(7717):E8-E9.

38. Mianne J, Chessum L, Kumar S, Aguilar C, Codner G, Hutchison M, et al. Correction of the auditory phenotype in C57BL/6N mice via CRISPR/Cas9-mediated homology directed repair. Genome Med. 2016;8(1):16.

39. Codner GF, Mianne J, Caulder A, Loeffler J, Fell R, King R, et al. Application of long single-stranded DNA donors in genome editing: generation and validation of mouse mutants. BMC Biol. 2018;16(1):70.

40. Skryabin BV, Kummerfeld DM, Gubar L, Seeger B, Kaiser H, Stegemann A, et al. Pervasive head-to-tail insertions of DNA templates mask desired CRISPR-Cas9-mediated genome editing events. Sci Adv. 2020;6(7):eaax2941.

41. Park J, Lim K, Kim JS, Bae S. Cas-analyzer: an online tool for assessing genome editing results using NGS data. Bioinformatics. 2017;33(2):286-8.

42. Clement K, Rees H, Canver MC, Gehrke JM, Farouni R, Hsu JY, et al. CRISPResso2 provides accurate and rapid genome editing sequence analysis. Nat Biotechnol. 2019;37(3):224-6.

43. Iida M, Suzuki M, Sakane Y, Nishide H, Uchiyama I, Yamamoto T, et al. A simple and practical workflow for genotyping of CRISPR-Cas9-based knockout phenotypes using multiplexed amplicon sequencing. Genes Cells. 2020;25(7):498-509.

44. Schrinner SD, Mari RS, Ebler J, Rautiainen M, Seillier L, Reimer JJ, et al. Haplotype threading: accurate polyploid phasing from long reads. Genome Biol. 2020;21(1):252.

45. Xie M, Wu Q, Wang J, Jiang T. H-PoP and H-PoPG: heuristic partitioning algorithms for single individual haplotyping of polyploids. Bioinformatics. 2016;32(24):3735-44.

46. Saada AO, Tsouris A, Friedrich A, Schacherer J. nPhase: An accurate and contiguous phasing method for polyploids. doi: https://doi.org/10.1101/2020.07.24.219105 (2020).

47. Bi C, Wang L, Yuan B, Zhou X, Li Y, Wang S, et al. Long-read individual-molecule sequencing reveals CRISPR-induced genetic heterogeneity in human ESCs. Genome Biol. 2020;21(1):213.

48. Gilpatrick T, Lee I, Graham JE, Raimondeau E, Bowen R, Heron A, et al. Targeted nanopore sequencing with Cas9-guided adapter ligation. Nat Biotechnol. 2020;38(4):433-8.

49. Kaneko T, Mashimo T. Simple Genome Editing of Rodent Intact Embryos by Electroporation. PLoS One. 2015;10(11):e0142755.

50. Sato Y, Tsukaguchi H, Morita H, Higasa K, Tran MTN, Hamada M, et al. A mutation in transcription factor MAFB causes Focal Segmental Glomerulosclerosis with Duane Retraction Syndrome. Kidney Int. 2018;94(2):396-407.

51. Li H. Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics. 2018;34(18):3094-100.

52. McInnes L, Healy J, Astels S. hdbscan: Hierarchical density based clustering. doi: 10.21105/joss.00205 1 (2017).

53. Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, et al. The UCSC Table Browser data retrieval tool. Nucleic Acids Res. 2004;32(Database issue):D493-6.

54. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. Bioinformatics. 2009;25(16):2078-9.

55. Robinson JT, Thorvaldsdottir H, Winckler W, Guttman M, Lander ES, Getz G, et al. Integrative genomics viewer. Nat Biotechnol. 2011;29(1):24-6.

Supplementary Information

Additional file 1.

Supplementary Figures S1–S21 with their legends (MS Word).

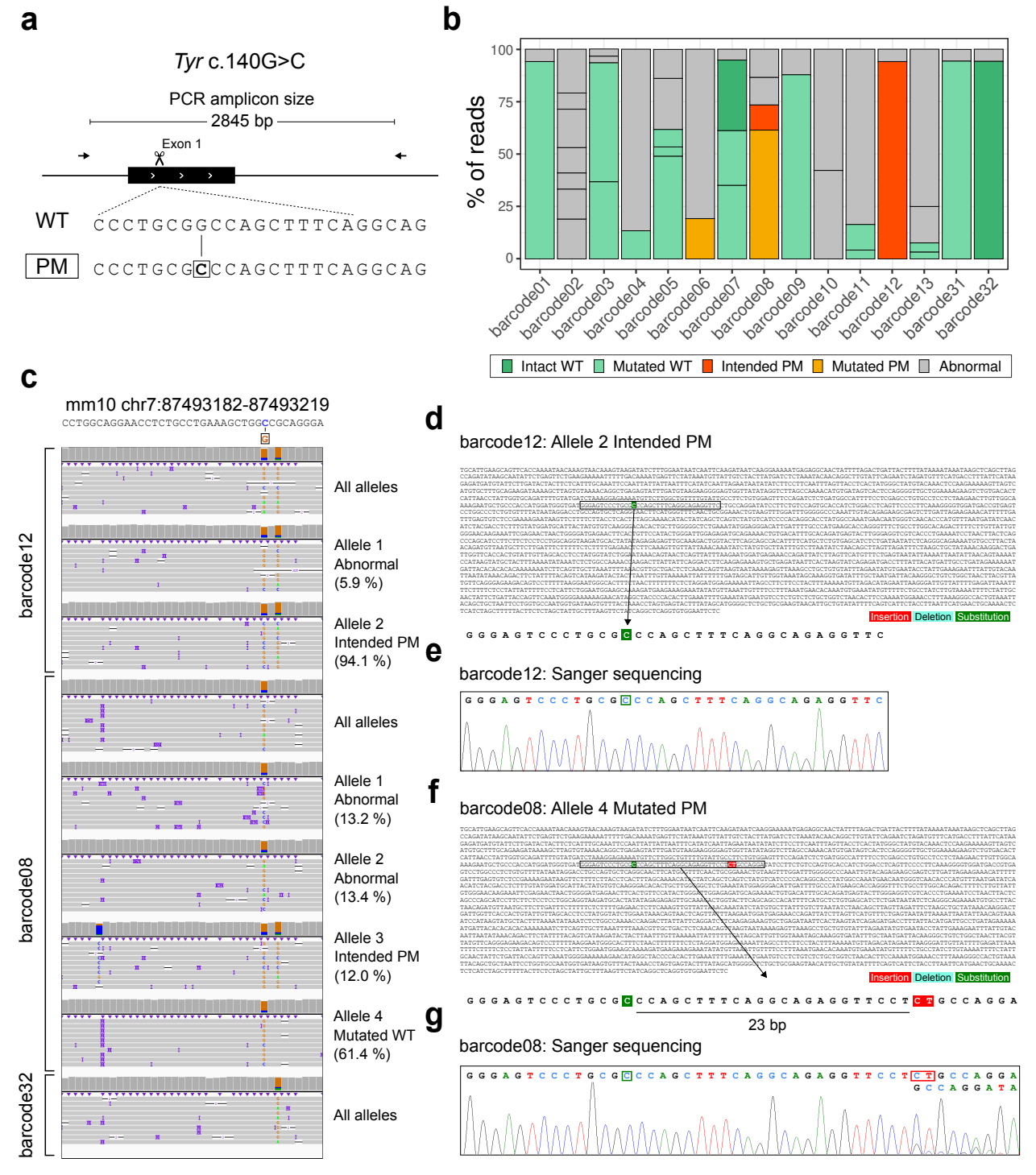
Additional file 2.

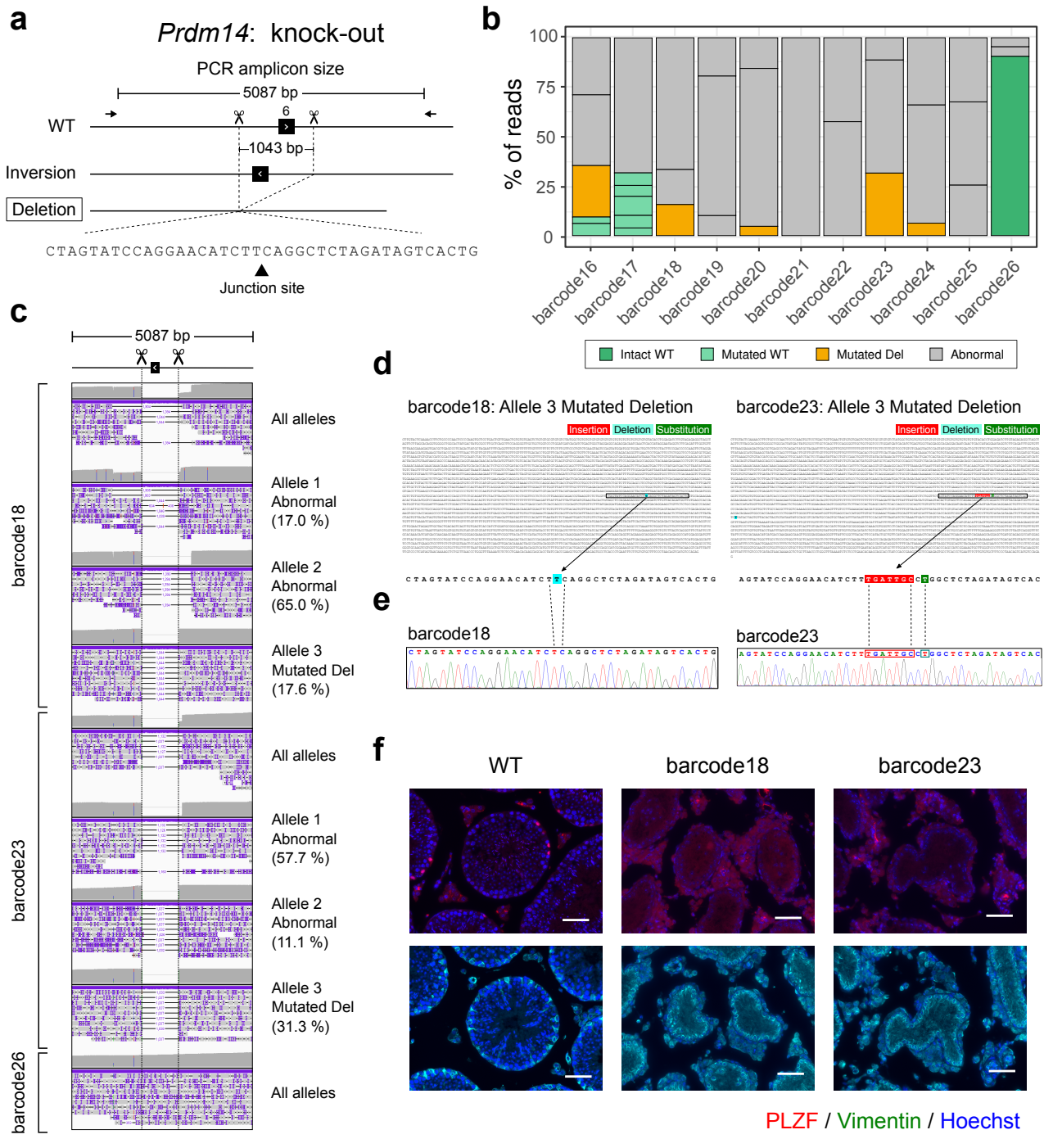
Tables S1–S8 (MS Excel).

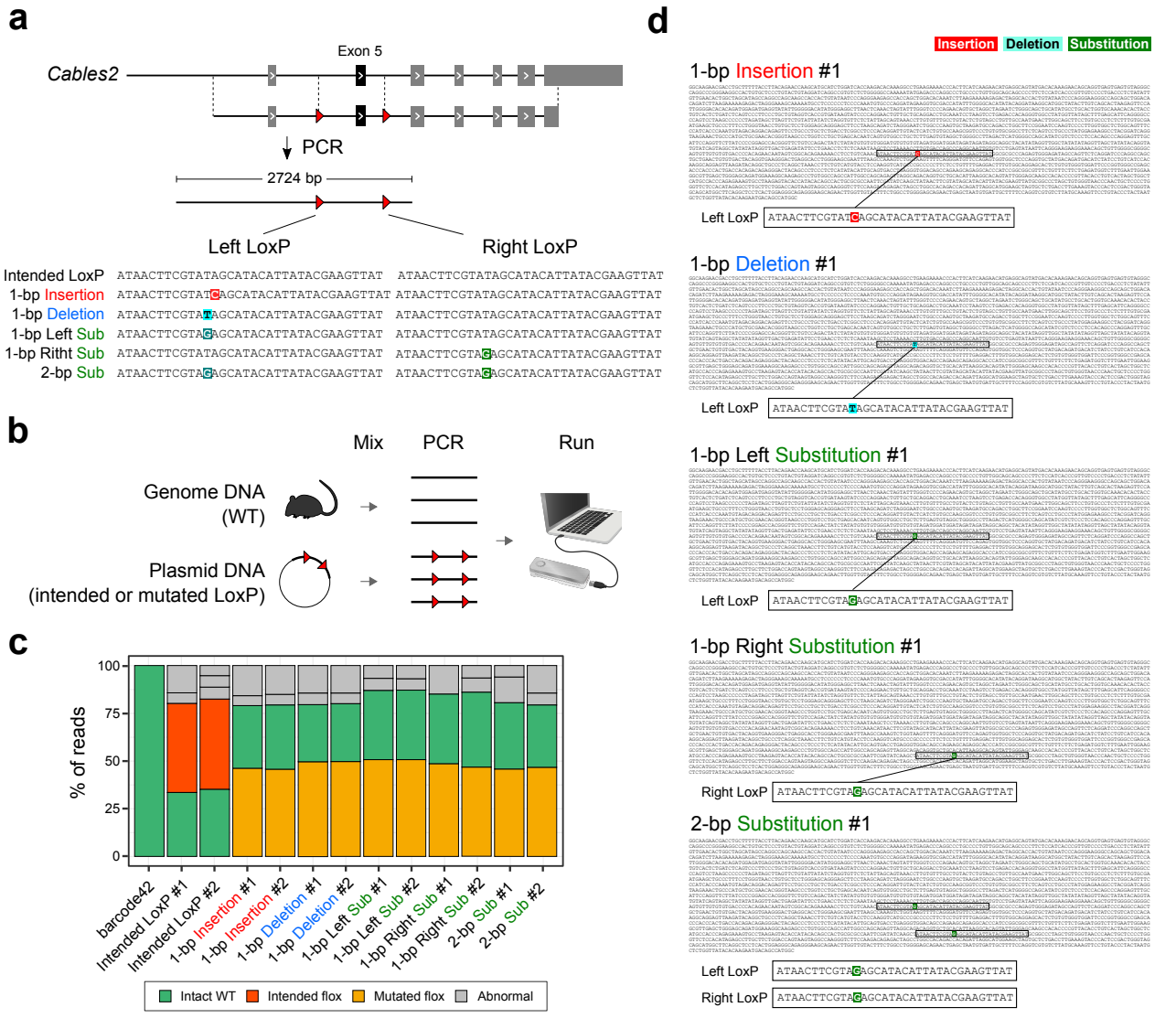
Additional file 3.

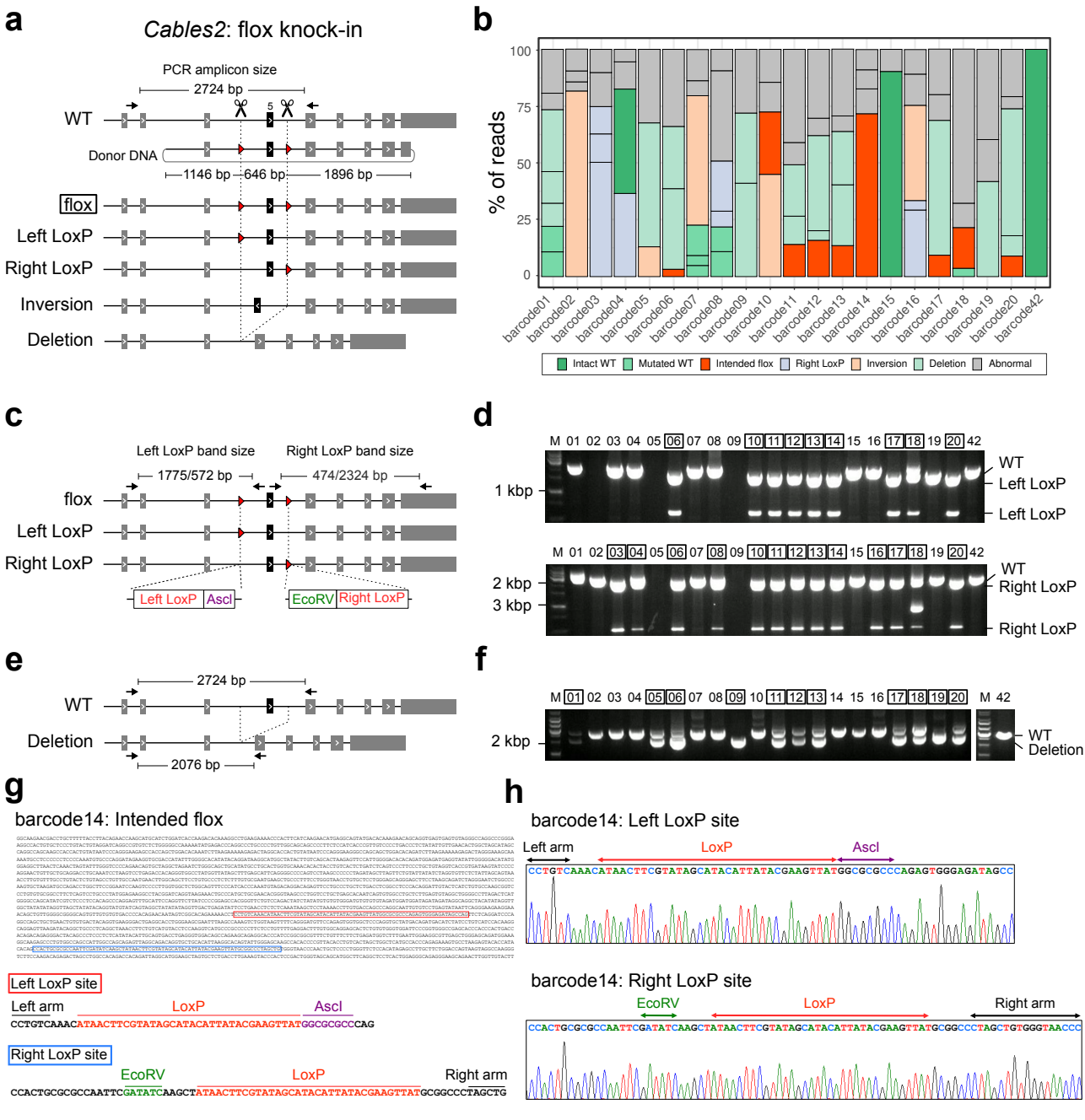
DAJIN consensus sequences (HTML).

Sample	Allele ID	% of reads	Allele type	Indel	Large Indel	Design
barcode01	1	21.5	Mutated WT	-	-	-
barcode01	2	19.3	Abnormal	+	+	-
barcode01	3	35.5	Intended Target	-	-	+
barcode01	4	23.7	Mutated Target	+	-	-
barcode02	1	95.5	Intact WT	-	-	-
barcode02	2	4.5	Abnormal	+	+	-


a
Tyr c.140G>C
PCR amplicon size
2845 bp
Exon 1
WT
CCCTGCGGCCAGCTTTCAGGCAG
PM
CCCTGCGCCAGCTTTCAGGCAG
b
% of reads
100
75
50
25
0
barcode01
barcode02
barcode03
barcode04
barcode05
barcode06
barcode07
barcode08
barcode09
barcode10
barcode11
barcode12
barcode13
barcode31
barcode32
Intact WT
Mutated WT
Intended PM
Mutated PM
Abnormal
c
mm10 chr7:87493182-87493219
CCTGGCAGGAACCTCTGCCTGAAAGCTGGCCGCAGGGA
barcode12
All alleles
Allele 1
Abnormal
(5.9 %)
Allele 2
Intended PM
(94.1 %)
barcode08
All alleles
Allele 1
Abnormal
(13.2 %)
Allele 2
Abnormal
(13.4 %)
Allele 3
Intended PM
(12.0 %)
Allele 4
Mutated WT
(61.4 %)
barcode32
All alleles
d
barcode12: Allele 2 Intended PM
Insertion
Deletion
Substitution
GGGAGTCCCTGCGCCCAGCTTTCAGGCAGAGGTTC
e
barcode12: Sanger sequencing
GGGAGTCCCTGCGCCCAGCTTTCAGGCAGAGGTTC
f
barcode08: Allele 4 Mutated PM
Insertion
Deletion
Substitution
GGGAGTCCCTGCGCCCAGCTTTCAGGCAGAGGTTCCTCTGCCAGGA
23 bp
g
barcode08: Sanger sequencing
GGGAGTCCCTGCGCCCAGCTTTCAGGCAGAGGTTCCTCTGCCAGGA
GCCAGGATA


a
Prdm14: knock-out
PCR amplicon size
5087 bp
6
WT
1043 bp
Inversion
Deletion
CTAGTATCCAGGAACATCTTCAGGCTCTAGATAGTCACTG
Junction site
b
% of reads
100
75
50
25
0
barcode16
barcode17
barcode18
barcode19
barcode20
barcode21
barcode22
barcode23
barcode24
barcode25
barcode26
Intact WT
Mutated WT
Mutated Del
Abnormal
c
5087 bp
barcode18
barcode23
barcode26
All alleles
Allele 1 Abnormal (17.0 %)
Allele 2 Abnormal (65.0 %)
Allele 3 Mutated Del (17.6 %)
All alleles
Allele 1 Abnormal (57.7 %)
Allele 2 Abnormal (11.1 %)
Allele 3 Mutated Del (31.3 %)
All alleles
d
barcode18: Allele 3 Mutated Deletion
barcode23: Allele 3 Mutated Deletion
Insertion
Deletion
Substitution
e
barcode18
barcode23
f
WT
barcode18
barcode23
PLZF / Vimentin / Hoechst


a
Cables2
Exon 5
PCR
2724 bp
Left LoxP
Right LoxP
Intended LoxP ATAACTTCGTATAGCATACATTATACGAAGTTAT ATAACTTCGTATAGCATACATTATACGAAGTTAT
1-bp Insertion ATAACTTCGTATGAGCATACATTATACGAAGTTAT ATAACTTCGTATAGCATACATTATACGAAGTTAT
1-bp Deletion ATAACTTCGTATAGCATACATTATACGAAGTTAT ATAACTTCGTATAGCATACATTATACGAAGTTAT
1-bp Left Sub ATAACTTCGTAGAGCATACATTATACGAAGTTAT ATAACTTCGTATAGCATACATTATACGAAGTTAT
1-bp Ritht Sub ATAACTTCGTATAGCATACATTATACGAAGTTAT ATAACTTCGTAGAGCATACATTATACGAAGTTAT
2-bp Sub ATAACTTCGTAGAGCATACATTATACGAAGTTAT ATAACTTCGTAGAGCATACATTATACGAAGTTAT
b
Mix
PCR
Run
Genome DNA (WT)
Plasmid DNA (intended or mutated LoxP)
c
% of reads
100
75
50
25
0
barcoded42
Intended LoxP #1
Intended LoxP #2
1-bp Insertion #1
1-bp Insertion #2
1-bp Deletion #1
1-bp Deletion #2
1-bp Left Sub #1
1-bp Left Sub #2
1-bp Right Sub #1
1-bp Right Sub #2
2-bp Sub #1
2-bp Sub #2
Intact WT
Intended flox
Mutated flox
Abnormal
d
Insertion
Deletion
Substitution
1-bp Insertion #1
Left LoxP ATAACTTCGTATGAGCATACATTATACGAAGTTAT
1-bp Deletion #1
Left LoxP ATAACTTCGTATAGCATACATTATACGAAGTTAT
1-bp Left Substitution #1
Left LoxP ATAACTTCGTAGAGCATACATTATACGAAGTTAT
1-bp Right Substitution #1
Right LoxP ATAACTTCGTAGAGCATACATTATACGAAGTTAT
2-bp Substitution #1
Left LoxP ATAACTTCGTAGAGCATACATTATACGAAGTTAT
Right LoxP ATAACTTCGTAGAGCATACATTATACGAAGTTAT


a
Cables2: flox knock-in
PCR amplicon size
2724 bp
WT
Donor DNA
1146 bp
646 bp
1896 bp
flox
Left LoxP
Right LoxP
Inversion
Deletion
b
% of reads
100
75
50
25
0
barcode01
barcode02
barcode03
barcode04
barcode05
barcode06
barcode07
barcode08
barcode09
barcode10
barcode11
barcode12
barcode13
barcode14
barcode15
barcode16
barcode17
barcode18
barcode19
barcode20
barcode42
Intact WT
Mutated WT
Intended flox
Right LoxP
Inversion
Deletion
Abnormal
c
Left LoxP band size
Right LoxP band size
1775/572 bp
474/2324 bp
flox
Left LoxP
Right LoxP
Left LoxP
AscI
EcoRV
Right LoxP
d
M 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 42
1 kbp
WT
Left LoxP
Left LoxP
M 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 42
2 kbp
3 kbp
WT
Right LoxP
Right LoxP
e
2724 bp
WT
Deletion
2076 bp
f
M 01 02 03 04 05 06 07 08 09 10 11 12 13 14 15 16 17 18 19 20 M 42
2 kbp
WT
Deletion
g
barcode14: Intended flox
Left LoxP site
Left arm
LoxP
AscI
CCTGTCAAACATAACTTCGTATAGCATACATTATACGAAGTTATGGCGCGCCCAG
Right LoxP site
EcoRV
LoxP
Right arm
CCACTGCGCGCCAATTCGATATCAAGCTATAACTTCGTATAGCATACATTATACGAAGTTATGCGGCCCTAGCTG
h
barcode14: Left LoxP site
Left arm
LoxP
AscI
CCTGTCAAACATAACTTCGTATAGCATACATTATACGAAGTTATGGCGCGCCCAGAGTGGGAGATAGCC
barcode14: Right LoxP site
EcoRV
LoxP
Right arm
CCACTGCGCGCCAATTCGATATCAAGCTATAACTTCGTATAGCATACATTATACGAAGTTATGCGGCCCTAGCTGTGGGTAACCC