The complex architecture of plant transgene insertions

Florian Jupe[1,2,*], Todd P. Michael[3*], Angeline C. Rivkin[1*], Mark Zander[1], S. Timothy Motley[3], Justin P. Sandoval[1], R. Keith Slotkin[4], Huaming Chen[1], Rosa Castagnon[1], Joseph R. Nery[1], Joseph R. Ecker[1,5,$]

[1]Genomic Analysis Laboratory, Salk Institute for Biological Studies, La Jolla, CA 92037 USA

[2]Current address: Monsanto Company, Creve Coeur, MO 63141 USA

[3]J. Craig Venter Institute, La Jolla, CA 92037 USA

[4]Department of Molecular Genetics, Ohio State University, Columbus, OH 43210 USA

[5]Howard Hughes Medical Institute, The Salk Institute for Biological Studies, La Jolla, CA 92037 USA

[*]These authors contributed equally to this work

[$]Correspondence: ecker@salk.edu

Keywords: T-DNA, Agrobacterium, *Arabidopsis thaliana*, nanopore sequencing, optical mapping, siRNA, DNA methylation, transgenic, silencing

Abstract

Over the last 35 years the soil bacterium *Agrobacterium tumefaciens* has been the workhorse tool for plant genome engineering. Replacement of native tumor-inducing (Ti) plasmid elements with customizable cassettes enabled insertion of a sequence of interest called Transfer DNA (T-DNA) into any plant genome. Although these T-DNA transfer mechanisms are well understood, detailed understanding of structure and epigenomic status of insertion events was limited by current technologies. To fill this gap, we analyzed transgenic *Arabidopsis thaliana* lines from three widely used collections (SALK, SAIL and WISC) with two single molecule technologies, optical genome mapping and nanopore sequencing. Optical maps for four randomly selected T-DNA lines revealed between one and seven insertions/rearrangements, and for the first time the actual length of individual transgene insertions from 27 to 236 kilobases. *De novo* nanopore sequencing-based genome assemblies for two segregating lines resolved T-DNA structures up to 36 kb into the insertions and revealed large-scale T-DNA associated translocations and exchange of chromosome arm ends. The multiple internally rearranged nature of T-DNA arrays made full assembly impossible, even with long nanopore reads. For the current TAIR10 reference genome, nanopore contigs corrected 83% of non-centromeric misassemblies. This unprecedented nucleotide-level definition of T-DNA insertions enabled the mapping of epigenome data. We identify variable small RNA transgene targeting and DNA methylation. SALK_059379 T-DNA insertions were enriched for 24nt siRNAs and contained dense cytosine DNA methylation. Transgene silencing via the RNA-directed DNA methylation pathway was confirmed by *in planta* assays. In contrast, SAIL_232 T-DNA insertions are predominantly targeted by 21/22nt siRNAs, with DNA methylation and silencing limited to a reporter, but not the resistance gene. With the emergence of genome editing technologies that rely on *Agrobacterium* for gene delivery, this study provides new insights into the structural impact of engineering plant genomes and demonstrates the utility of state-of-the-art long-range sequencing technologies to rapidly identify unanticipated genomic changes.

Introduction

Plant genome engineering using the soil microorganism *Agrobacterium tumefaciens* has revolutionized plant science and agriculture by enabling identification and testing of gene functions and providing a mechanism to equip plants with superior traits [1, 2, 3]. Transfer (T)-DNA insertional mutant projects have been conducted in important dicot and monocot models, and over 700,000 lines with gene affecting insertions have been generated in *Arabidopsis thaliana* (*Arabidopsis* henceforth) alone (reviewed in O'Malley [4]). Targeted T-DNA sequencing approaches were conducted on approximately 325,000 of these lines to identify the disruptive transgene insertions and to link genotype with phenotype.

This wealth of information has been made available on this website http://signal.salk.edu/Source/AtTOME_Data_Source.html, which was iteratively updated since 2001, and which was accessed over 10 million times by 2017.

The *Agrobacterium* strains used in research projects are no longer harmful to the plant because the oncogenic elements of the tumor-inducing (Ti) plasmid have been replaced by a customizable cassette that includes a diverse set of *in planta* regulatory elements. *Agrobacterium*-mediated transgene integration occurs through excision of the T-DNA strand between two imperfect terminal repeat sequences [5], the left border (LB) and right border (RB) [6], and translocation into the host cell nucleus (reviewed in Nester [7]). Hijacking the plant molecular machinery, the T-DNA is integrated at naturally occurring double strand breaks through annealing and repair at sites of microhomology [8, 9]. While the exact mechanisms behind this error prone integration are poorly understood, it is known that insertion events generally occur at multiple locations throughout the genome [5, 10]. T-DNA insertions also frequently contain the vector backbone, and occur as direct or inverted repeats of the T-DNA, resulting in large intra- and inter-chromosomal rearrangements [6, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].

Knowledge of insertion site, copy number and potential backbone insertions is crucial from scientific as well as regulatory perspectives. These aspects are routinely assessed using laborious Southern blotting, Thermal Asymmetric Interlaced (TAIL) PCR or targeted short-read sequencing [4]. One of the few attempts to gain deeper insight into an engineered genome was for transgenic Papaya, using a Sanger sequencing approach [14]. This work identified three insertion events, each less than 10 kilobases (kb) in length, however large repeat structures with high sequence identity are generally impossible to assemble using short-read sequences [21].

Current knowledge of structural genome changes and epigenetic stability in transgenic plants is limited. In this study we report on the genome structures of four *Arabidopsis* T-DNA floral-dip transformed plants, and are for the first time able to report the lengths of T-DNA insertions up to 236 kilobases. We have long-molecule evidence for further genome structural rearrangements including chromosomal translocations. To study such large insertions and rearrangements at the sequence level, we *de novo* assembled the genomes of two multi-insert lines (SALK_059379 and SAIL_232) and the reference accession Columbia-0 (Col-0) using Oxford Nanopore MinION (ONT) reads to very high contiguity. We present polished contigs that span chromosome arms and reveal the scrambled nature of T-DNA and vector backbone insertions and rearrangements in high detail. We subsequently tested transgene expression and functionality, and show differential epigenetic effects of the insertions on the transgenes between the two tested vector backbones. Small interfering RNA (siRNAs) species induced transgene silencing through the RNA-directed DNA methylation (RdDM) pathway of the entire T-DNA strand in the SALK-vector background, and in contrast the transgene remained active in the SAIL-line.

In this work, we report how technological advances enabled us to establish the size of T-DNA insertions in four independent mutant lines, and to assemble and analyse the genomes and epigenomes of two such lines to unprecedented detail.

Results

Optical maps reveal size and structure of T-DNA insertions

To study both the location and size of T-DNA insertions into the genome, we randomly selected four transgenic *Arabidopsis* lines and generated optical genome maps using the Bionano Genomics Irys system (BNG; San Diego, CA) from pooled leaf tissue. These Columbia-based plant lines have previously been transformed by *Agrobacterium* using different binary vector constructs: SALK_059379 and SALK_075892 with pROK2 (T-DNA strand 4.4 kb; [22]), SAIL_232 with pCSA110 (T-DNA strand 7.1 kb; [23, 24]), and WiscDsLox_449D11 with pDSLox (T-DNA strand 9.1 kb; [19]). These segregating lines had no visual mutation-induced phenotype, and resembled their Columbia parent plants phenotypically. The fluorescently-labeled DNA molecules used for optical mapping had an average length of up to 288 kb which enables conservation of long-range information across transgene insertion sites (Table 1 and Fig. 1, 2; Supplementary Table 1). We compared assembled optical maps for each

of the four lines to the *Arabidopsis* TAIR10 reference genome sequence and observed up to four structural variations, here T-DNA insertions, with sizes ranging from 27 kb to 236 kb (Fig. 1, 2, Table 1, and Supplementary Table 2). Although insertion sizes larger than the actual T-DNA cassette have been reported [6, 11, 12, 17, 19, 20], the here measured length for insertions derived from short T-strands exceeded our expectations by ~20-60 fold. Moreover, optical genome maps of the SAIL_232 line identified a total of seven genomic changes including three insertion events, one inversion involving ~500 kb on chromosome 1, an inverted translocation on chromosome 3 that involves the exchange of two adjacent regions between 2.6-3.4Mb (847 kb) with 8.9-10.1 Mb (1,193 kb), as well as a swap between chromosome arm ends (chromosomes 3 and 5; Fig. 2). Previous short-read sequencing projects to identify insertion numbers (TDNAseq; http://signal.salk.edu/cgi-bin/tdnaexpress) provided evidence for two (WiscDsLox_449D11), three (SALK_075892), four (SALK_059379) and five (SAIL_232) insertion sites, thus less than what we have observed through the optical genome mapping approach (Supplementary Table S5).

Table 1: Optical genome maps and nanopore genome assemblies identified T-DNA insertions from segregant samples. Optical maps and MinION assemblies were aligned against the TAIR10 reference genome to detect coverage, number of insertions, and maximum insertion site. Individual ONT reads identified further insertions or rearrangements that were absent from the segregant assemblies.

	Optical Maps			MinION Assemblies			MinION individual reads	
	TAIR coverage	# Insertions/ Rearrangements	max. insert kb	TAIR coverage	# Insertions/ Rearrangements	max. resolved insert kb	# reads	# Insertions/ Rearrangements
Salk_059379	98 %	4	206	98 %	3	39	503,388	2
Salk_075892	96 %	3	96	-	-	-	-	-
SAIL_232	97 %	7	236	98 %	9	50	1,297,359	3
Wisc_DsLox449D11	94 %	1	164	-	-	-	-	-

Assembly of highly contiguous genomes from MinION data

The number of insertions in SALK_059379 and the type of rearrangements observed in SAIL_232 sparked our interest in analyzing these genomes at greater (nucleotide) resolution. We sequenced these engineered genomes, alongside the reference Col-0 (ABRC accession CS70000) using the ONT MinION device. We performed nanopore sequencing on each line using a single R9.4 flow cell (Table 1 and Supplementary Table 1) and assembled each genome using minimap/miniasm followed by three rounds of racon [25] and one round of Pilon [26]. We assembled the three lines into 40 contigs (Col-0; longest 16,115,063 bp), 59 contigs (SAIL_232; longest 16,070,966 bp) and 139 contigs (SALK_059379; longest 8,784,268 bp) (Supplementary Table 1). Individual whole genome alignments to the TAIR10 reference show over 98% coverage with 39 and 57 contigs for SAIL_232 and SALK_059379, respectively (Table 1; Supplementary Fig. 1; Supplementary Table 2). The remaining short contigs (< 50 kb) encode only highly repetitive sequences such as rDNA and centromeric repeats that cannot be placed onto the reference. Due to the high contiguity, and if not disrupted by T-DNA insertions, completed chromosome arms were contained within one or two contigs, while contiguity declined with repeat content towards the centromere (Supplementary Fig. 1). Chromosome arm spanning contigs covered telomere repeats and at least the first centromeric repeat, thus capturing 100% of the genic content. The remaining ~3.9 Mb alignment-free reference regions were exclusively centromeric, and are also not necessarily correct in the TAIR10 reference. The Col-0 contigs covered over 99% of the TAIR10 reference, and the only discrepancies occur at the centromeres (Supplementary Fig. 1). High contiguity and quality of this assembly closed 38 of 46 previously identified non-centromeric misassemblies in the TAIR10 reference genome (Fig. 1a, Supplementary Table 3). Optical genome map alignments confirmed that all nanopore sequence contigs were chimera-free, while only eleven (Col-0 and SALK_059379) or three contigs (SAIL_232) contained misassembled non-T-DNA repeats (Supplementary Table 4).

One aim of creating near complete genome assemblies was to enable the structural resolution of transgene insertions at nucleotide level, rather than with genome scaffolding alone. We next assessed contiguity at sites of T-DNA insertion and after alignment to optical maps concluded that the shorter insertions, SALK:chr2_18Mb (28 kb), SAIL:chr3_9Mb (11 kb), SAIL:chr3_21Mb (25 kb; Fig. 2c) and SAIL:chr5_22Mb (11 kb) are completely assembled. Because of extensive repeats, much larger T-DNA insertions collapse upon themselves,

although contigs reaching up to 39 kb into the insertions from the flanking genomic sequences could be assembled.

T-DNA independent chromosomal inversion in the SAIL Col-3 background

Chromosome 1 in the SAIL_232 line was assembled into a single contig (SAIL_contig_20), spanning an entire chromosome arm from telomere to the first centromeric repeat arrays. Compared with the Col-0 reference genome, we found a 512 kb inversion in the upper arm (SAIL_chr1:11,703,634-12,215,749). Because we could not find a signature of T-DNA insertion at the inversion edges, we posited that this event may have resulted from a preexisting structural variation of the particular Columbia strain used for the SAIL project (Col-3; ABRC accession CS873942). To test this hypothesis, we genotyped SAIL_232 alongside two randomly selected lines of the same collection (SAIL_59 and SAIL_107) using primers specific to the reference Col-0 CS7000 genome and the SAIL-inverted state (see Methods). PCR analysis confirmed that this “inversion” was common to all three independent SAIL-lines tested, and absent from Col-0 CS70000. Thus the event was not due to the T-DNA mutagenesis, rather is an example of the genomic “drift” among strains used as Columbia “reference”.

SALK_059379 T-DNA insertions are conglomerates of T-strand and vector backbone

To annotate T-DNA insertion sites within the assembled genomes, we searched for pROK2 and pCSA110 plasmid vector sequence fragments within the assembled contigs (Supplementary Table 5). The assembled SALK_059379 genome contained three of the four optical map identified T-DNA insertions: SALK:chr1_5Mb, chr2_15Mb and chr2_18Mb (Table 1, Fig. 1 a,c,e). Specifically, SALK:chr2_18Mb, the shortest identified insertion with 28,356 bp, was completely assembled (contig_7:3,690,254-3,719,373) and included a genomic deletion of 5,497 bp (chr2:18,864,678-18,870,175). Annotation of the insertion revealed two independent insertions; a T-DNA/backbone-concatemer (11,838 bp) from the centromere proximal end, and a T-DNA/backbone/T-DNA-concatemer (16,463 bp) from the centromere distal end, both linked by a guanine-rich segment of 55 bp (26 G’s) (Fig. 1e). Two independent insertion events, 5,497 bp apart, potentially created a double-hairpin through sequence homology, that was eventually excised and removed the intermediate chromosomal stretch (Supplementary Fig. 2).

The second and third insertions, SALK:chr1_5Mb (131 kb) and SALK:chr2_15Mb (207 kb), were partially assembled into contigs. SALK:chr1_5Mb contig_10 and contig_5 contain 25,132 bp and 33,736 bp T-DNA segments. Similarly, the extremely long chr2_15Mb insertion was partially contained within contig_4 and contig_7 (Fig. 1c), leaving an unassembled gap of 131 kb. The recovered structure of this insertion is noteworthy as it represents a conglomerate of intact T-DNA/backbone concatemers, as well as various breakpoints that introduced partial vector fragments with frequent changes of the insertion direction (Fig. 1c). Finally, the fourth insertion SALK_059379 (SALK:chr4_10Mb) was absent from the assembly. However, we observed a single ONT read (length 10,118 bp) supporting the presence of an insertion at this location. We also recovered a further ONT read (length 15,758 bp) that anchors at position chr3:20,141,394 and extends 14,103 bp into a previously unidentified T-DNA insertion (Supplementary Fig. 3). PCR amplification from DNA samples using the genomic/T-DNA junctions sequences from segregant and homozygous seeds, in fact, confirmed the presence of all five insertions, revealing heterozygosity within the ABRC-sourced seed material.

While optical maps were successful in placing long “T-DNA only” sequence contigs into the large gaps (e.g. SALK:chr1_5Mb), the four “short” T-DNA-only sequence contigs of ~50 kb or less did not contain sufficiently unique nicking pattern to confidently facilitate contig placement.

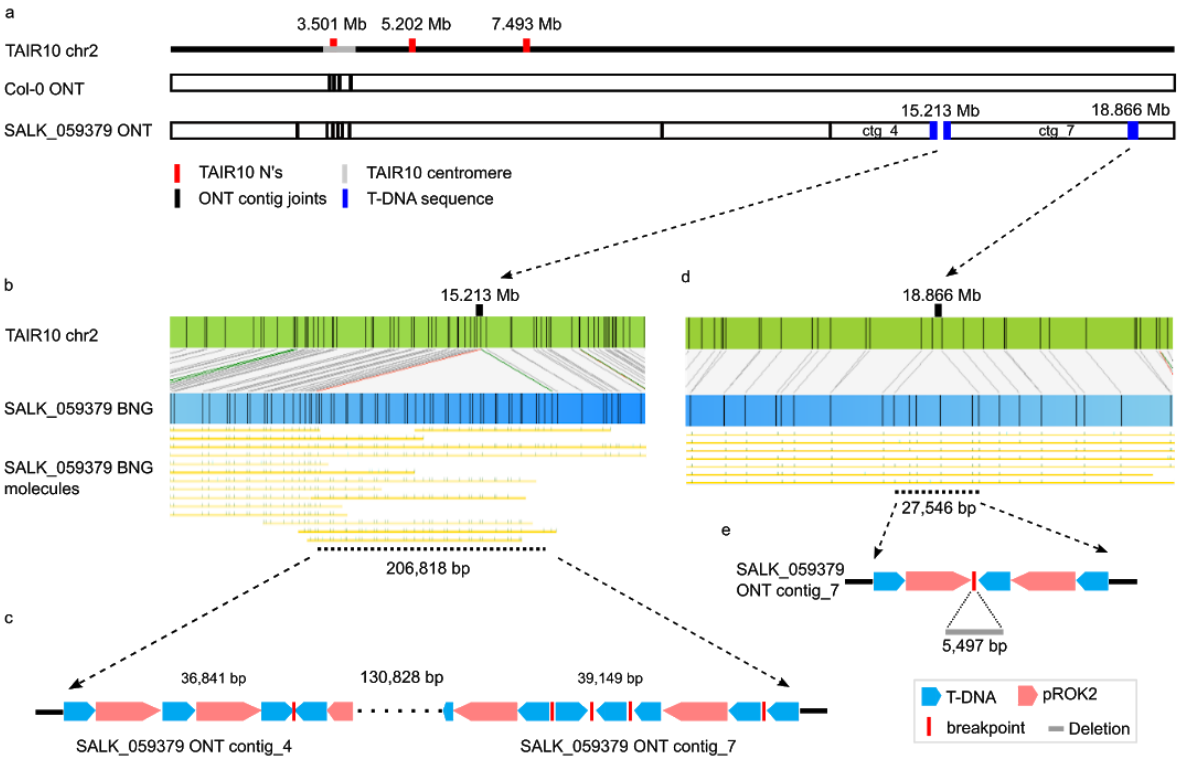


Fig. 1: SALK_059379 T-DNA insertions are T-strand and backbone conglomerations. Schematic representation of *Arabidopsis* mutant line SALK_059379 chromosome 2 T-DNA insertions as identified from (a, c, e) ONT sequencing *de novo* genome assemblies and (b, d) optical genome maps. (a) Graphical alignment of Col-0 and SALK_059379 ONT contigs to TAIR10 chromosome 2; two of the three TAIR10 misassemblies (red) are resolved within a single Col-0 ONT contig. Contig joints are represented as vertical black bars. Blue boxes indicate a broken (chr2_15Mb) and an assembled T-DNA insertion (chr2_18Mb). (b, d) represent the optical maps that identify T-DNA insertions including an alignment of individual BNG molecules. (c) ONT contigs 4 and 7 have the chr2_15Mb insertion site partially assembled and identify a mix of direct T-DNA (blue) and pROK2 backbone (red) concatemers or scrambled stretches with breakpoints as red bars. (e) The chr2_18Mb insertion is assembled and identifies a 5,497 bp chromosomal deletion.

Large-scale rearrangements reshape the SAIL_232 genome

We next searched the SAIL_232 ONT contigs for pCSA110 vector fragments, and were able to confirm all optical map observed genome insertions (Table 1). This search additionally identified a further T-DNA insertion at chr5:20,476,509 (Supplementary Table 5) that was not assembled in the optical genome maps.

We found that chromosome 3 harbored two major rearrangements (Fig. 2). The first was a translocation of a 1.19 Mb fragment (chr3:8,902,305-10,095,395), which split at an internal T-DNA insertion at reference position 9,343,053 bp (Fig. 2a). The resulting two fragments were independently inverted prior to integration just before position chr3:2,586,494 (Fig. 2 a,b). The second major change was a swap between the distal arms of chromosomes 3 and 5, which is supported by two optical maps as well as two ONT contigs (Fig. 2a-d). Here, chromosome 3 broke at 21,094,402 nt, and chromosome 5 at 18,959,379 nt and 20,476,664 nt, and the larger chromosomal fragments swapped places. Specifically, this translocation was captured in SAIL_contig_31, showing the fusion of chr5:20,476,664-end to chromosome 3 after position 21,094,402 nt. The reciprocal event joined (SAIL_contig_2) chromosome 3 fragment (21,094,407-end), almost seamlessly to chromosome 5 after reference position 18,959,379 nt (Fig. 2c,d). The genomic location of an excised fragment of chromosome 5 (chr5:18,959,380 - 20,476,663 found within SAIL_contig_11), was not determined (Fig. 2d).

Finally, the 81-kb insertion at SAIL:chr1_19Mb, consisting of four tandem T-DNA copies (~30 kb) followed by ~20 kb of breakpoint interspersed T-DNA and vector backbone was partially assembled (50,676 bp) at the 5' end of contig_5 (Fig. 2e). Although not assembled as part of the flanking SAIL_contig_47, we recovered single ONT reads that contain T-DNA as well as genomic DNA fragments. In summary, our optical maps perfectly aligned with the sequenced-based assembly of the T-DNA insertion haplotype (Fig. 2e).

T-DNA Integration occurs independently from both double-strand break ends

While both sequenced lines share similar numbers of T-DNA insertion events, the genome of the SAIL_232 plant line underwent more significant changes to its architecture. All genome insertion sites began and ended with the left border (LB) of the T-DNA strand, providing evidence for independent transgene integration at both ends of the DNA double-strand break. We did not recover any LB sequences at the chromosome/T-DNA junction, in line with literature reports that usually 73 - 113 bp are missing from the LB sequence inwards [15, 27]. Internal T-DNA sequence deletions were also seen at breakpoints within the insertion (Fig. 1c, Fig. 2c).

As observed for the SALK:chr2_18Mb chromosomal deletion (Fig. 1e and Supplementary Fig. 2), we cannot exclude that long homologous stretches between the independently inserted T-DNA/vector backbone concatemers represent inverted repeats.

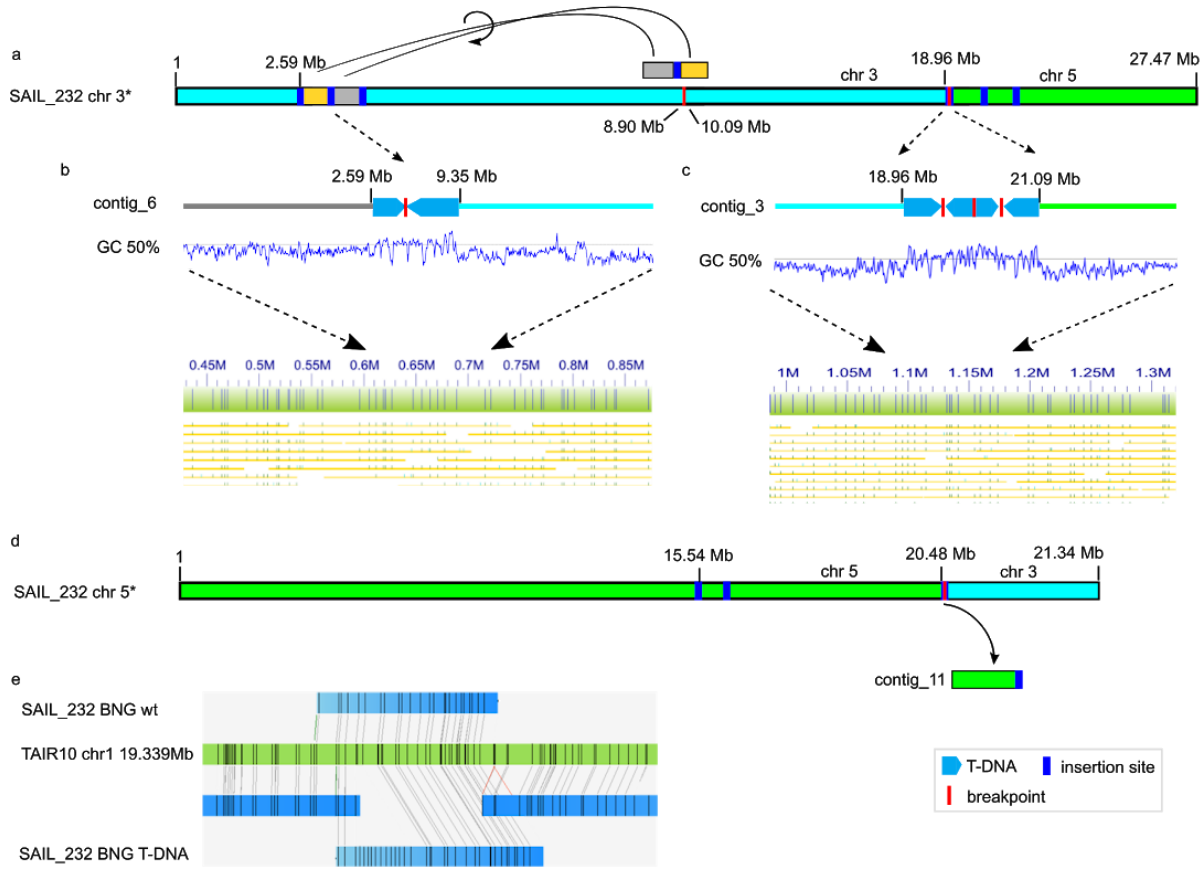


Fig. 2: T-DNA integration induced large scale rearrangements in the SAIL_232 genome. Two T-DNA induced translocations occurred on chromosomes 3 (light blue) and 5 (lime green). (a) Inverted translocation of a 1.1 Mb segment split at a T-DNA insertion site (blue bar) was moved to a distal part of the same chromosome arm. The other chromosome arm was swapped with chromosome 5 (d). (b) The ONT assembly identified the underlying T-DNA inserts (boxed arrows) including a breakpoint (red line). c) The chromosome swap joint site contains four pCSA110 vector fragments (boxed arrows) interspersed with breakpoints. These insertions change the CG signature in the respective region, and are confirmed by optical molecule alignment. (e) Optical maps (blue), aligned to *Arabidopsis* reference TAIR10 (green), are able to phase the wild-type (wt) and T-DNA haplotypes (T-DNA) for this particular insertion site. Black lines indicate Nt.BspQI nicking sites.

Transgenes are functional in SAIL-lines but are silenced in SALK-lines

We next wanted to assess the effects of T-DNA insertions on the epigenomic landscape. The pROK2 T-DNA strand contains the kanamycin antibiotic-resistance gene *nptII* under control of the bacterial nopaline synthase promoter (NOSp) and terminator (NOSt), and an empty multiple cloning site under control of the widely used Cauliflower mosaic virus 35S (CaMV 35S) promoter and NOSt. The CaMV 35S constitutive overexpression promoter has previously been described to cause homology-dependent transcriptional gene silencing (TGS) in crosses with other mutant plants that already contain a CaMV 35S promoter driven transgene [28, 29]. In germination assays, we confirmed that the kanamycin selective marker is not functional in SALK lines propagated for more than a few generations [4]. While ~75% of SALK_059379 seed initially germinated on kanamycin-containing plates, these seedlings stopped growth with the first root and cotyledons emerging and we were not able to recover adult plants (Fig. 3a). The SAIL_232 pCSA110 T-DNA segment encodes the herbicide resistance gene *bar* (phosphinothricin acetyltransferase), under control of a mannopine synthase promoter [23]. In contrast to SALK_059379, we confirmed proper transgene function by applying herbicide to soil-germinated plants [30, 31] (Fig. 3b). We corroborated these differential phenotypes by mapping RNAseq reads to the corresponding transformation plasmid and found that the SAIL_232 *bar* gene was expressed, while in SALK_059379 the *nptII* gene was not expressed, most likely due to epigenetic silencing (Fig. 3c,d).

Differential siRNA species define transgene silencing

We hypothesized that the observed plasmid dependent transgene expression is dictated by gene silencing via RNA-directed DNA methylation (RdDM) pathways [32]. To test the effects of small RNA and cytosine DNA methylation, we sequenced small RNA populations, as well as bisulfite converted whole genome DNA libraries for both lines. Our analyses identified abundant, yet divergent small RNA species (15-20nt, 21nt, 22nt, 23nt, 24nt) that mapped to genomic insertions of the *nptII* (pROK2) and *bar* (pCSA110) vector sequences in the SALK and SAIL lines, respectively (Fig. 3c,d). The *nptII* gene was highly targeted by 21/22nt short interfering (si)RNAs, as well as RdDM promoting 24nt siRNAs. The NOSp and duplicated NOSt were equally targeted by 22nt and 24nt siRNAs, although at lower numbers than the *nptII* gene itself (Fig. 3e). In contrast, the SAIL_232 *bar* gene and its promoter were targeted by high numbers of 21nt siRNAs, and lacking 24-mers (Fig. 3f). Interestingly, the pCSA110 encoded pollen specific promoter pLAT52, promoting GUS, was the only element highly targeted by 24nt siRNAs as well

as high cytosine methylation levels in all sequence contexts (CG, CHG and CHH methylation, where H is A, C or T) in the SAIL background (Fig. 3d, f; Supplementary Table 6). In contrast, the entire pROK2 T-DNA region has continuously high cytosine methylation in all three contexts. We found all elements to be fully CG and CHG methylated, and between 30% (*nptII*) and 71% (NOSt) of CHH's were methylated (Supplementary Table 7).

In summary, the GUS reporter in pCSA110 and all pROK2 transgenic elements are targeted by RdDM gene silencing pathway 24nt siRNAs. In contrast, the SAIL_232 pCSA110 *bar* gene was only targeted by 21/22nt siRNAs without any apparent effects on expression and the chemical resistance phenotype.

We were curious whether the 24nt siRNA and DNA methylation of the GUS gene observed in leave tissue suppresses expression in pollen, where it is driven by the strong pLAT52 promoter. Indeed a GUS-staining of mature pollen identified GUS signal in the tested SAIL line, thus suggesting that strong epigenetic marks can be broken by strong promoters (Fig. 3g-i).

While these observations are limited to the transformation vectors, we specifically looked at the individual junctions between genome and T-DNA. We only found few siRNA reads at two SALK_059379 (of eight) and six SAIL_232 (of 11), and we were not able to draw any conclusions. Similarly for bisulfite converted DNA reads, where we were able to identify DNA methylation signatures at only one junction in each background. In our experiments, we could hence not observe any signatures for siRNA or DNA methylation spreading outside of the T-DNA insertions.

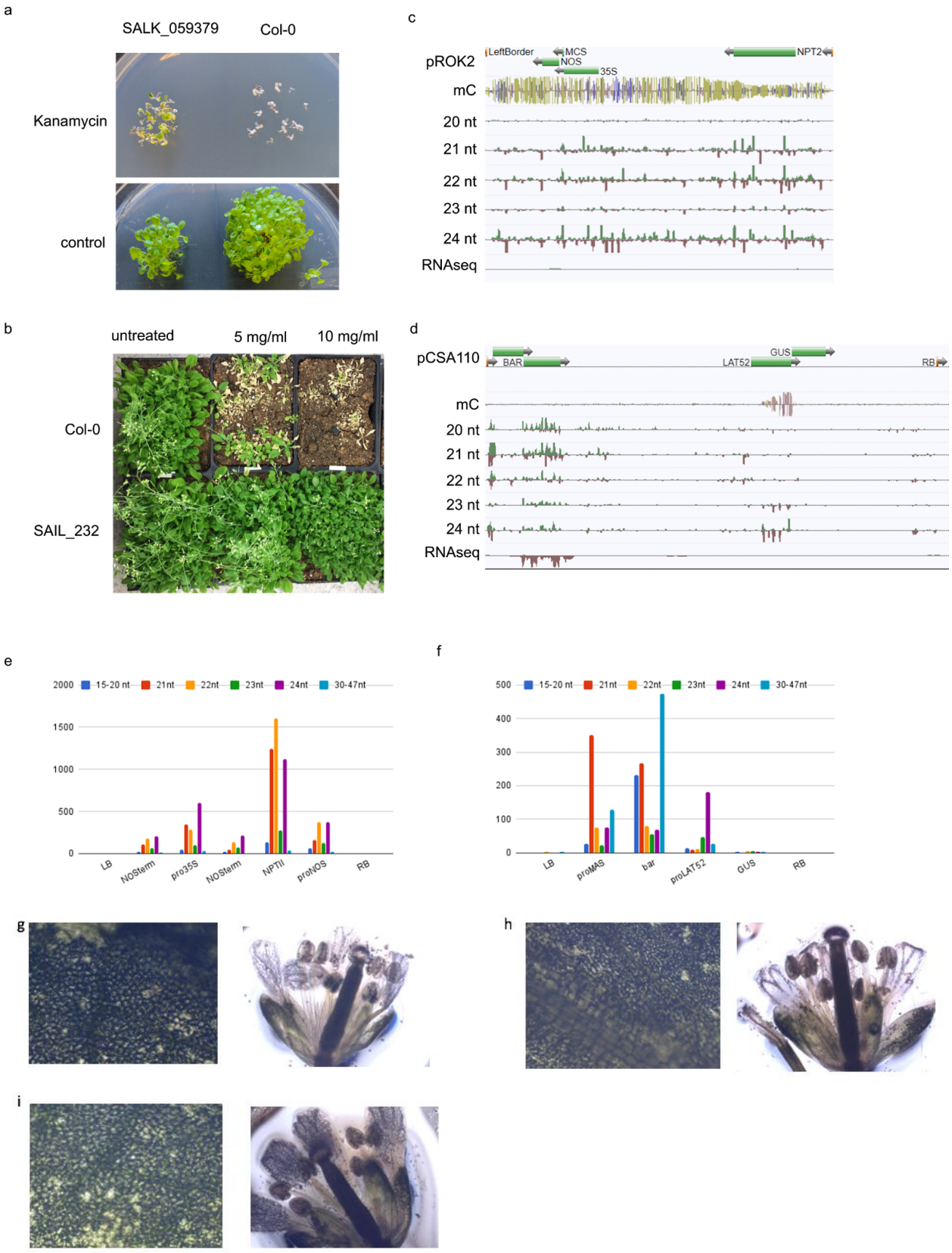


Fig. 3: SALK and SAIL T-strand sequences show extremely divergent epigenomic signatures. Transgene activity was tested by exposing plants to the respective selective agent: (a) SALK_059379 grown on media containing the antibiotic kanamycin (or empty control) and (b) the SAIL_232 sprayed with the herbicide Finale(™) at two different concentrations through the middle of the tray. All phenotypes are compared to the wild type Col-0 (a,b). Analysis of expression and epigenetic signatures on the corresponding T-DNA sequence is captured in browser shots for SALK_059379 plasmid pROK2 (c) and SAIL_232 plasmid pCSA110 (d): Illumina read mapping of bisulfite sequencing, RNAseq and different small RNA species. Quantification of individual siRNA read length against individual parts of the two plasmid sequences are reported in (e = pROK2) and (f = pCSA110). GUS staining of leaves and flowers of Sail_232 (g), with Col-0 (h) and Salk_059379 (i) as control.

Discussion

The engineering of a gene of interest into another plant genome is, in theory, a straightforward process where a single copy of this gene is repaired in between the ends of randomly occurring, or artificially induced (e.g. through nuclease activity), chromosomal double-strand breaks. Sure enough, the theory is far from the practice, as shown in numerous experiments since the first successful plant transformations in the early 1980s [1, 2, 3]. Although in many cases *Agrobacterium* transformation resulted in the expected outcome where the gene of interest is present and functional, it is common knowledge (yet usually unpublished) that the majority of transformation events are unsuccessful (excellent review by Gelvin [5]). For example, transgenes can be only partially present, or present but differentially or not expressed (e.g. Gelvin [10], Peach and Velten [35]). The majority of studied cases enforce the conclusion that a transgene's destiny is determined by alterations to the genome structure at the site of insertion or the structure of the insertion itself, whereby both can induce epigenetic features with detrimental effects on the transgene function. In order to understand these structural effects better, these need to be resolved. However, all attempts were limited due to the short-read length of sequencing technologies [36] and the barely proven repetivity of concatenated identical T-DNA (transgene) and vector backbone insertions [13]. Recent advances in the DNA sequencing space enabled the here presented detailed study of transgene insertions in the model plant *Arabidopsis thaliana*. We identified and analysed perturbations to the genome

structures of four randomly selected transgenic *Arabidopsis* T-DNA insertion lines from three of the most widely utilized plant mutant collections (SALK, SAIL, WISC). Optical genome maps created from DNA molecules with an average size of up to 288 kb unveiled genomic transgene insertions up to 236 kb in length. This is the first time that the exact size of such a large insertion from a T-strand with only a fraction of this size is reported. While two lines had only minimal T-DNA-induced genomic changes, SALK_059379 and SAIL_232 showed large-scale genome rearrangements. To better understand these, we utilized the latest generation nanopore long read DNA sequencing technology, and *de novo* assembled these two highly disturbed genomes along with the reference Col-0 genome. In line with earlier results [21], we generated sequence contigs up to chromosome arm length that enabled us to close 83% of non-centromeric (26% of all) misassemblies within the gold-standard reference genome (TAIR10) [37]. Moreover, sequence contigs for the two transgenic lines (SAIL_232 and SALK_059379) captured up to 39 kb of assembled T-DNA insertion sequence, providing sufficient information to better understand the complexity of such *Agrobacterium*-mediated transgene insertions (e.g. rearrangements insertions, deletions, etc.). Less repetitive rearrangements as found in SAIL_232, which included a 512 kb inversion on chromosome 1, were perfectly captured within chromosome arm spanning contigs.

Yet, the long read sequences were still not sufficient to provide complete contiguity of the highly repetitive T-DNA insertions or the centromeres. Although optical maps had a shorter N50 compared with ONT sequence contigs, the biological triggers that “disrupt” contig extension are different for both applications. Around the T-DNA insertion sites, we were able to take advantage of this, as optical genome maps preserved contiguity over all insertions and rearrangements, and thus provided absolute proof for these events in contrast to e.g. split-read mapping of short Illumina reads [36]. In addition, these maps provided a valuable size measure for the transgene insertion events. In summary, the combination of optical genome maps and nanopore sequencing provided unprecedented insight not only into the size of T-DNA insertions, but also helped us to describe the scrambled nature of transgene insertions.

Long reads provide invaluable means to analyze repetitive genome regions with higher certainty. We showed that the T-DNA sequence at the insertion start and end is in all cases derived from the same T-strand region, around 60 nucleotides into the left border sequence. Loss of the LB-end is most likely the result of endonuclease activity on the unprotected T-strand

end [38]. This observation, together with the highly scrambled insertion structure, led us to hypothesize that the final T-DNA insertion derives from two independent T-DNA insertions at each end of the genomic breakpoint, subsequent daisy-chaining into the observed long T-DNA concatemers, and possible interaction among the many resulting homologous regions. Long homologous stretches between the independently inserted T-DNA/vector backbone concatemers represent inverted repeats which can lead to hairpin formation, which stalls the recombination fork and excision during DNA repair [39]. While internal breakpoints and the scrambled insertion pattern that we observed are in support of this, we have no data to confirm whether this happened before, during, or after integration (Fig. 1, 2, Supplementary Fig. 2). We identified a single chimeric vector/vector insertion (SALK:chr2_15Mb) that most likely occurred through homology-dependent recombination of the two identical NOSt sequences in the T-DNA vector (pROK2:6778-7038 bp and 8607-8867 bp). Further experiments will be necessary to ultimately identify how and why T-DNA strands concatenate, and whether internal rearrangements and excisions occur during the insertion process or at later meiotic stages. Understanding why multi-insertions occur so frequently, and how these can be genetically or biochemically eliminated, will be of great interest also to the agricultural biotech industry.

Previous reports, spoken communication, and anecdotal experiences provided plenty of evidence that the antibiotic resistance marker genes, part of the transgene insertion cassette, are non-functional, most likely due to transcriptional or post-transcriptional DNA silencing [4]. To study these epigenetic features, we mapped short-read RNAseq, smallRNAseq, and bisulfite-converted DNA reads to each of the respective vector plasmids as reference. The high repetivity of the T-DNA insertions and the short reads utilized here made this necessary. Future studies can rely on sequencing a transcript in a single read, as well as extract DNA methylation information from kinetic information inherent to the long read nanopore data [40]. Our analyses revealed distinct epigenomic features for the two phenotypically indistinct transgenic lines; both transgenes are targeted for small RNA-level silencing. For pROK2, the small RNAs have progressed to the RdDM phase, and are targeting the entire transgene for methylation and functional shut-down of the resistance gene. For pCSA110, besides the Lat52 promoter, the sRNAs have been held at bay at the RNAi level, and RdDM has not set in, which results in a still functional herbicide tolerance gene.

On pROK2, the epigenetic pattern suggests that *nptII* is undergoing a mixture of expression-dependent silencing and conventional Pol IV-RdDM. It lacks heavy 24mers at the promoter, and makes 21-22mers within the coding region. These even levels of 21-22 and 24nt siRNAs further suggest that the *nptII* gene is transcribed at least at some level, but degraded very efficiently, and targeted for RdDM. This transgene is a heavy target of some type of RdDM (equal levels of CHH and CG) and produces no steady state mRNA, confirming the antibiotic sensitive phenotype.

Analyzing the epigenomic landscape of pCSA110 showed that the *bar* transcript is targeted by RNAi, likely RDR6-dependent since 21mers are derived from both strands. The few observed 24nt siRNA traces are potentially degradation products that occur during the normal life cycle of a highly expressed gene such as pCSA110 *bar* gene. However, the siRNAs are not successfully targeting RdDM, most likely due to absence of Pol V expression, so there is likely a reduced level of steady state mRNA, but enough to make a protein and provide the observed herbicide tolerance. It is not surprising that the pollen LAT52 promoter is heavily targeted by Pol IV-RdDM. This tomato-derived promoter is made out of retrotransposon sequences that have enough sequence conservation among plant species, that this promoter is always recognized and targeted by Pol IV-RdDM. And in addition, were able to show that this promoter is still active in pollen, and thus drives GUS expression in mature pollen.

We are intrigued by the distinct mechanisms that result in silencing of the entire pROK2 element, but only partial silencing of the pCSA110 element. An increased GC content of a transgenic construct, above whole genome GC content, was shown to decrease transgene DNA methylation and siRNA production from the *Agrobacterium*-introduced transgene constructs [41]. In our study, all reporter genes had considerably higher GC levels (*bar* 69%, *nptII* 60%, *GUS* 48%) compared to the *A. thaliana* genome (~36%), however the *bar* gene had the least siRNA and methylC-seq coverage, providing a possible explanation for silencing of *GUS* and *nptII*, but not *bar*. The methylated segments observed in the pCSA110 vector are in some distance to the left border, which might explain why we have not seen any spreading of DNA methylation into the flanking genomic regions. Similarly, although the SALK vector is overall heavily methylated, it is much less so towards the left border. Again we were not able to see spreading of DNA methylation into the adjacent chromosome in our data.

The relationship between *Agrobacterium*-mediated T-DNA integration and the host epigenome is mainly uncharted territory. Does the host chromatin state potentially guide the T-DNA integration, does the T-DNA affect the local chromatin environment at the site of insertion and if so, what are the molecular determinants for this phenomenon? Evidence for a clear guidance function of the chromatin state for T-DNA insertions is relatively sparse. There is indeed a bias of kanamycin-selected sites for genomic regions with low levels of DNA methylation and nucleosome occupancy in *Arabidopsis thaliana* but this preference cannot be observed in non-selected junctions indicating that T-DNA integrations occur rather randomly and uniformly distributed throughout the host genome [42].

Although the limited sample size examined here does not allow us to make accurate predictions of the average number of insertion events per plant, several key observations can be made about these important T-DNA mutant collections. First, individuals within the T-DNA mutant seed stocks can contain up to ten or more mutated genes without noticeable phenotypes, and many insertions are likely still segregating even in the background of identified homozygous gene insertion lines [4]. Some of these might also not have been picked up by previous sequencing approaches. Secondly, additional events such as the observed 512-kb inversion of chromosome 1 of SAIL_232 clearly show that caution should also be taken for “natural” genome variations between the different Columbia lab strains. In the genomics era, labs should make an effort and identify their *Arabidopsis* Columbia lab strain, or shift towards the recognized reference Col-0 CS70000 (available at ABRC). Third, we have characterized enormous rearrangements, such as the chromosome arm end swap, that have not produced any measurable visual phenotype. These could however have adverse effects on not immediately obvious biological pathways, and thus interfere with the mutation of interest, leading to wrong conclusions. Genetic and biochemical measures thus need to be taken to ensure the actual phenotype is from the insertion of interest. Fourth, 24-nt siRNA guided RdDM transgene silencing has occurred in both sequenced lines; in SAIL_232 this however targeted only one of the two transgenes, leaving the herbicide tolerance gene expressed. While we do not quite understand the biology behind this, it can occur for any transgene of interest in any plant species. To avoid unfavorable outcomes, and make the best out of this invaluable resource, best practice measures such as those described in O’Malley and Barragan [43] should be taken into account. Last but not least, the T-DNA mutant seed stocks are highly heterozygous and still segregating, which was reflected in the optical maps and genome

assemblies, where some insertions were missing, most likely due to lower allele frequency. In case of optical maps, we discovered a single instance of a phased genomic region, where two maps represent the wild type and the insertion allele (Fig. 2e). Although we did not observe phasing among the ONT contigs, we found several unassembled reads from T-DNA insertions absent from the ONT contigs, or that are not covered by optical maps and/or were not identified by TDNAseq (http://signal.salk.edu/cgi-bin/tdnaexpress). Existence of these insertions, such as SALK:chr3_20Mb and SALK:chr4_10Mb (Supplementary Fig. 3), were confirmed through genotyping of the seed pool. Our utilization of these reads shows great potential for ONT sequencing for explorative low coverage sequencing of larger crop genomes to identify structural genome variation, that can then be further explored using more targeted methods.

Small genome size and widespread utilization of T-DNA mutant collections to understand fundamental questions in plant biology made *Arabidopsis* the perfect organism to study the structural and transgenerational effects of transgene insertions. Here, we have paved the way and enabled the study of other transgenic model and crop plant species. Whether the observed extreme insertion sizes are unique to the model *Arabidopsis*, the Brassicaceae or common among all plants remains to be seen. But we have finally gained some understanding for how plant genomes cope with transgene insertions, similar to transposable elements, using siRNA and DNA methylation.

Material and Methods

Plant Material

Seeds were ordered from ABRC (https://abrc.osu.edu): SALK_059379 (segregant and homozygous lines), SALK_075892, SAIL_232 (seg.), SAIL_59 (seg.), SAIL_107 (seg.) and WiscDsLox_449D11 (seg.), SALKseq_061267. Plants were grown in a 20°C growth room with 13h light/11h dark cycles in peat-based soil supplemented with nutrients.

Testing Selection Markers

Kanamycin resistant lines- SALK and Wisc

MS/MES media was prepared (1L NP H2O, 2.2g MS, 0.5g MES, pH5.7) and autoclave sterilized. Kanamycin (50ug/mL) was added to half of the media, and plates with and without the antibiotic were poured. Seeds from the lines SALK_059379 (homozygous), SAIL_232, WiscDsLox_449D11, and Col-0 CS70000 control were sterilized and spotted on the plates, which were placed in the same growth conditions as the plant material in soil. Survival rates were measured 5 days after seeds were spotted and growth was observed for further 14 days.

Finale resistant lines- SAIL

One flat each of Col-0 (control) and SAIL_232 were grown in soil as outlined in *Plant Material*. Plants were sprayed with Finale at three different concentrations (5mg/L, 10mg/L, 20mg/L) at 5, 14, and 21 days after germination [1, 2].

GUS reporter staining- SALK and SAIL

Leaf and flower tissue was submerged in 1.5mL staining buffer (0.5mM ferrocyanide, 0.5mM ferricyanide, 0.5% Triton, 1mg/mL X-Gluc, 100mM Sodium Phosphate Buffer pH 7) and incubated at 37C overnight. The tissues were washed with decreasing concentrations of ethanol (100%, 80%, 65%, 50%). A light microscope was used to visualize the staining at 10x magnification for the leaves and 4x for the flowers.

MinION Nanopore Library Prep and Sequencing

5g of flash-frozen leaf tissue, pooled from the segregant seed stocks, was ground in liquid nitrogen and extracted with 20mL CTAB/Carlson lysis buffer (100mM Tris-HCl, 2% CTAB,

1.4M NaCl, 20mM EDTA, pH 8.0) containing 20µg/mL proteinase K for 20 minutes at 55℃. To purify the DNA, 0.5x volume chloroform was added, mixed by inversion, and centrifuged for 30 minutes at 3000 RCF. Purification was followed by a 1x volume 1:1 phenol: [24:1 chloroform:isoamyl alcohol] extraction. The DNA was further purified by ethanol precipitation (1/10 volume 3M sodium acetate pH 5.3, 2.5 volumes 100% ethanol) for 30 minutes on ice. The resulting pellet was washed with freshly-prepared ice-cold 70% ethanol, dried, and resuspended in 350µL 1x TE buffer (10mM Tris-HCl, 1mM EDTA, pH 8.0) with 5µL RNase A (Qiagen, Hilden) at 37℃ for 30 minutes, then at 4℃ overnight. The RNase A was removed by double extraction with 24:1 chloroform:isoamyl alcohol, centrifuging at 22,600x*g* for 20 minutes at 4℃ each time to pellet. An ethanol precipitation was performed as before for 3 hours at 4℃, washed, and resuspended overnight in 350µL 1x TE buffer. The genomic DNA samples were purified with the Zymogen Genomic DNA Clean and Concentrator-10 column (Zymo Research, Irvine). The purified DNA was prepared for sequencing with the Ligation Sequencing Kit 1D (SQK-LSK108, ONT, Oxford, UK) sequencing kit protocol. Briefly, approximately 2 µg of purified DNA was repaired with NEBNext FFPE Repair Mix for 60 min at 20^0C. The DNA was purified with 0.5X Ampure XP beads (Beckman Coulter, Brea). The repaired DNA was End Prepped with NEBNExt Ultra II End-repair/dA tail module and purified with 0.5X Ampure XP beads. Adapter mix (ONT, Oxford, UK) was added to the purified DNA along with Blunt/TA Ligase Master Mix (NEB) and incubated at 20^0C for 30 min followed by 10 min at 65^0C. Ampure XP beads and ABB wash buffer (ONT, Oxford, UK) were used to purify the library molecules, which were recovered in Elution buffer (ONT, Oxford, UK). The purified library was combined with RBF (ONT, Oxford, UK) and Library Loading Beads (ONT, Oxford, UK) and loaded onto a primed R9.4 Spot-On Flow cell. Sequencing was performed with a MinION Mk1B sequencer running for 48 hrs. Resulting FAST5 (HDF5) files were base-called using the Oxford Nanopore Albacore software (0.8.4) for the SQK-LSK108 library type.

Sequence extraction, assembly, consensus and correction

Raw ONT reads (fastq) were extracted from base-called FAST5 files using poretools [3]. Overlaps were generated using minimap [4] with the recommended parameters (-Sw5 -L100 -m0). Genome assembly graphs (GFA) were generated using miniasm [4]. Unitig sequences were extracted from GFA files. Three rounds of consensus correction was performed using Racon [5] based on minimap overlaps, and the resulting assembly was polished using Illumina

PCR-free 2x250 bp reads mapped with bwa [6] and pilon [7]. Genome stats were generated using QUAST [8].

The ONT reads were also assembled with the CANU assembler [9]; The minimap/miniasm assemblies were however of higher contiguity and resolved longer stretches of the T-DNA insertions. In addition, the minimap/miniasm assemblies were computed within 4-6 hours, and in contrast the CANU assemblies took between one (Col-0) and four weeks (SAIL). We hypothesize that the SAIL line inherent complex repeat structure caused the long compute time, while the reference Col-0 required only the expected time of one week. Also, we hypothesize that minimap/miniasm resolved the T-DNA structures more fully due to the fact it does not have a read correction step, which could lead to the collapsing of highly repetitive yet distinct T-DNA insertions.

Identification of individual T-DNA matching reads and annotation

Reads obtained via MinION sequencing were imported into Geneious R10 [10], along with the pROK2, and pCSA110 plasmid sequences. BLAST databases, were created for the complete set of SALK_059379 and SAIL_232 MinION reads, and their respective plasmid sequences were BLASTn searched against these databases. Reads with hits on the plasmid sequences were extracted, and BLASTn searched against the TAIR10 reference. NCBI megaBLAST [11] and Geneious 'map to reference' functions were used to annotate regions on each read corresponding to either TAIR10 or the plasmid sequence and the chromosomal regions were compared to the regions identified by TDNAseq and optical mapping to verify and refine T-DNA insertion start and stop coordinates.

Optical Genome Mapping

HMW DNA for optical genome mapping and nanopore sequencing was extracted as outlined in Kawakatsu [12]. Briefly, up to 5g of fresh mixed leaf and flower tissue (excluding chlorotic leaves and stems) pooled from the segregant seed stocks were homogenized in 50mL nuclei isolation buffer. Nuclei were separated from debris using a Percoll layer. Extracted nuclei were subsequently embedded in low melting agarose plugs and exposed to lysis buffer overnight. DNA was released by digesting the agarose with Agarase enzyme (Thermo Fisher Scientific).

HMW DNA was nicked with the enzyme Nt.BspQI (New England Biolabs, Ipswich, MA), fluorescently labeled, repaired and stained overnight following the Bionano Genomics nick-labeling protocol and accompanying reagents (Bionano Genomics, San Diego, CA) [13]. Each *A. thaliana* T-DNA insertion line was run for up to 120 cycles on a single flow cell on the Irys platform (Bionano Genomics, San Diego, CA). Collected data was filtered (SNR = 2.75; min length 100kb) using IrysView 2.5.1 software, and assembled using default "small" parameters. Average molecule length for assembly varied between 201 kb and 288 kb, and resulted in optical map N50 0.97 to 1.03 Mb. Derived assembled maps were anchored to converted TAIR10 chromosomes using the RefAligner tool and standardized parameters (-maxthreads 32 -output-veto-filter _intervals.txt$ -res 2.9 -FP 0.6 -FN 0.06 -sf 0.20 -sd 0.0 -sr 0.01 -extend 1 -outlier 0.0001 -endoutlier 0.01 -PVendoutlier -deltaX 12 -deltaY 12 -xmapchim 12 -hashgen 5 7 2.4 1.5 0.05 5.0 1 1 1 -hash -hashdelta 50 -mres 1e-3 -hashMultiMatch 100 -insertThreads 4 -nosplit 2 -biaswt 0 -T 1e-10 -S -1000 -indel -PVres 2 -rres 0.9 -MaxSE 0.5 -HSDrange 1.0 -outlierBC -xmapUnique 12 -AlignRes 2 -outlierExtend 12 24 -Kmax 12 -f -maxmem 128 -stdout -stderr). We utilized the resulting *.xmap, *_q.cmap and *_r.cmap files in the structomeIndel.py script (https://github.com/RyanONeil/structome) to identify T-DNA insertion locations and sizes. We used this script to further determine misassemblies in the ONT genome assemblies. Known Col-0 mis-assemblies [12] were subtracted from the list of derived locations.

Aligning contigs to TAIR10

We aligned ONT contigs to the TAIR10 reference using the BioNano Genomics RefAligner as above (Supplementary Table 3). ONT contigs were in silico digested with Nt.BspQI and used as template in RefAligner. The alignment output was manually derived from the .xmap output file.

Genotyping of structural genome variations

Primer3 was used to create oligos (Supplementary Table 5) for each insertion identified in the SALK_059379 and SAIL_232 lines. The forward primers correspond to the chromosome sequence ~500bp before the insertion start site, and the reverse to the plasmid sequence ~500bp after the insertion start site. A second set of reverse primers corresponding to the original wild-type chromosome ~1kb after the original forward primers. Individuals and pooled

tissue from SALK_059379, SAIL_232, SAIL_59, SAIL_107 and Col_0 CS7000 DNA was extracted using Qiagen DNeasy Plant Mini kit (following the protocol) and genotyped using these primers, with Col-0 as the negative control.

Bisulfite Sequencing

DNA was extracted using the DNeasy Plant Mini Kit (Qiagen, Hilden, D) according to the manufacturer's protocol from segregant plant pools, and quantified using the Qubit dsDNA BR assay kit. Illumina sequencing library preparation and bisulfite conversion was conducted as described in Kawakatsu [12]. Briefly, DNA was End Repaired using the End-It kit (Epicentre, Madison, WI), A-tailed using dA-Tailing buffer and 3µL Klenow (3' to 5' exo minus) (NEB, Ipswich, MA) and Truseq Indexed Adapters (Illumina, San Diego, CA) were ligated overnight. Bead purified DNA was quantified using the Qubit dsDNA BR assay kit, and stored at -20℃. At least 450ng of adapter ligated DNA was taken into bisulfite conversion which was performed according to the protocol provided with the MethylCode Bisulfite Conversion kit (Thermo Fisher, Waltham, MA). Cleaned, converted single stranded DNA was amplified by PCR using the KAPA U+ 2x Readymix (Roche Holding AG, Basel, CH): 2 min at 95℃, 30 sec at 98℃ [15 sec at 98℃, 30 sec at 60℃, and 4 min at 72℃] x 9, 10 min at 72℃, hold at 4℃. After amplification, the DNA was bead purified, the concentration was assessed using the Qubit dsDNA BR assay kit, and the samples were stored at -20℃. WGBS library was sequenced as part of a large multiplexed pool paired-end 150 bp on an Illumina HiSeq 4000.

RNA Libraries

RNA extractions performed using the RNeasy Plant Mini Kit (Qiagen, Hilden, D) from segregant plant pools. RNA seq libraries were prepared manually (SAIL_232) as described in Kawakatsu or using the Illumina NeoPrep Library Prep System (SALK_059379) (Illumina, San Diego, CA), following the NeoPrep control software protocol. The completed libraries were quantified using the Qubit dsDNA HS assay kit and stored at -20℃. SALK_059379 was sequenced in duplicate as single end 150 bp in a multiplexed pool on two lanes of an Illumina HiSeq 2500. SAIL_232 was sequenced as part of a large multiplexed pool paired-end 150 bp on an Illumina HiSeq 4000.

small RNA Libraries

We have extracted sRNA according to the protocol described in Vandivier [14] with modifications to the RNA extraction and size selection steps. In brief, 300mg flash-frozen leaf tissue from segregant plant pools was ground in liquid nitrogen and extracted with 700µL QIAzol lysis reagent (Qiagen, Hilden, D). The RNA was separated from the lysate using QIAshredder columns (Qiagen, Hilden, D) and purified with a ⅛ volume 24:1 chloroform:isoamyl alcohol (Sigma-Aldrich, St. Louis, MO) extraction followed by a 1mL wash of the aqueous phase with 100% ethanol. The RNA was further purified using miRNeasy columns (Qiagen, Hilden, D) with two 82µL DEPC-treated water washes. 37µL DNase was added to the flow-through and incubated at RT for 25 minutes. The DNase-treated RNA was precipitated with 20µL 3M sodium acetate (pH 5.5) and 750µL 100% ethanol overnight at -80℃. The pellet was washed with 750µL ice-cold 80% ethanol and resuspended in a 12:1 DEPC-treated water:RNase OUT Recombinant Ribonuclease Inhibitor (Thermo Fisher, Waltham, MA) mixture on ice for 30 minutes and quantified (Qubit RNA HS assay kit).

Size selection and library prep were conducted exactly as described in Vandivier [14]. The RNA from 15-35bp was cut out of the gel and purified. Small RNA libraries were sequenced in duplicate as single end 150 bp in a multiplexed pool on two lanes of an Illumina HiSeq 2500.

Short read analysis

RNAseq and sRNA Illumina reads were adapter trimmed and mapped against the junction sequences as well as the corresponding plasmid sequences, using Bowtie2 and Geneious aligner (Geneious R10.2.3; custom sensitivity, iterate up to 5 times, zero mismatches or gaps per read, word length 14) at highest stringency. RNAseq reads were mapped against AtACT1 as control, and found this gene expressed. RNAseq reads were mapped against Pol V AT2G40030 as control, and found this gene not expressed. Resulting reads were extracted and re-mapped against the TAIR10 reference genome to ensure that no off-target reads mapped to the plasmid sequences. Because we analyzed various genomic insertions against a single reference, we did not normalize read mapping.

The MethylC-Seq data (paired-end) reads in FastQ format and the second pair reads were converted to their reverse complementary strand nucleotides. Then reads were aligned to the *Arabidopsis thaliana* reference genome (araport 11) and pCSA110/pROK2 genome. The Chloroplast genome was used for quality control (< 0.35% non-conversion rate). After mapping,

overlapping bases in the paired-end reads were trimmed. The base calls per reference position on each strand were used to identify methylated cytosines at 1% FDR.

Data Availability

Raw MinION sequencing data was deposited in the European Nucleotide Archive (ENA) under project PRJEB23977 (ERP105765). Final polished assemblies were deposited in the ENA Genome Assembly Database PRJEB23977. Raw Bionano Genomics molecules and assembled maps are deposited under BioProjects PRJNA387199, PRJNA387199, PRJNA387199 and PRJNA387199. Short-read datasets are deposited under GEO accession GSE108401.

Author contributions

FJ, TPM, JRE conceptualized the study; AR, FJ, STM, MZ, RC, JN and JS created sequencing libraries and/or conducted sequencing; FJ, AR, HC, MZ and TPM analyzed sequencing data; AR, JS and FJ conducted plant studies; FJ, AR, TPM, RKS and JRE wrote the manuscript.

Acknowledgments

We would like to thank Cesar Barragan and Dr. Ronan O'Malley for insights into the T-DNA project, and Bruce Jow and Christopher Santos for excellent greenhouse support. We thank Detlef Weigel and Christa Lanz, both Max Planck Institute for Developmental Biology (Tuebingen, Germany) for performing and providing Illumina short read sequencing for Col-0 CS70000. FJ was supported through a Human Frontier Science Program Organization long-term fellowship. JRE is an Investigator of the Howard Hughes Medical Institute.

Conflict of Interests

The authors declare that they have no conflicting interests relating to this work.

References

[1] Baulcombe, David C., Graham R. Saunders, Michael W. Bevan, Michael A. Mayo, and Bryan D. Harrison. 1986. “Expression of Biologically Active Viral Satellite RNA from the Nuclear Genome of Transformed Plants.” *Nature* 321 (6068): 446–49. doi:10.1038/321446a0.

[2] Caplan, A, L Herrera-Estrella, D Inzé, E Van Haute, M Van Montagu, J Schell, and P Zambryski. 1983. “Introduction of Genetic Material into Plant Cells.” *Science* 222 (4625): 815–21. doi:10.1126/science.222.4625.815.

[3] Fraley, R T, S G Rogers, R B Horsch, P R Sanders, J S Flick, S P Adams, M L Bittner, et al. 1983. “Expression of Bacterial Genes in Plant Cells.” *Proceedings of the National Academy of Sciences of the United States of America* 80 (15): 4803–7.

[4] O’Malley, Ronan C, and Joseph R Ecker. 2010. “Linking Genotype to Phenotype Using the Arabidopsis Unimutant Collection.” *The Plant Journal: For Cell and Molecular Biology* 61 (6): 928–40. doi:10.1111/j.1365-313X.2010.04119.x.

[5] Gelvin, Stanton B. 2003. “Agrobacterium-Mediated Plant Transformation: The Biology Behind the ‘Gene-Jockeying’ Tool.” *Microbiology and Molecular Biology Reviews* 67 (1): 16–37, table of contents.

[6] Zambryski, P, A Depicker, K Kruger, and H M Goodman. 1982. “Tumor Induction by Agrobacterium Tumefaciens: Analysis of the Boundaries of T-DNA.” *Journal of Molecular and Applied Genetics* 1 (4): 361–70.

[7] Nester, Eugene W. 2014. “Agrobacterium: Nature’s Genetic Engineer.” *Frontiers in Plant Science* 5: 730. doi:10.3389/fpls.2014.00730.

[8] Van Kregten, Maartje, Sylvia de Pater, Ron Romeijn, Robin van Schendel, Paul J J Hooykaas, and Marcel Tijsterman. 2016. “T-DNA Integration in Plants Results from Polymerase-θ-Mediated DNA Repair.” *Nature Plants* 2 (11): 16164. doi:10.1038/nplants.2016.164.

[9] Zelensky, A. N., Schimmel, J., Kool, H., Kanaar, R., & Tijsterman, M. (2017). Inactivation of Pol θ and C-NHEJ eliminates off-target integration of exogenous DNA. *Nature Communications*, *8*(1), 66. doi:10.1038/s41467-017-00124-3

[10] Gelvin, S. B. (2017). Integration of Agrobacterium T-DNA into the Plant Genome. *Annual Review of Genetics*, *51*, 195–217. doi:10.1146/annurev-genet-120215-035320

[11] Clark, Katie A, and Patrick J Krysan. 2010. "Chromosomal Translocations Are a Common Phenomenon in Arabidopsis Thaliana T-DNA Insertion Lines." *The Plant Journal: For Cell and Molecular Biology* 64 (6): 990–1001. doi:10.1111/j.1365-313X.2010.04386.x.

[12] Feldmann, Kenneth A. 1991. "T-DNA Insertion Mutagenesis in Arabidopsis: Mutational Spectrum." *The Plant Journal: For Cell and Molecular Biology* 1 (1): 71–82. doi:10.1111/j.1365-313X.1991.00071.x.

[13] Jorgensen, R., Snyder, C., & Jones, J. D. G. (1987). T-DNA is organized predominantly in inverted repeat structures in plants transformed with Agrobacterium tumefaciens C58 derivatives.*Molecular & General Genetics : MGG*, *207*(2-3), 471–477. doi:10.1007/BF00331617

[14] Ming, Ray, Shaobin Hou, Yun Feng, Qingyi Yu, Alexandre Dionne-Laporte, Jimmy H Saw, Pavel Senin, et al. 2008. "The Draft Genome of the Transgenic Tropical Fruit Tree Papaya (Carica Papaya Linnaeus)." *Nature* 452 (7190): 991–96. doi:10.1038/nature06856.

[15] Nacry, P, C Camilleri, B Courtial, M Caboche, and D Bouchez. 1998. "Major Chromosomal Rearrangements Induced by T-DNA Transformation in Arabidopsis." *Genetics* 149 (2): 641–50.

[16] Ooms, G, A Bakker, L Molendijk, G J Wullems, M P Gordon, E W Nester, and R A Schilperoort. 1982. "T-DNA Organization in Homogeneous and Heterogeneous Octopine-Type Crown Gall Tissues of Nicotiana Tabacum." *Cell* 30 (2): 589–97. doi:10.1016/0092-8674(82)90255-0.

[17] Thomashow, M F, R Nutter, A L Montoya, M P Gordon, and E W Nester. 1980. “Integration and Organization of Ti Plasmid Sequences in Crown Gall Tumors.” *Cell* 19 (3): 729–39. doi:10.1016/S0092-8674(80)80049-3.

[18] Ulker, Bekir, Edgar Peiter, David P Dixon, Caroline Moffat, Richard Capper, Nicolas Bouché, Robert Edwards, Dale Sanders, Heather Knight, and Marc R Knight. 2008. “Getting the Most Out of Publicly Available T-DNA Insertion Lines.” *The Plant Journal: For Cell and Molecular Biology* 56 (4): 665–77. doi:10.1111/j.1365-313X.2008.03608.x.

[19] Woody, Scott T, Sandra Austin-Phillips, Richard M Amasino, and Patrick J Krysan. 2007. “The WiscDsLox T-DNA Collection: An Arabidopsis Community Resource Generated by Using an Improved High-Throughput T-DNA Sequencing Pipeline.” *Journal of Plant Research* 120 (1): 157–65. doi:10.1007/s10265-006-0048-x.

[20] Zhu, Qian-Hao, Kerrie Ramm, Andrew L Eamens, Elizabeth S Dennis, and Narayana M Upadhyaya. 2006. “Transgene Structures Suggest That Multiple Mechanisms Are Involved in T-DNA Integration in Plants.” *Plant Science* 171 (3): 308–22. doi:10.1016/j.plantsci.2006.03.019.

[21] Michael, T. P., Jupe, F., Bemm, F., Motley, S. T., Sandoval, J. P., Lanz, C., … Ecker, J. R. (2018). High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature Communications*, *9*(1), 541. doi:10.1038/s41467-018-03016-2.

[22] Alonso, José M, Anna N Stepanova, Thomas J Leisse, Christopher J Kim, Huaming Chen, Paul Shinn, Denise K Stevenson, et al. 2003. “Genome-Wide Insertional Mutagenesis of *Arabidopsis Thaliana*.” *Science* 301 (5633): 653–57. doi:10.1126/science.1086391.

[23] McElver, J, I Tzafrir, G Aux, R Rogers, C Ashby, K Smith, C Thomas, et al. 2001. “Insertional Mutagenesis of Genes Required for Seed Development in Arabidopsis Thaliana.” *Genetics* 159 (4): 1751–63.

[24] Sessions, Allen, Ellen Burke, Gernot Presting, George Aux, John McElver, David Patton, Bob Dietrich, et al. 2002. “A High-Throughput Arabidopsis Reverse Genetics System.” *The Plant Cell* 14 (12): 2985–94.

[25] Vaser, R., Sović, I., Nagarajan, N., & Šikić, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Research*, *27*(5), 737–746. doi:10.1101/gr.214270.116

[26] Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., … Earl, A. M. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *Plos One*, *9*(11), e112963. doi:10.1371/journal.pone.0112963

[27] Kleinboelting, Nils, Gunnar Huep, Ingo Appelhagen, Prisca Viehoever, Yong Li, and Bernd Weisshaar. 2015. “The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism.” *Molecular Plant* 8 (11): 1651–64. doi:10.1016/j.molp.2015.08.011.

[28] Daxinger, L., Hunter, B., Sheikh, M., Jauvion, V., Gasciolli, V., Vaucheret, H., … Furner, I. (2008). Unexpected silencing effects from T-DNA tags in Arabidopsis. *Trends in Plant Science*, *13*(1), 4–6. doi:10.1016/j.tplants.2007.10.007

[29] Gao, Y., & Zhao, Y. (2013). Epigenetic suppression of T-DNA insertion mutants in Arabidopsis. *Molecular Plant*, *6*(2), 539–545. doi:10.1093/mp/sss093

[30] LeClere, S., & Bartel, B. (2001). A library of Arabidopsis 35S-cDNA lines for identifying novel mutants. *Plant Molecular Biology*, *46*(6), 695–703.

[31] Nakamura, S., Mano, S., Tanaka, Y., Ohnishi, M., Nakamori, C., Araki, M., … Ishiguro, S. (2010). Gateway binary vectors with the bialaphos resistance gene, bar, as a selection marker for plant transformation. *Bioscience, Biotechnology, and Biochemistry*, *74*(6), 1315–1319. doi:10.1271/bbb.100184

[32] Law, J. A., & Jacobsen, S. E. (2010). Establishing, maintaining and modifying DNA methylation patterns in plants and animals. *Nature Reviews. Genetics*, *11*(3), 204–220. doi:10.1038/nrg2719

[33] Fritsch, O., Benvenuto, G., Bowler, C., Molinier, J., & Hohn, B. (2004). The INO80 protein controls homologous recombination in Arabidopsis thaliana. *Molecular Cell*, *16*(3), 479–485. doi:10.1016/j.molcel.2004.09.034

[34] Coleman-Derr, D., & Zilberman, D. (2012). Deposition of histone variant H2A.Z within gene bodies regulates responsive genes. *PLoS Genetics*, *8*(10), e1002988. doi:10.1371/journal.pgen.1002988

[35] Peach, C., & Velten, J. (1991). Transgene expression variability (position effect) of CAT and GUS reporter genes driven by linked divergent T-DNA promoters. *Plant Molecular Biology*,*17*(1), 49–60. doi:10.1007/BF00036805

[36] Schouten, H. J., Vande Geest, H., Papadimitriou, S., Bemer, M., Schaart, J. G., Smulders, M. J. M., … Schijlen, E. (2017). Re-sequencing transgenic plants revealed rearrangements at T-DNA inserts, and integration of a short T-DNA fragment, but no increase of small mutations elsewhere. *Plant Cell Reports*, *36*(3), 493–504. doi:10.1007/s00299-017-2098-z

[37] Arabidopsis Genome Initiative. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. *Nature*, *408*(6814), 796–815. doi:10.1038/35048692

[38] Dürrenberger, F., Crameri, A., Hohn, B., & Koukolíková-Nicola, Z. (1989). Covalently bound VirD2 protein of Agrobacterium tumefaciens protects the T-DNA from exonucleolytic degradation.*Proceedings of the National Academy of Sciences of the United States of America*, *86*(23), 9154–9158.

[39] Schuermann, D., Fritsch, O., Lucht, J. M., & Hohn, B. (2009). Replication stress leads to genome instabilities in Arabidopsis DNA polymerase delta mutants. *The Plant Cell*, *21*(9), 2700–2714. doi:10.1105/tpc.109.069682

[40] Simpson, J. T., Workman, R. E., Zuzarte, P. C., David, M., Dursi, L. J., & Timp, W. (2017). Detecting DNA cytosine methylation using nanopore sequencing. *Nature Methods*, *14*(4), 407–410. doi:10.1038/nmeth.4184

[41] Sidorenko, L. V., Lee, T.-F., Woosley, A., Moskal, W. A., Bevan, S. A., Merlo, P. A. O., … Meyers, B. C. (2017). GC-rich coding sequences reduce transposon-like, small RNA-mediated transgene silencing. *Nature Plants*. doi:10.1038/s41477-017-0040-6

[42] Shilo, S., Tripathi, P., Melamed-Bessudo, C., Tzfadia, O., Muth, T. R., & Levy, A. A. (2017). T-DNA-genome junctions form early after infection and are influenced by the chromatin state of the host genome. *PLoS Genetics*, *13*(7), e1006875. doi:10.1371/journal.pgen.1006875

[43] O'Malley, R. C., Barragan, C. C., & Ecker, J. R. (2015). A user's guide to the Arabidopsis T-DNA insertion mutant collections. *Methods in Molecular Biology*, *1284*, 323–342. doi:10.1007/978-1-4939-2444-8_16

Methods References

[1] LeClere, S., & Bartel, B. (2001). A library of Arabidopsis 35S-cDNA lines for identifying novel mutants. *Plant Molecular Biology*, *46*(6), 695–703.

[2] Nakamura, S., Mano, S., Tanaka, Y., Ohnishi, M., Nakamori, C., Araki, M., … Ishiguro, S. (2010). Gateway binary vectors with the bialaphos resistance gene, bar, as a selection marker for plant transformation. *Bioscience, Biotechnology, and Biochemistry*, *74*(6), 1315–1319. doi:10.1271/bbb.100184

[3] Loman, Nicholas J, and Aaron R Quinlan. 2014. "Poretools: a Toolkit for Analyzing Nanopore Sequence Data." *Bioinformatics* 30 (23): 3399–3401. doi:10.1093/bioinformatics/btu555.

[4] Li, Heng. 2016. "Minimap and Miniasm: Fast Mapping and de Novo Assembly for Noisy Long Sequences." *Bioinformatics* 32 (14): 2103–10. doi:10.1093/bioinformatics/btw152.

[5] Vaser, Robert, Ivan Sovic, Niranjan Nagarajan, and Mile Sikic. 2016. “Fast and Accurate de Novo Genome Assembly from Long Uncorrected Reads.” *BioRxiv*, August. doi:10.1101/068122.

[6] Li, Heng, and Richard Durbin. 2009. “Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform.” *Bioinformatics* 25 (14): 1754–60. doi:10.1093/bioinformatics/btp324.

[7] Walker, Bruce J, Thomas Abeel, Terrance Shea, Margaret Priest, Amr Abouelliel, Sharadha Sakthikumar, Christina A Cuomo, et al. 2014. “Pilon: An Integrated Tool for Comprehensive Microbial Variant Detection and Genome Assembly Improvement.” *Plos One* 9 (11): e112963. doi:10.1371/journal.pone.0112963.

[8] Gurevich, Alexey, Vladislav Saveliev, Nikolay Vyahhi, and Glenn Tesler. 2013. “QUAST: Quality Assessment Tool for Genome Assemblies.” *Bioinformatics* 29 (8): 1072–75. doi:10.1093/bioinformatics/btt086.

[9] Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., & Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research*, *27*(5), 722–736. doi:10.1101/gr.215087.116

[10] Kearse, Matthew, Richard Moir, Amy Wilson, Steven Stones-Havas, Matthew Cheung, Shane Sturrock, Simon Buxton, et al. 2012. “Geneious Basic: An Integrated and Extendable Desktop Software Platform for the Organization and Analysis of Sequence Data.” *Bioinformatics* 28 (12): 1647–49. doi:10.1093/bioinformatics/bts199.

[11] Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*, *215*(3), 403–410. doi:10.1016/S0022-2836(05)80360-2

[12] Kawakatsu, Taiji, Shao-Shan Carol Huang, Florian Jupe, Eriko Sasaki, Robert J Schmitz, Mark A Urich, Rosa Castanon, et al. 2016. “Epigenomic Diversity in a Global Collection of Arabidopsis Thaliana Accessions.” *Cell* 166 (2): 492–505. doi:10.1016/j.cell.2016.06.044.

[13] Lam, Hugo Y K, Michael J Clark, Rui Chen, Rong Chen, Georges Natsoulis, Maeve O'Huallachain, Frederick E Dewey, et al. 2011. "Performance Comparison of Whole-Genome Sequencing Platforms." *Nature Biotechnology* 30 (1): 78–82. doi:10.1038/nbt.2065.

[14] Vandivier, L. E., Li, F., & Gregory, B. D. (2015). High-throughput nuclease-mediated probing of RNA secondary structure in plant transcriptomes. *Methods in Molecular Biology*, *1284*, 41–70. doi:10.1007/978-1-4939-2444-8_3

Supplementary Table 1: Bionano Genomics and Oxford Nanopore sequencing and assembly statistics.

Supplementary Table 2: Genome Alignment of Oxford Nanopore assembled contigs against TAIR10 using the Bionano Genomics RefAligner algorithm.

Supplementary Table 3: Identification of 'N' regions in the TAIR10 reference genome and the corresponding Col-0 ONT contigs.

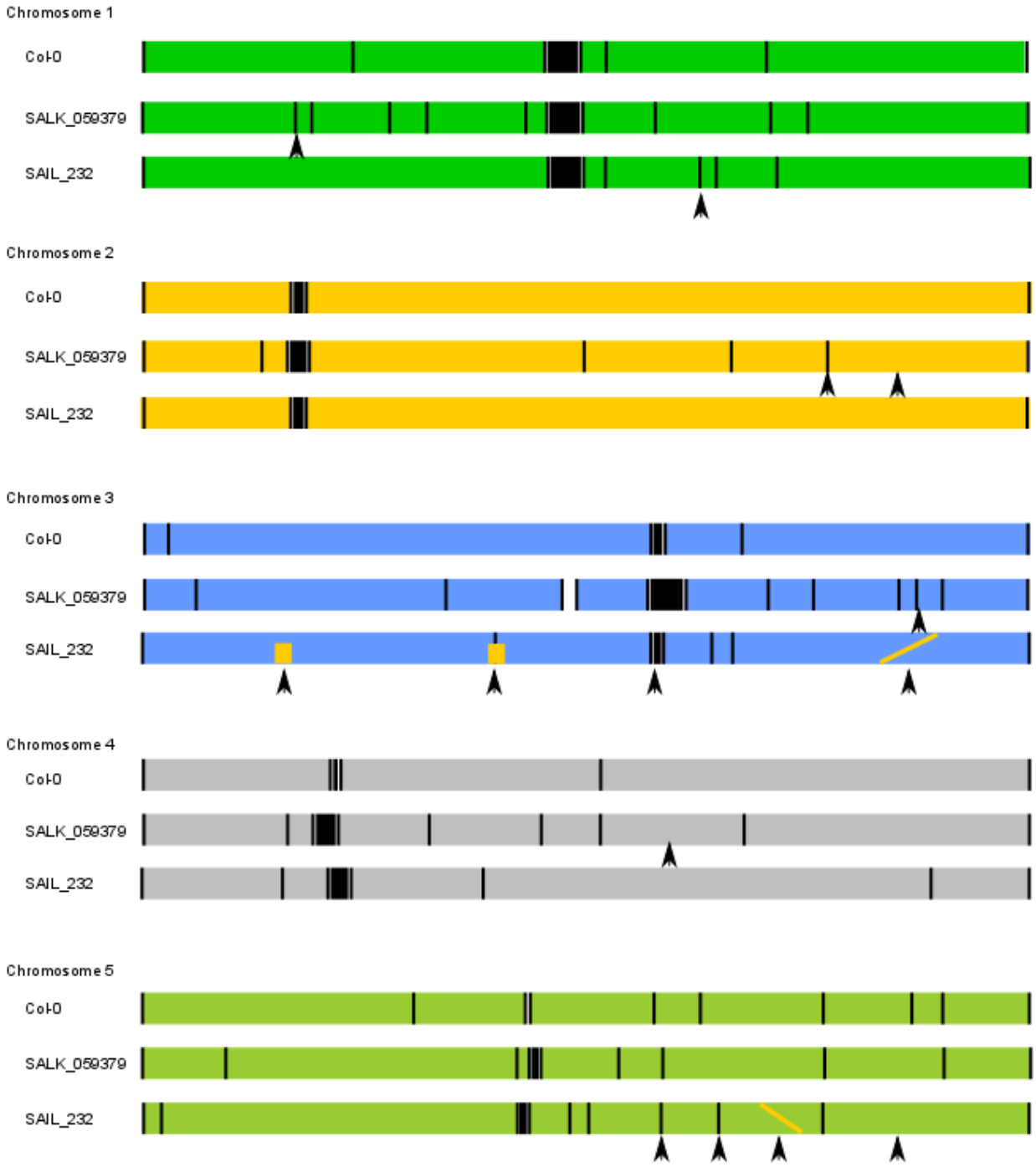
Supplementary Table 4: Optical maps identify misassembled regions in ONT contigs as 'False Duplications' or 'False Deletions'.

Supplementary Table 5: List and analysis of genomic T-DNA insertion sites.

Supplementary Table 6: Analysis of methylated cytosines on the pCSA110 plasmid for SAIL_232.

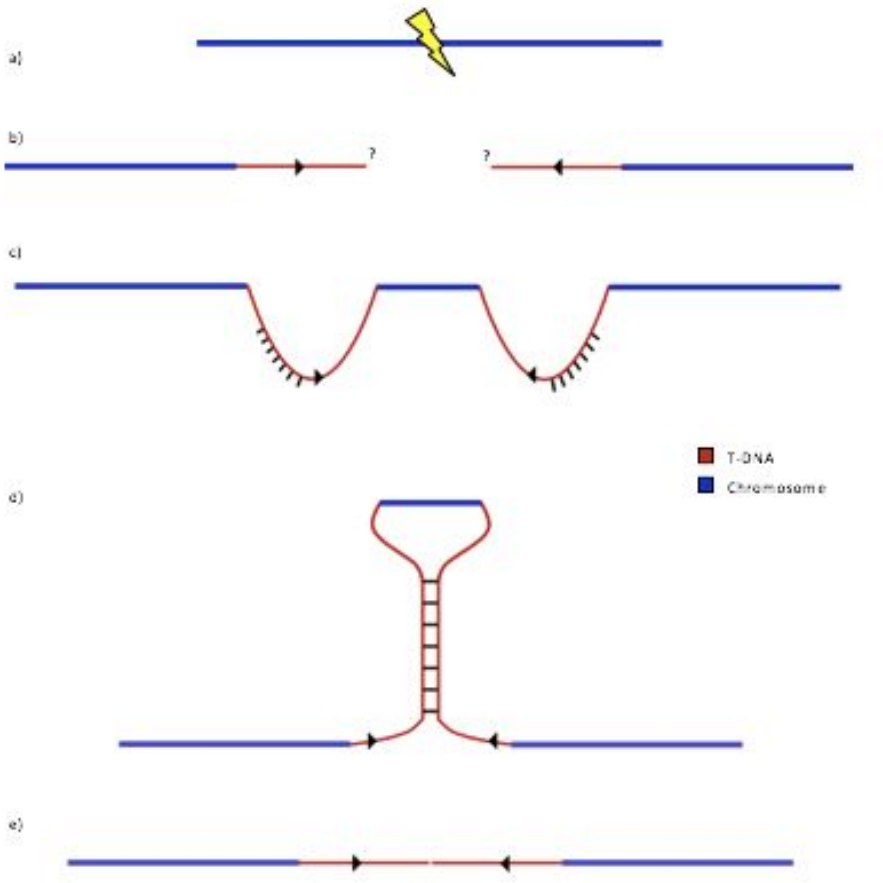
Supplementary Table 7: Analysis of methylated cytosines on the pROK2 plasmid for SALK_059379.

Methods Table 1: PCR oligo sequences used for insert genotyping

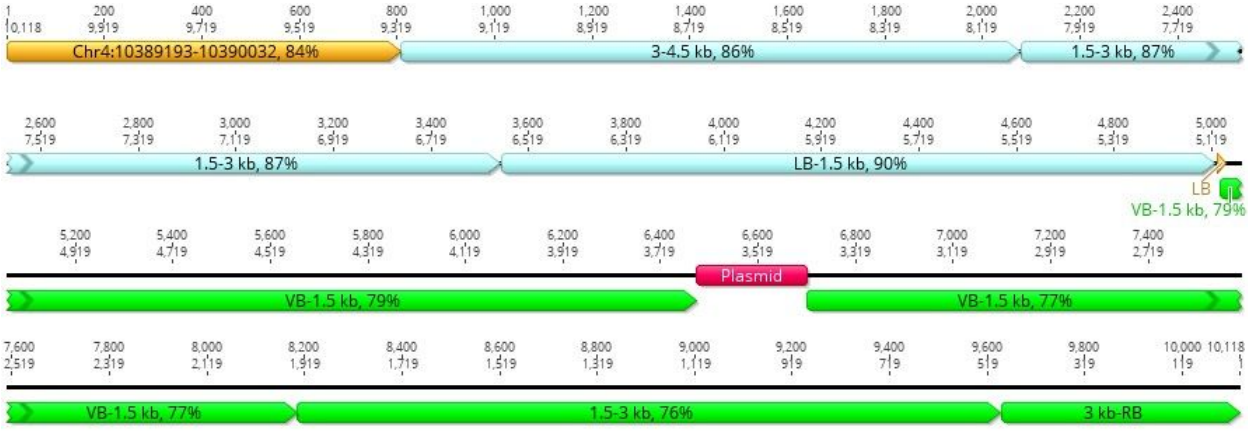


Supplementary Fig. 1: Alignment of *de novo* assembled ONT contigs for Col-0 (CS70000) and the transgenic lines SALK_059379 and SAIL_232 relative to the reference TAIR10. Uninterrupted colored

blocks indicate contig length, black bars indicate the start and end of contigs, black boxes indicate centromere gaps. Arrowheads represent T-DNA insertion sites; orange lines and boxes indicate sites of translocations. Drawn to scale for each chromosome individually.



Supplementary Fig. 2: Model for microhomology dependent excision of T-DNA/genome fragments. (a) A double strand chromosome break leads to (b) two multi-copy T-DNA strand insertions in opposing directions. The exact structure and length of the inserted sequence is unknown, as indicated by question marks. (c) SALK_chr2:18Mb insertion features two individual double strand breaks, around 5 kb apart. High homology between the T-DNA strands as well as the hairpin forming original DNA piece created a secondary structure (d), that was potentially excised (e) and resulted in the deletion of the ~5kb chromosomal fragment, as shown in main text Fig. 2. Arrowheads on the red T-DNA strand show direction. The blue line represents double stranded DNA.



Supplementary Fig. 3: ONT reads identify insertions with lower allele frequency that are not part of assembled contigs. We applied blastn searches of all unassembled ONT reads against the utilized vector sequence, and subsequently the TAIR10 reference genome. This identified reads such as the depicted, that confirm insertion events (like SALK_059379 on Chr 4 at 10.4 Mb) not present in the assembly or optical maps. Our blastn strategy identified chromosomal sequence (yellow), and individual alignments with the pROK plasmid sequence revealed T-strand (blue) and vector backbone (green) sequence. The pink plasmid sequence within the vector backbone shows an internal breakpoint, which was likely caused by multiple independent insertion events within the same region. Percent identity of the raw read stretch to the reference sequence are listed within the annotations.