Large scale genomic rearrangements in selected *Arabidopsis thaliana* T-DNA lines are caused by T-DNA insertion mutagenesis

Boas Pucker[1,2+], Nils Kleinbölting[3+], and Bernd Weisshaar[1*]

1 Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University, Sequenz 1, 33615 Bielefeld, Germany

2 Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge, United Kingdom

3 Bioinformatics Resource Facility, Center for Biotechnology (CeBiTec, Bielefeld University, Sequenz 1, 33615 Bielefeld, Germany

+ authors contributed equally

* corresponding author: Bernd Weisshaar

BP: bpucker@cebitec.uni-bielefeld.de, ORCID: 0000-0002-3321-7471
NK: nkleinbo@cebitec.uni-bielefeld.de, ORCID: 0000-0001-9124-5203
BW: bernd.weisshaar@uni-bielefeld.de, ORCID: 0000-0002-7635-3473

Abstract

Background

Experimental proof of gene function assignments in plants is heavily based on mutant analyses. T-DNA insertion lines provided an invaluable resource of mutants and enabled systematic reverse genetics-based investigation of the functions of *Arabidopsis thaliana* genes during the last decades.

Results

We sequenced the genomes of 14 *A. thaliana* GABI-Kat T-DNA insertion lines, which eluded flanking sequence tag-based attempts to characterize their insertion loci, with Oxford Nanopore Technologies (ONT) long reads. Complex T-DNA insertions were resolved and 11 previously unknown T-DNA loci identified, suggesting that the number of T-DNA insertions per line was underestimated. T-DNA mutagenesis caused fusions of chromosomes along with compensating translocations to keep the gene set complete throughout meiosis. Also, an inverted duplication of 800 kbp was detected. About 10% of GABI-Kat lines might be affected by chromosomal rearrangements, some of which do not involve T-DNA. Local assembly of selected reads was shown to be a computationally effective method to resolve the structure of T-DNA insertion loci. We developed an automated workflow to support investigation of long read data from T-DNA insertion lines. All steps from DNA extraction to assembly of T-DNA loci can be completed within days.

Conclusion

Long read sequencing was demonstrated to be a very effective way to resolve complex T-DNA insertions and chromosome fusions. Many T-DNA insertions comprise not just a single T-DNA, but complex arrays of multiple T-DNAs. It is becoming obvious that T-DNA insertion alleles must be characterized by exact identification of both T-DNA::genome junctions to generate clear genotype-to-phenotype relations.

Keywords: long read sequencing, genome assembly, structural variants, translocations, chromosome fusions, reverse genetics, chromosomal rearrangements, GABI-Kat

Background

T-DNA insertion lines contributed substantially to the high-value knowledge about the functions of plant genes that has been produced by the plant research community on the basis of gene structures predicted from genome sequences. In addition to the application of T-DNA as activation tags to cause overexpression of flanking genes, T-DNA insertions turned out as an effective mechanism for the generation of knock-out alleles for use in reverse genetics and targeted gene function search [1, 2]. Since targeted integration of DNA into plant genomes via homologous recombination was difficult or at least technically very challenging [3], large collections of sequence-indexed T-DNA integration lines with random insertion sites were used to provide knock-out alleles for the majority of genes [4]. Knowledge about the inserted sequences is an advantage over other mutagenesis methods, because localization of the insertion within the mutagenized genome based on the generation of flanking sequence tags (FSTs) is possible [5, 6]. While the CRISPR/Cas technology now offers technically feasible alternatives for access to mutant alleles for reverse genetics [7], thousands of T-DNA insertion mutants have been characterized and represent today the main or reference mutant allele for (lack of) a given gene function.

Agrobacterium tumefaciens is a Gram-negative soil bacterium with the ability to transfer DNA into plant cells and to integrate this T-DNA stably and at random positions into the nuclear genome [8, 9]. A specific tumor inducing (Ti) plasmid, that is naturally occurring in *Agrobacteria* and that enables them to induce the formation of crown galls in plants, contains the T-DNA which is transferred into plant cells [10]. The T-DNA is enclosed by 25 bp long imperfect repeats that were designated left (LB) and right border (RB) [9]. The T-DNA sequence between LB and RB can be modified to contain resistance genes for selection of successfully transformed plants [11]. T-DNAs

from optimized binary plasmids are transformed into *A. thaliana* plants via floral dip to generate stable lines [12]. T-DNA transfer into the nucleus of a plant cell is supported by several VIR proteins which are, in the biotechnologically optimized system, encoded on a separate helper plasmid. It is assumed that host proteins are responsible for integration of the T-DNA into the genome, most likely as a DNA double strand into a double strand break (DSB) using host DNA repair pathways and DNA polymerase theta [9, 13, 14]. T-DNA integration resembles DNA break repair through non-homologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ) and is often accompanied by the presence of filler DNA or microhomology at both T-DNA::genome junctions [9, 13]. Chromosomal inversions and translocations are commonly associated with T-DNA insertions [15-19], suggesting that often more than just one DSB is associated with T-DNA integration [9].

The most important collections of T-DNA lines for the model plant *Arabidopsis thaliana* are SALK (150,000 lines) [6], GABI-Kat (92,000 lines) [20, 21], SAIL (54,000 lines) [22], and WISC (60,000 lines) [23]. In total, over 700,000 insertion lines have been constructed [4]. GABI-Kat lines were generated through the integration of a T-DNA harboring a sulfadiazine resistance gene for selection of transformed lines [20]. Additionally, the T-DNA contains a 35S promoter at RB causing transcriptional up-regulation of genes next to the integration site if the right part of the T-DNA next to RB stays intact during integration [1]. Integration sites were predicted based on FSTs and allowed access to knock-out alleles of numerous genes. At GABI-Kat, T-DNA insertion alleles were confirmed by an additional "confirmation PCR" using DNA from the T2 generation [24] prior to the release of a mutant line and submission of the line to the Nottingham Arabidopsis Stock Centre (NASC). Researchers could identify suitable and available T-DNA insertion lines via SimpleSearch on the GABI-Kat website [25]. Since 2017, SimpleSearch uses Araport11 annotation data [26]. Araport11 is based on the *A. thaliana* Col-0 reference genome sequence from TAIR9 which includes about 96 annotated gaps filled with Ns [27], among them the centromers and several gaps in the pericentromeric regions.

The prediction of integration sites based on bioinformatic evaluations using FST data does often not reveal the complete picture. Insertions might be masked from FST predictions due to truncated

borders [13], because of repetitive sequences or paralogous regions in the genome [28], or even lack of the true insertion site in the reference sequence used for FST mapping [29, 30]. Also, confirmation by sequencing an amplicon that spans the predicted insertion site at one of the two expected T-DNA::genome junctions is not fully informative. Deletions and target site duplications at the integration site can occur and can only be detected by examining both borders of the inserted T-DNA [13]. In addition, more complex insertions have been reported by several studies that include large deletions, insertions, inversions or even chromosomal translocations [13, 18, 31-34]. Also, binary vector backbone (BVB) sequences have been detected at insertion sites [35] as well as fragments of *A. tumefaciens* chromosomal DNA [36]. In addition, recombination between two T-DNA loci was described as a mechanism for deletion of an enclosed genomic fragment [37].

Plant genomes are dynamic and often show whole genome doubling followed by purging processes [38, 39]. Transposable elements (TE) play an important role in restructuring genomes [38], but chromosomal rearrangement events not involving TEs also lead to large structural variation [33, 40, 41]. The karyotype of *A. thaliana* is the result of chromosome fusion events which reduced the chromosome number from the ancestral eight to five [40]. Recent advances in long read sequencing pave the way for comprehensive synteny analyses with Brassica species related to *A. thaliana*. A recent study reported 13-17 Mbp of rearranged sequence between pairs of geographically diverse *A. thaliana* accessions [42]. Also, structural variants have the potential to contribute to speciation [39]. Chromosomal rearrangements can occur during the repair of DSBs, e.g. via microhomology-mediated end joining or non-allelic homologous recombination [reviewed by 43, 44]. Evidently, regions with high sequence similarity like duplications are especially prone to chromosomal rearrangements [43].

While usually one T-DNA locus per line was identified by FSTs, the number of T-DNA insertion loci per line is usually higher. For GABI-Kat, it was estimated that about 50% of all lines (12,018 of 21,049 tested, according to numbers from the end of 2019) display a single insertion locus. This estimation is based on segregation analyses using sulfadiazine resistance as a selection marker [20]. Other insertion mutant collections report similar numbers [4]. The average number of T-DNA

insertions per line was reported to be about 1.5, but this is probably a significant underestimation since the kanamycin and BASTA selection marker genes applied to determine the numbers are known to be silenced quite often [4]. For these reasons, it is required that insertion mutants (similar to mutants created by e.g. chemical mutagenesis) are backcrossed to wild type prior to phenotyping a homozygous line.

The FSTs produced for the different mutant populations by individual PCR and Sanger-sequencing allowed usually access to a single T-DNA insertion locus per line, although for GABI-Kat there are several examples with up to three confirmed insertion loci based on FST data (e.g. line GK-011F01, see [25]). This leaves a significant potential of undiscovered T-DNA insertions in lines already available at the stock centers, which has been exploited by the group of Joe Ecker by applying TDNA-Seq (Illumina technology) to the SALK and a part of the GABI-Kat mutant populations. Essentially the same technology has later also been used to set up a sequence indexed insertion mutant library of *Chlamydomonas reinhardtii* [45]. With the fast development of new DNA sequencing technologies, the comprehensive characterization of T-DNA insertion lines comes into reach.

Several studies already harnessed high throughput sequencing technologies to investigate T-DNA insertion and other mutant lines [46, 47]. Oxford Nanopore Technologies (ONT) provides a cost-effective and fast approach to study *A. thaliana* genomes, since a single MinION/GridION Flow Cell delivers sufficient data to assemble one genotype [48]. Here, we present a method to fully characterize T-DNA insertion loci and additional genomic changes of T-DNA insertion lines through ONT long read sequencing. We selected 14 lines that contain confirmed T-DNA insertion alleles (first border or first T-DNA::genome junction confirmed by sequencing an amplicon across one junction), but which escaped characterization of the second T-DNA::genome junction (we refer to the T-DNA::genome junction that is expected to exist after confirmation of one T-DNA::genome junction as "2nd border"). Within this biased set of lines, we detected several chromosome fragment or chromosome arm translocations, a duplication of 800 kbp and also an insertion of DNA from the chloroplast (plastome), all related to T-DNA insertion events. The results clearly demonstrate the

166 importance of characterizing both T-DNA::genome junctions for reliable selection of suitable alleles
167 for setting up genotype/phenotype relations for gene function search. In parallel to data evaluation,
168 we created an automated workflow to support long-read-based analyses of T-DNA insertion lines
169 and alleles.
170
171

172 ## Results

173 In total, 14 GABI-Kat T-DNA insertion lines (Table 1, Additional file 1) were selected for genomic
174 analysis via ONT long read sequencing. This set of lines was selected based on prior knowledge
175 which indicated that the insertion locus addressed in the respective line was potentially somehow
176 unusual. The specific feature used for selection was the (negative) observation that creation of
177 confirmation amplicons which span the T-DNA::genome junction failed for one of the two junctions,
178 operationally that means that the 2nd border could not be confirmed. T-DNA insertion loci in the
179 selected lines were assessed by *de novo* assembly of the 14 individual genome sequences, and by
180 a computationally more effective local assembly of selected reads.
181
182

183 Table 1: Key findings summary of ONT-sequenced GABI-Kat T-DNA insertion lines.

Line ID[a]	**number of insertions**				**Summary of observation**
	FST pred.	PCR-conf.	total found	ONT found	
GK-038B07	2	0 (1)[b]	4[c]	3	FST-predicted insertion in Chr5 is part of a fusion of Chr3 and Chr5, 2 T-DNA arrays detected at translocation fusion points of which one contains in addition an inversion of ~2 Mbp fused with another T-DNA array, additional insertion of a complex T-DNA array in Chr1.
GK-089D12	1	1	2	1	Fusion of Chr3 and Chr5, FSTs are derived from the single DNA::genome junction that contains LB, both translocation fusions contain mostly canonical T-DNAs, failure to confirm 2nd T-DNA::genome junction explained by shortened T-DNA.

GK-430F05	1	1	3	2	Three T-DNA insertion sites, two complex T-DNA arrays with one containing more than 5 kbp BVB[d], the confirmed T-DNA insertion in Chr3 is complex, no theoretical explanation for failure of confirmation PCR at 2nd DNA::genome junction, one insertion in the pericentromeric region of (probably) Chr4 containing a rearranged RB region.
GK-654A12[d]	1	1	2[c]	1	Fusion of Chr1 and Chr4 with a complex T-DNA array that also contains BVB, Chr1-part of the fusion predicted by FST, the compensating fusion of Chr4 and Chr1 does not contain T-DNA at the translocation fusion point, translocation explains failure to confirm 2nd T-DNA::genome junction, additional insertion of a T-DNA array in Chr2 with a 162 bp duplicated inversion at the integration site.
GK-767D12	1	1	2	1	Large segmental duplication and inversion at the predicted insertion site, a long T-DNA array also containing BVB present at the southern end of the segmental duplication, inversion explains failure to confirm 2nd T-DNA::genome junction, the northern fusion point of the inverted segmental duplication does not contain T-DNA, another canonical T-DNA insertion in Chr2 but with 4 Mbp distance.
GK-909H04	1	1	3	2	The predicted T-DNA insertion contains an inverted duplication of 20 kbp at the integration site explains why the 2nd T-DNA::genome junction could not be confirmed, integration of 652 bp derived from the plastome at the northern fusion point of the duplication but no T-DNA, two additional canonical T-DNA insertions in Chr1 and Chr2.
GK-947B06	1	1	2	1	Complex T-DNA array insertion of 3 T-DNAs at the predicted insertion site in Chr1, no theoretical explanation for failure of the confirmation PCR at the 2nd DNA::genome junction, additional complex insertion containing 2 T-DNAs and BVB on Chr2.

[a] Lines with newly detected insertions, all lines are listed in Additional file 2.
[b] The T-DNA insertion used for line selection was a false positive case not detected by ONT seq
[c] One locus in a line with a chromosome translocation cause the FSTs from this single locus to map to two places in the reference sequence. These two places might also be the source of potentially existing FST from the compensating chromosome fusion. If the compensating chromosome fusion does not contain a T-DNA, the number of insertions is lower than expected.
[d] BVB, binary vector backbone

A tool designated "loreta" (long read-based t-DNA analysis) has been developed during the analyses and might be helpful for similar studies (see methods for details). The results of both approaches demonstrate that a full *de novo* assembly is not always required if only certain regions in the genome are of interest. The 14 GABI-Kat lines harbor a total of 26 T-DNA insertions resulting in an average of 1.86 insertions per line. A total of 11 insertion loci detected in seven of 14 lines were not revealed by previous attempts to detect T-DNA insertions that were based on FSTs (Table

1, Additional file 2 and 3). In case of GK-038B07, the lack of re-detection of the expected insertion allele of At4g19510 was explained by a PCR template contamination during the initial confirmation, the line that contains the real insertion (source of the contamination) is most probably GK-159D11. A similar explanation is true in case of GK-040A12 where the expected insertion allele of At1g52720 was also not found in the ONT data. At least, the error detected fits to the selection criteria, because the 2nd border or 2nd T-DNA::genome junction can obviously not be detected if the T-DNA insertion as such is not present in the line.

In six of the 14 lines studied by ONT whole genome sequencing, chromosomal rearrangements were found. To visualize these results, we created ideograms of the five *A. thaliana* chromosomes with a color code for each of the chromosomes. The colors allow to visually perceive information on chromosome arm translocations, and the changeover points indicate presence or absence of T-DNA sequences.

Chromosome fusions

In four lines, fusions of different chromosomes were detected. These fusions result from chromosome arm translocations which were, in all four cases, compensated within the line by reciprocal translocations. The T-DNA insertion on chromosome 5 (Chr5) of GK-038B07 is part of a complex chromosome arm translocation (Fig. 1). A part of Chr5 is fused to Chr3, the replaced part of Chr3 is fused to an inversion of 2 Mbp on Chr5. This inversion contains T-DNAs at both ends, one of which is the insertion predicted by FSTs.

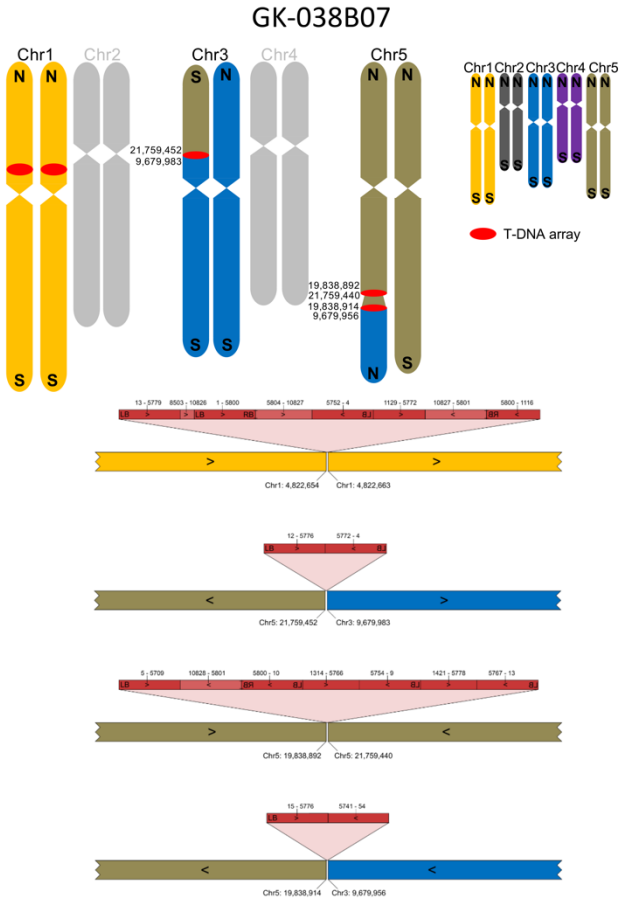
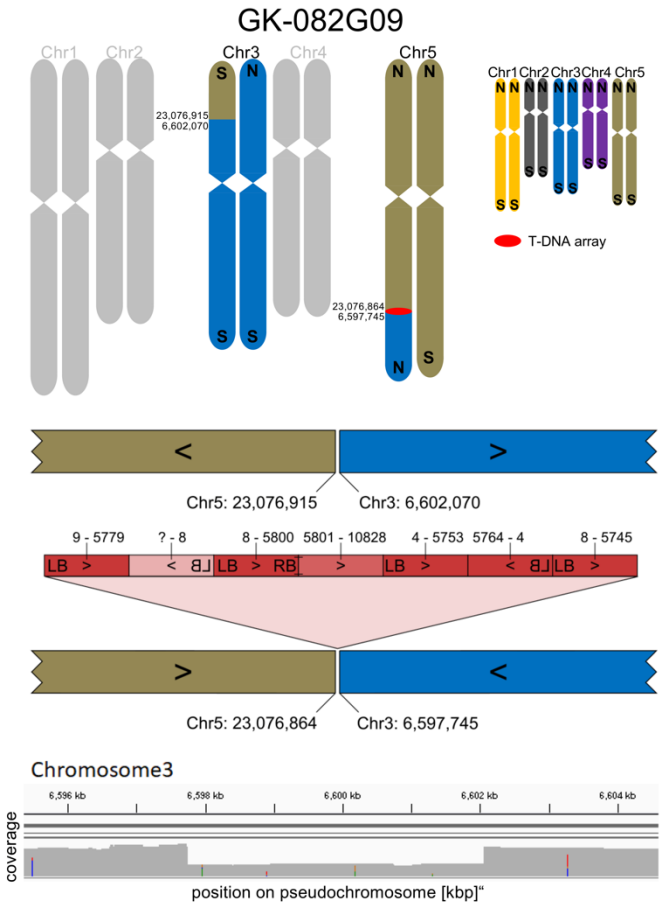


Fig. 1: Structure of the nuclear genome of GK-038B07 with a focus on translocations, inversions and T-DNA structures. Upper right: color codes used for the five chromosomes; N, northern end of chromosome; S, southern end of chromosome. Upper left: ideograms of the chromosomes that display the reciprocal fusion of Chr3 and Chr5 as well as a 2 Mbp inversion between two T-DNA arrays at the fusion sites; numbers indicate end points of pseudochromosome fragments according to TAIR9. Lower part: visualization of the four T-DNA insertion loci of GK-038B07 resolved by local assembly. LB and RB, T-DNA left and right border; dark red, *bona fide* T-DNA sequences located between the borders; light red, sequence parts from the binary vector backbone (BVB); numbers above the red bar indicate nucleotide positions with position 1 placed at the left end of LB in the binary vector which makes position 4 the start of the transferred DNA [13]; numbers below the colored bars indicate pseudochromosome positions according to TAIR9.

For line GK-082G09, two FST predictions had been generated and one FST lead to the prediction of an insertion at Chr3:6,597,745 which was confirmed by PCR. Confirmation of the expected corresponding 2nd border failed. Another FST-based prediction at Chr5:23,076,864 was not addressed by PCR. ONT sequencing confirmed both predictions (Fig. 2). There is only one

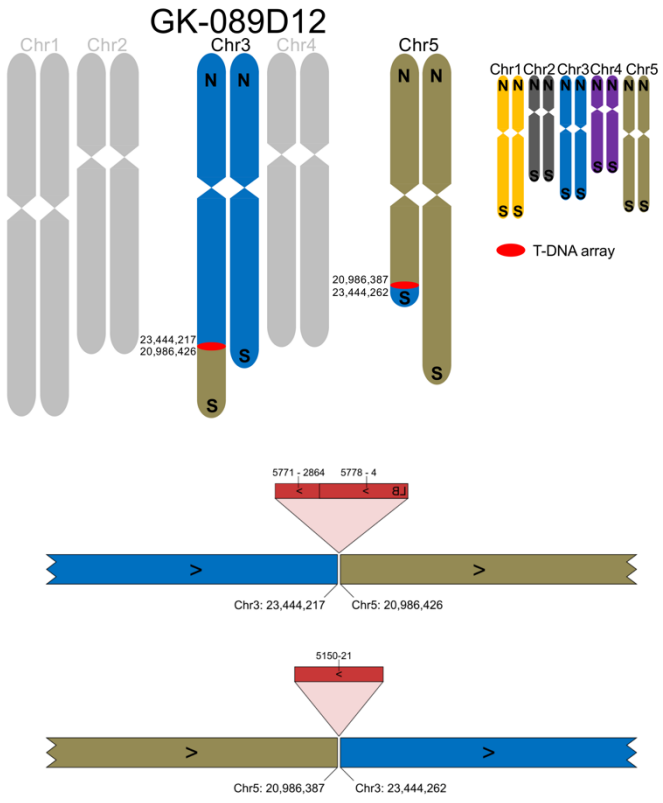
237 complex insertion consisting of multiple T-DNA copies and BVB in GK-082G09 that fuses the south
238 of Chr5 to an about 6.6 Mbp long fragment from the north of Chr3. This translocation is
239 compensated by a fusion of the corresponding parts of both chromosomes without a T-DNA. The
240 second fusion point of Chr3 and Chr5, that was detected in the *de novo* assembly of the genome
241 sequence of GK-082G09, was validated by generating and sequencing a PCR amplicon spanning
242 the translocation fusion point (see Additional file 1 for sequences/accession numbers and
243 Additional file 4 for the primer sequences).
244



245

246 **Fig. 2**: Structure of the nuclear genome of GK-082G09 with a focus on translocations, inversions
247 and T-DNA structures. For a description of the figure elements see legend to Fig. 1. Bottom: read
248 coverage depth analyses of the region of Chr3 that is involved in the fusions which confirms a
249 deletion of about 4 kbp from Chr3. The reads that cover the deleted part were derived from the wild
250 type allele present in the segregating population (see methods).
251

252 Line GK-089D12 harbors two T-DNA insertions (Fig. 3) and both were predicted by FSTs, one in
253 Chr3 and one in Chr5. Since fragments of Chr3 and Chr5 are exchanged in a reciprocal way with
254 no change in sequence direction (southern telomeres stay at the southern ends of the
255 chromosomes), PCR confirmation would have usually resulted in "fully confirmed" insertion alleles.
256 Only long read sequencing allowed to determine the involvement of translocations. The line was
257 studied because the shortened T-DNA at 089D12-At5g51660-At3g63490 (see Additional file 3 for
258 designations of insertions) caused failure of formation of the confirmation amplicon.
259

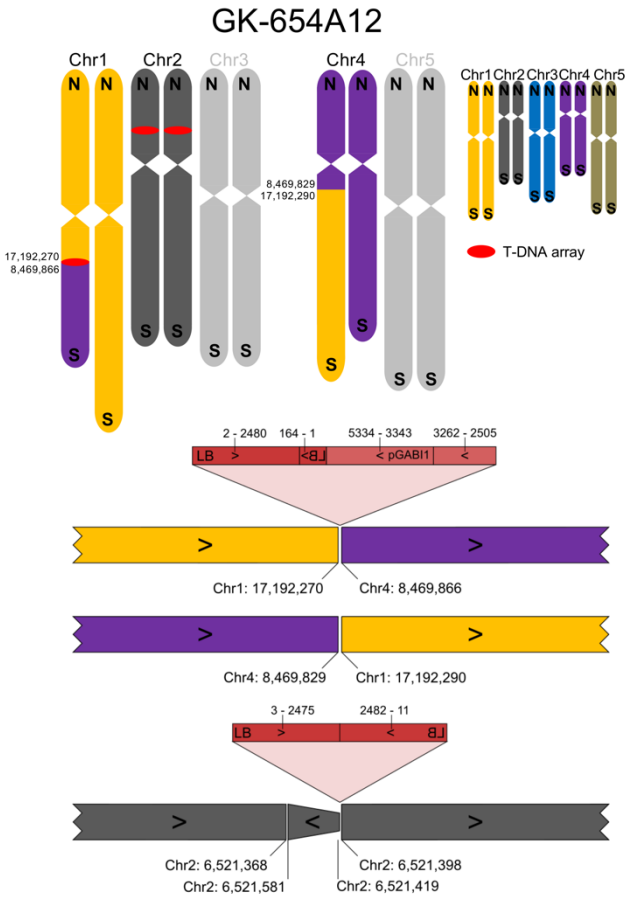


260

261 **Fig. 3**: Structure of the nuclear genome of GK-089D12 with a focus on translocations, inversions
262 and T-DNA structures. For a description of the figure elements see legend to Fig. 1.
263

264 FSTs from line GK-654A12 indicated a T-DNA insertion on Chr1. ONT sequencing revealed a
265 translocation between Chr1 and Chr4 that explained failure to generate the confirmation amplicon
266 at the 2nd border (Fig. 4). The southern arms of Chr1 and Chr4 are exchanged, with a T-DNA array
267 inserted at the fusion point of the new chromosome that contains CEN1 (centromere of Chr1). The

268 fusion point of the new chromosome that contains CEN4 does not contain T-DNA sequences. Also
269 this T-DNA-free fusion point (654A12-FCAALL-0-At1g45688) was validated by generating and
270 sequencing a PCR amplicon which spanned the fusion site (Additional files 1 and 4). The T-DNA
271 array at 654A12-At1g45688-FCAALL contains BVB sequences, interestingly as an independent
272 fragment and not in an arrangement that is similar to the binary plasmid construction which
273 provided the T-DNA.
274

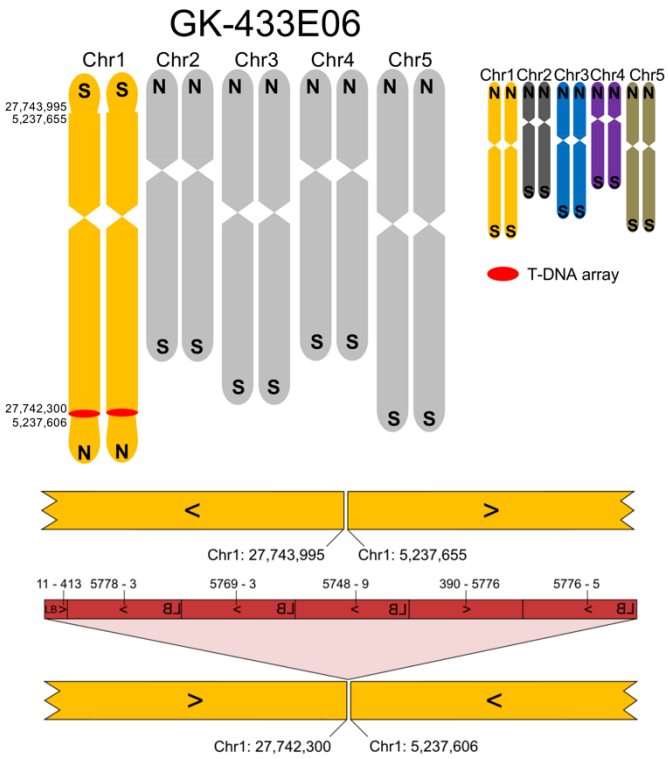


275

276 **Fig. 4**: Structure of the nuclear genome of GK-654A12 with a focus on translocations, inversions
277 and T-DNA structures. For a description of the figure elements see legend to Fig. 1. See Additional
278 File 2 for an explanation of pGABI1. The T-DNA insertion in Chr2 is associated with a small
279 duplicated inversion of about 160 bp as already described for a fraction of all T-DNA::genome
280 junctions [13].
281

282 **Intrachromosomal rearrangements and a large duplication**

283 For line GK-433E06 the FST data indicated four insertions, one T-DNA insertion (433E06-
284 At1g73770-F9L1) at Chr1:27,742,275 has been confirmed by amplicon sequencing. ONT
285 sequencing revealed an intrachromosomal translocation that exchanged the two telomeres of Chr1
286 together with about 5 Mbp DNA. The FSTs that indicated two T-DNA insertions in chromosome 1
287 were derived from one T-DNA array (Fig. 5). Once more, the compensating fusion point,
288 designated 433E06-F9L1-0-At1g73770, does not contain T-DNA sequences which was validated
289 by amplicon sequencing (Additional files 1 and 4).
290



291

292 **Fig. 5**: Structure of the nuclear genome of GK-433E06 with a focus on translocations, inversions
293 and T-DNA structures. For a description of the figure elements see legend to Fig. 1.
294

295 In line GK-767D12 a large duplication of a part of Chr2 that covers about 800 kbp was detected
296 (Fig. 6). The duplication is apparent from read coverage analyses based on read mapping against
297 the TAIR9 reference genome sequence (Col-0) which was performed for all lines studied
298 (Additional file 5). The duplicated region is inserted in reverted orientation (inversion) next to the T-

DNA insertion 767D12-At2g19210-At2g21385. This insertion was predicted by an FST at Chr2:8,338,072 and has been confirmed by PCR, the 2nd border confirmation for the T-DNA insertion failed because of reversed orientation. The other end of the duplicated inversion of Chr2 is fused to Chr2:8,338,361 (designated 767D12-At2g21385-0-At2g19210) without T-DNA sequences. Also this T-DNA-free fusion point was validated by a PCR amplicon which spanned the fusion site (Additional files 1 and 4). The T-DNA array at 767D12-At2g19210-At2g21385 is the largest we detected in this study and consists of 8 almost complete T-DNA copies and a BVB fragment arranged in diversified configurations (Fig. 6).

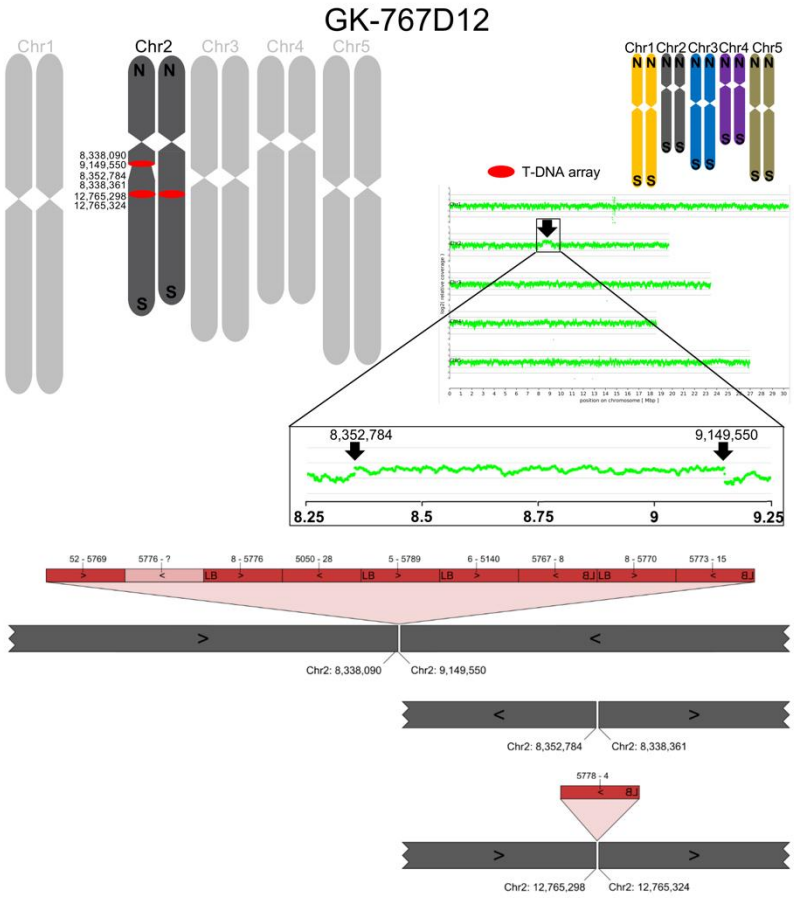


Fig. 6: Structure of the nuclear genome of GK-767D12 with a focus on translocations, inversions and T-DNA structures. For a description of the figure elements see legend to Fig. 1. On the right, results from a read coverage depth analysis are depicted that revealed a large duplication compared to the TAIR9 Col-0 reference sequence. We used read coverage depth data to decide

for the selection of the zygosity of the insertions and rearrangements displayed for Chr2 in the ideograms.

FST analyses detected only one T-DNA insertion in line GK-909H04. This insertion, designated 909H04-At1g54390, had been confirmed by PCR but failed for the 2nd border. ONT sequencing revealed an inverted duplication of about 20 kbp next to the T-DNA insertion site (Fig. 7). The fusion between this inverted duplication and the remaining part of Chr1 does not contain T-DNA sequences, but a 652 bp fragment derived from the plastome. The cpDNA insertion was validated by generating and sequencing a PCR amplicon spanning the insertion and both junctions to the genome (Additional files 1 and 4). ONT sequencing also revealed an additional insertion of a truncated T-DNA (909H04-At1g38212 at about 14.3 Mbp of Chr1) which is in the pericentromeric region not far from CEN1 (CEN1 is located at 15,086,046 to 15,087,045 and marked in the reference sequence by a gap of 1,000 Ns). Initial analyses indicated that this insertion might be associated with a deletion of about 45 kbp. However, the predicted deletion was less obvious in the read coverage depth analyses and the region is rich in TEs (also At1g38212 is annotated as "transposable element gene"). We assembled a new genome sequence of the Col-0 wild type used at GABI-Kat (assembly designated Col-0_GKat-wt, see below) and studied the structure of 909H04-At1g38212 on the basis of this assembly. The results indicated that the deletion predicted on the basis of the TAIR9 assembly is a tandemly repeated sequence region in TAIR9 which is differently represented in Col-0_GKat-wt (Additional file 6). The 3'-end of an example read from line GK-909H04 maps continuously to Col-0_GKat-wt and also to a sequence further downstream in TAIR9. The evidence collected clearly shows that there are only 13 bp deleted at the T-DNA insertion at 14.3 Mbp of Chr1 (Fig. 7), and that the initially predicted deletion is caused by errors in the TAIR9 assembly in this pericentromeric region.

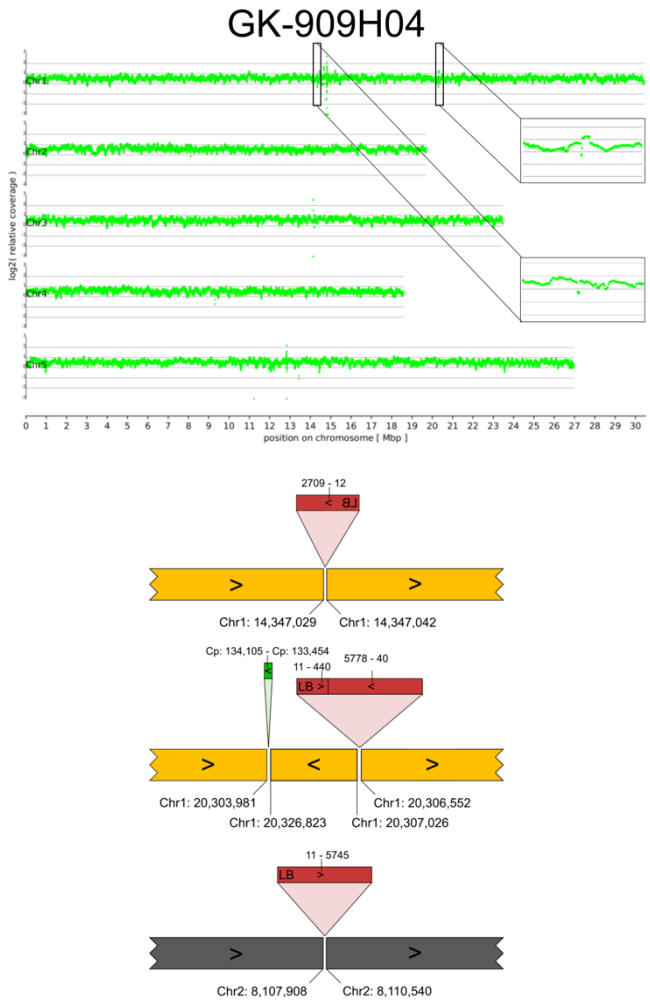


Fig. 7: Structure of the nuclear genome of GK-909H04 with a focus on insertions and T-DNA structures. For a description of the figure elements see legend to Fig. 6. The read coverage depth plot includes zoom-in enlargements of the regions at 14.3 and 20.3 Mbp of Chr1. These display variable coverage in the region of the truncated T-DNA insertion 909H04-At1g38212 (see text), and increased coverage next to the T-DNA insertion 909H04-At1g54390 which fits to the duplicated inversion detected in the local assembly of GK-909H04. Green block, sequence part from the plastome (cpDNA).

The six junction sequences that contained no T-DNA, three from compensating chromosome fusions, one from the 800 kbp inversion and two at both ends of the cpDNA insertion (see Additional file 3), were analyzed for specific features at the junctions. The observations made were fully in line with what has already been described for T-DNA insertion junctions: some short filler DNA and microhomology was found (Additional file 7). A visual overview over the T-DNA insertion

structures of all 14 lines, including those not displaying chromosomal rearrangements, is presented in Additional file 8.

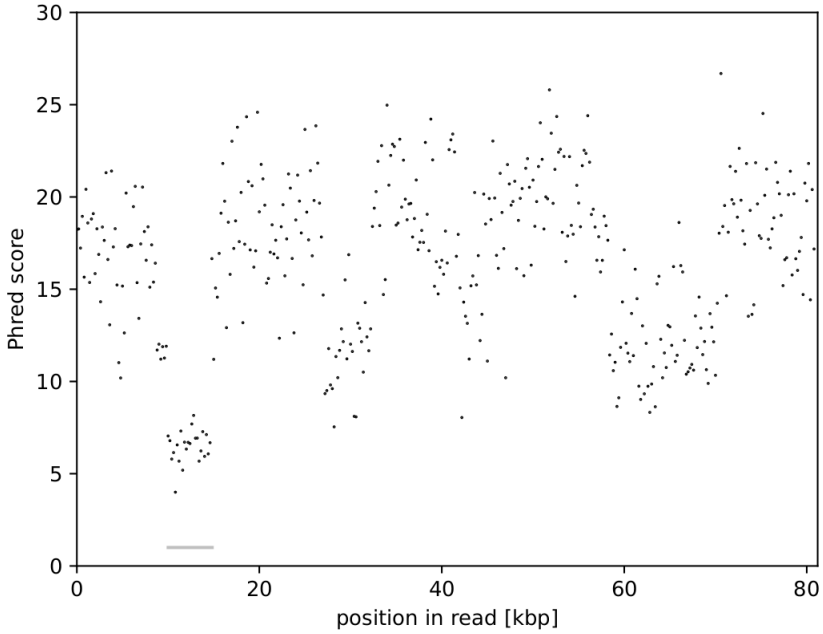
Detection of novel T-DNA insertions and T-DNA array structures

As mentioned above, 11 T-DNA insertion loci were newly detected in 7 of 14 lines studied, indicating that these were missed by FST-based studies (Table 1, Additional file 2 and 3). The primer annealing sites for FST generation at LB seem to be present in all 11 T-DNA insertions only found by ONT sequencing. Analysis of the data on T-DNA::genome junctions summarized in Additional file 3 revealed that a majority of the T-DNA structures have LB sequences at both T-DNA::genome junctions (14 of 26). The bias for the T-DNA::genome junctions involving LB is increased by the fact that several of the RB junctions were truncated, and also by some other junctions which involve BVB sequences. True T-DNA::genome junctions involving intact RB were not in the dataset, and in 14 out of 26 cases an internal RB::RB fusion was detected.
While about 40% (8 to 10 of 26, depending on judgement of small discontinuous parts) of the insertions contain a single T-DNA copy (here referred to as "canonical" insertions), some of which even further truncated and shortened, there are often cases of complex arrays of T-DNA copies inserted as T-DNA arrays. We observed a wide variety of configurations of the individual T-DNA copies within the complex arrays. In six out the 26 cases BVB sequences were detected, in the case of 038B07-At1g14080, 082G09-At5g57020-At3g19080, 430F05-At4g23850 and 947B06-T7M7 even almost complete vector sequences.

Sequence read quality decreased in T-DNA arrays

During the analyses of T-DNA insertion sequences, we frequently faced regions without sequence similarity to any sequence in the *A. thaliana* genome sequence, the sequence of the Ti-plasmid (T-DNA and BVB), or the *A. tumefaciens* genome sequence. Analysis of the read quality (Phred score) in these regions compared to other regions on the same read revealed a substantial quality drop (Fig. 8). Consequently, the number of miscalled bases in these regions is especially high. These

379 miscalls prevent matches in BLAST searches where a perfect match of several consecutive bases
380 is required as seed for a larger alignment. In some cases, the entire read displayed an extremely
381 low quality thus masking/hiding quality drops. Reads that display such locally increased error rates
382 were found in the context of T-DNA array structures which involve head-to-head or tail-to-tail
383 configurations that have the ability to form foldback structures.
384



385

386 **Fig. 8**: Decrease of Phred score in ONT reads when moving from genomic sequence into a T-DNA
387 array. Grey bar indicates the position of unclassified sequence in a T-DNA array. ID of example
388 read: a8275ad0-dce2-4dd0-a54c-947da1d8d483.
389

390 **Independent Col-0 assembly resolves misassemblies**

391 As mentioned above for the insertion allele 909H04-At1g38212, the detection of rearrangements in
392 the insertion lines is not only dependent on the quality of the reads from the genomes of the lines to
393 be studied and the assemblies that can be generated from these reads, but also from the
394 correctness of the reference sequence. While the quality of the Col-0 reference sequence (the
395 sequence from TAIR9 is still the most recent, see Introduction) is generally of very high quality,
396 there are some sequence regions that are not fully resolved. We used a subset of our ONT data,
397 namely very long (> 100 kbp, see Methods) T-DNA free reads, to *de novo* assemble the Col-0

genome sequence. The assembly, designated Col-0_GKat-wt, comprises 35 contigs after polishing and displays an N50 of 14.3 Mbp (GCA_905067165, see Additional file 9). The Col-0_GKat-wt assembly is about 4 Mbp longer than TAIR9 but still does not reach through any of the centromers. Comparison to the TAIR9 sequence indicated that the main gain in assembly length was reached in the pericentromeric regions.

Our collection of ONT sequencing datasets from the GABI-Kat lines provides a combined coverage of over 500x for the TAIR9 reference genome sequence of Col-0. In addition to using very long reads for generating an assembly, the reads were also used for identification of potentially problematic regions in the reference sequence. We identified conflicting regions by evaluating read alignments to assemblies and obtained a list of 383 candidate regions (Additional file 10). We compared selected regions against our *de novo* genome assembly and focused first on the locus At1g38212 (at about 14.3 Mbp of Chr1, see Fig. 7). The differences in this region of the TAIR9 assembly, which were detected when analyzing the T-DNA insertion allele 909H04-At1g38212 (Additional file 6), did show up again. Together with nine other examples selected across all chromosomes, Additional file 11 displays regional comparisons of TAIR9 to Col-0_GKat-wt. Not surprisingly, the 96 gaps containing various numbers of Ns which are reported for TAIR9 are frequently detected (Additional files 10 and 11).

Discussion

By sequencing GABI-Kat T-DNA insertion lines with ONT technology, we demonstrate the power of long read sequencing for the characterization of complex T-DNA insertion lines. The complexity of these lines has, at least, four aspects: (i) the number of different insertion loci present in a given line in different regions of the nuclear genome, (ii) the variance of the structures of one or several T-DNA copies appearing at a given insertion locus, (iii) the changes in the genome sequence in the direct vicinity of the T-DNA, and (iv) the changes at the chromosomal or genome level related to T-DNA integration.

425

426 **Number of T-DNA insertion loci per *A. thaliana* insertion line**

427 The average number of T-DNA insertions per *A. thaliana* T-DNA insertion line is assumed to be
428 about 1.5 [4]. However, the insertion lines available at the stock centers like NASC or ABRC list in
429 almost all cases only one insertion per line. In our limited dataset of 14 lines, 11 new insertions
430 were detected among a total of 26, indicating that one should expect an average of about 2
431 insertions per line. The 11 new insertions all contain sufficiently intact LB sequences that should
432 have allowed generation of FSTs. The reason for the lack of detection is probably that the FST data
433 generated at GABI-Kat in total have not reached the saturation level, although several insertions
434 are predicted per line at GABI-Kat [21]. The potential of existing T-DNA insertion lines for finding
435 additional knock-out alleles in existing T-DNA insertion lines is also indicated by the fact that TDNA-
436 Seq revealed additional insertion loci in established and FST-indexed lines (see Introduction).
437 Clearly, analysis by ONT sequencing can effectively reveal additional insertions and can very
438 successfully be used to fully characterize the genomes of T-DNA insertion lines. This approach is
439 faster, less laborious, more comprehensive and compared to the level of reliability also significantly
440 cheaper than PCR- or short-read based methods.

441

442 **Structure of the inserted T-DNA or T-DNA array**

443 The variance of the T-DNA structures that we were able to resolve by ONT sequencing spans a
444 really wide range of configurations and lengths. Tandem repeats as well as inverted repeats [49]
445 are occurring. Insertion length starts with 2.7 kbp for 909H04-At1g38212 and reaches up to about
446 50 kbp for 767D12-At2g19210-At2g21385. The lines were selected to contain a T-DNA by checking
447 for resistance to sulfadiazine [20] which is provided by the T-DNA used at GABI-Kat. However,
448 since there are regularly several insertions per line, also T-DNA fragments with a truncated
449 selection marker gene are to be expected - given that resistance is provided in trans. For SALK,
450 SAIL and WISC insertion lines, T-DNA arrays sizes of up to 236 kbp have been reported [32]. We
451 hypothesize that the complexity of T-DNA arrays might correlate with the tendency of selection

marker silencing, which could mechanistically be realized via siRNA [32]. The comparably reduced complexity of T-DNA arrays derived from pAC161 (the binary vectors mostly used at GABI-Kat) could thus explain why the sulfadiazine selection marker stays mostly active for many generations. Inclusion of BVB sequences in T-DNA array structures has been reported repeatedly for various species [35, 50, 51]. For the studied GABI-Kat lines, BVB sequences were structurally resolved as internal components of T-DNA arrays as well as at the junction to genomic sequences. A total of six T-DNA arrays with BVB sequences were detected among 26 cases, indicating that about 20% of all insertions, and an even higher percentage of lines, contain inserted BVB sequences.

We detected only few intact right border sequences in contrast to left border sequences, which fits to the empirical observation that FST-generation for characterization of insertion populations is much more productive at LB than at RB [4-6]. In turn, the lines studied here are selected from insertions detected by using LB for FST generation, which introduces a bias. When insertions accessed via FSTs from RB were studied, RB is found to be more precisely cut than LB [13, 52], which is explained by protection of RB by VirD2 [9]. Nevertheless, within the longer T-DNA arrays and also in the insertions newly detected by ONT sequencing in the lines studied, most of the RBs are lost. This does not fit well to current models for the integration mechanism and explanations for the observed internal "right end to right end" (without RB) fusions in T-DNA arrays and requires further investigation.

Changes in the genome sequence at the insertion site

Changes in the genome sequence in the direct vicinity of the T-DNA insertion site have already been described in detail [13]. However, this study relied on data from PCR amplicon sequences and could, therefore, not detect or analyze events that affect distances longer than the length of an average amplimer of about 2 kbp. In addition, amplicons from both T-DNA::genome junctions were required. Here, we addressed insertions that failed to fulfill the "amplicon sequences from both junctions available" criterion. This allowed to focus on a set of GABI-Kat lines that has a higher chance of showing genomic events (Table 1). The T-DNA::genome junctions studied here fall, with

one exception, generally into the range already described for DSB-based integration and repair by NHEJ, with filler sequences and microhomology at the insertion site [9]. The exception is 909H04-At1g54385-cp-At1g54440, an insertion allele that displays an about 20 kbp long duplicated inversion and in addition 652 bp derived from the plastome at the additional breakpoint that links the inversion back to the chromosome. It seems that during repair of the initial DSBs and in parallel to T-DNA integration, also cpDNA is used to join broken ends of DNA at the insertion locus. Inversions obviously require more than one repaired DSB in the DNA at the insertion site to be explained, and that cpDNA is available in the nucleus has been demonstrated experimentally [53] and in the context of horizontal gene transfer [54].

Genome level changes and translocations related to T-DNA integration

Our analyses revealed five lines with chromosome arm translocations, either exchanged within one chromosome (GK-433E06, Fig. 5) or moved to another chromosome (Figures 1 to 4). In addition, line GK-767D12 displayed a chromosomal rearrangement that resulted in an inverted duplication of 0.8 Mbp. In general, this aligns well with previous reports of interchromosomal structural variations, translocations, and chromosome fusions in T-DNA insertion lines [18, 32-34]. Because of the bias for complex cases in the criteria we used for selection of the lines investigated, we cannot deduce a reliable value for the frequency of chromosomal rearrangements in the GABI-Kat population. However, an approximation taking into account that the 6 cases are from 14 lines sequenced, and the 14 lines sequenced are a subset of 342 out of 1,818 lines with attempted confirmation of both borders but failure at the 2nd T-DNA::genome junction, ends up with about one of 10 GABI-Kat lines that may display chromosome-level rearrangements (~10%). It remains to be determined if this rough estimation holds true, but the approximation fits somehow to the percentage of T-DNA insertion lines that show Mendelian inheritance of mutant phenotypes (88%) while 12% do not [55]. For the SALK T-DNA population, 19% lines with chromosomal translocations have been reported [18], based on genetic markers and lack of linkage between markers from upstream and downstream of an insertion locus.

Although the number of investigated lines with chromosome arm translocations is small, the high proportion of fusions between Chr3 and Chr5 in our dataset is conspicuous (3 out of 5, see Additional file 3). Also for the line SAIL_232 a fusion of Chr3 and Chr5 was reported [32]. This work addressed 4 T-DNA insertion lines (two SALK, one WISC and SAIL_232) by ONT sequencing and Bionano Genomics (BNG) optical genome maps. Translocations involving chromosomes other than Chr3 and Chr5 were observed in our study and have also been reported before [16-18, 33], but it is possible that translocations between Chr3 and Chr5 occur with a higher rate than others. Full sequence characterization of the genomes of (many) more T-DNA insertion lines by long read sequencing have the potential to reveal hot spots of translocations and chromosome fusions, if these exist. It is worth nothing that T-DNA insertion related chromosomal translocations have also been reported for transgenic rice (*Oryza sativa*) [56] and transgenic birch (*Betula platyphylla* x *B. pendula*) plants [57].

Compensating translocations

The chromosome arm translocations detected are all "reciprocal" translocations, which involve two breakpoints and exchange parts of chromosomes. Both rearranged chromosomes are equally detected in the sequenced DNA of the line. Most probably, the combination of both rearranged chromosomes in the offspring is maintained because homozygosity of only one of the two rearranged chromosomes is lethal due to imbalance of gene dose for large chromosomal regions. However, if both rearranged chromosomes can be transmitted together in one gametophyte, both rearranged chromosomes might exist in offspring in homozygous state [58]. The fact that T-DNA insertion mutagenesis is accompanied by chromosome mutations, chromosomal rearrangements and chromosome arm translocations has since a long time received attention in *A. thaliana* genetic studies. One reason is that these types of mutations cause distorted segregation among offspring that are also indicative of genes essential for gametophyte development [58-60]. With regard to the chromosome arm translocations we detected, it is important to note that some of the arms are fused without integrated T-DNA. The case of line GK-082G09 (one T-DNA insertion locus, still

reciprocal chromosome arm translocation) is relevant in this context, because the presence of a single T-DNA insertion per line was used as a criterion to select valid candidates for gametophyte development mutants.

Our analyses of the sequences of T-DNA free chromosomal junctions (e.g. 082G09-At5g57020-0-At3g19080) did not result in the detection of specialties that make these junctions different from T-DNA::genome junctions. We cannot fully exclude that the T-DNA free junctions are the result of recombination of two loci that initially both contained T-DNA, and that one T-DNA got lost during recombination at one of two loci. However, it is also possible that the translocations are the direct result of DSB repair, similar to what has been realized by targeted introduction of DSBs [61]. We speculate that both, T-DNA containing and T-DNA free junction cases, result from DSB/integration/repair events that involve genome regions which happen to be in close contact, even if different chromosomes are involved. It is evident that several DSB breaks are required, and repair of these DSB can happen with the DNA that is locally available, might it be cpDNA (see above), T-DNA that must have been delivered to DSB repair sites, or different chromosomes that serve as template for fillers [13] or as target for fusion after a DSB happened to occur.

ONT sequencing of mutants offers relatively easy access to data on presence or absence of translocations. For example, the investigation of T-DNA insertion alleles/lines that display deformed pollen phenotypes, which was impacted by chromosome fusions and uncharacterized T-DNA insertions [60], can now be realized by long read sequencing to reveal all insertion and structural variation events with high resolution. Clearly, comprehensive characterization of T-DNA insertion lines, independent from the population from which the mutant originates, as well as other lines used for forward and reverse genetic experiments, can prevent unnecessary work and questionable results. While growing plants for DNA extraction can take a few weeks, the entire workflow from DNA extraction to the final genome sequence can be completed in less than a week. The application of "loreta" supports the inspection of T-DNA insertions as soon as the read data are generated.

Analyses of inverted duplicated DNA sequences by ONT sequencing

Decreased quality (Phred scores) was previously described for ONT sequence reads as consequence of inverted repeats which might form secondary structures and thus interfere with the DNA translocation through the nanopore [62]. Obviously, complex T-DNA arrays are a challenge to ONT sequencing and probably all other current sequencing technologies. We observed in such cases, which frequently occur in T-DNA arrays, that the first part of the inverted repeat has low sequence quality, while the second part (probably no longer forming a secondary structure) is of good sequence quality. The quality decrease needs to be considered especially when performing analyses at the single read level. These stretches of sequence with bad quality also pose a challenge for the assembly, especially since the orientation of the read determines which part of the inverted repeat is of good or poor quality. However, we were able to solve the problem to a satisfying level by manual consideration of reads from the opposite direction which contain the other part of the inverted repeat in good sequence quality.

Conclusions

This study presents a comprehensive characterization of multiple GABI-Kat lines by long read sequencing. The results argue very strongly for full characterization of mutant alleles to avoid misinterpretation and errors in gene function assignments. If an insertion mutant and the T-DNA insertion allele in question are not characterized well at the level of the genotype, the phenotype observed for the mutant might be due to a complex integration locus, and not causally related to the gene that is expected to be knocked-out by the insertion allele. Structural changes at the genome level, including chromosome translocations and other large rearrangements with junctions without T-DNA, may have confounding effects when studying the genotype to phenotype relations with T-DNA lines. This conclusion must also consider that during the last 20 to 30 years, many T-DNA alleles have been used in reverse genetic experiments. Finally, and similar to the ONT sequence data that resulted from the analyses of four SALK and SAIL/WISC T-DNA insertion lines [32], the

ONT sequence data from this study allowed to detect and correct many non-centromeric misassemblies in the current reference sequence.

Methods

Plant material

The lines subjected to ONT sequencing were chosen from a collection of GABI-Kat lines which were studied initially to collect statistically meaningful data about the structure of T-DNA insertion sites at both ends of the T-DNA insertions [13]. In this context and also after 2015, confirmation amplicon sequence data from both T-DNA::genome junctions of individual T-DNA insertions were created at GABI-Kat, which was successful for 1,481 cases from 1,476 lines by the end of 2019 (1,319 cases were successfully completed for both junctions in the beginning of 2015). To generate this dataset, 1,835 individual T-DNA insertions from 1,818 lines with one T-DNA::genome junction already confirmed were addressed, meaning that there were 354 cases from 342 lines which failed at the 2nd T-DNA::genome junction. From these 354 cases, we randomly selected the 14 insertions (in 14 different lines) that were studied here, with good germination as additional criterion for effective handling (Additional file 1). Since the focus of interest in insertion alleles was always NULL alleles of genes, all 14 insertions addressed are CDSi insertions (insertions in the CDS or enclosed introns). A total of 100 T2 seeds of each line were plated with sulfadiazine selection as described [24]. Surviving T2 plantlets should contain at least one integrated T-DNA, either in hemizygous or in homozygous state. Sulfadiazine-resistant plantlets were transferred to soil, grown to about 8-leaf stage and pooled for DNA extraction. For a single locus with normal heritability, statistically 66% of the chromosomes in the pool should contain the T-DNA. The T-DNA in GK-654A12 is from pGABI1 [36], the other lines contain T-DNA from pAC161 [20].

DNA extraction, size enrichment, and quality assessment

Genomic DNA was extracted from young plantlets or young leaves through a CTAB-based protocol (Additional File 12) modified from [20, 63]. We observed like others before [4] that the quality of extracted DNA decreased with the age of the leaf material processed, with very young leaves leading to best results in our hands. The cause might be increasing cell and vacuole size containing more harmful metabolites which might be responsible for reduced quality and yield in DNA extractions. As DNA quality for ONT sequencing decreases with storage time, we processed the DNA as soon as possible after extraction. DNA quantity and quality was initially assessed based on NanoDrop (Thermo Scientific) measurement, and on an agarose gel for DNA fragment size distribution. Precise DNA quantification was performed via Qubit (Thermo Fisher) measurement using the broad range buffer following the supplier's instructions. Up to 9 µg of genomic DNA were subjected to an enrichment of long fragments via Short Read Eliminator kit (Circulomics) according to the suppliers' instructions.

Library preparation and ONT sequencing

DNA solutions enriched for long fragments were quantified via Qubit again. One µg DNA (R9.4.1 flow cells) or two µg (R10 flow cells) were subjected to library preparation following the LSK109 protocol provided by ONT. Sequencing was performed on R9.4.1 and R10 flow cells on a GridION. Real time base calling was performed using Guppy v3.0 on the GridION (R9.4.1 flow cells) and on graphic cards in the de.NBI cloud [64] (R10 flow cells), respectively.

***De novo* genome sequence assemblies**

Reads of each GK line were assembled separately to allow validation of other analysis methods (see below). Canu v1.8 [65] was deployed with previously optimized parameters [66]. Assembly quality was assessed based on a previously developed Python script (Table 2). No polishing was performed for assemblies of individual GABI-Kat lines as these assemblies were only used to analyze large structural variants and specifically T-DNA insertions.

Through removal of all T-DNA reads from the combined ONT read dataset from all insertion lines (Col-0 background [20]) and size filtering, a comprehensive data set of very long reads (> 100.000 nt) was generated. This dataset is available from ENA/GenBank with the ID ERS5246674 (SAMEA7490021). The assembly of these very long reads was computed as described above. Polishing was performed with Racon v.1.4.7 [67] and medaka v.0.10.0 as previously described [63]. Potential contamination sequences were removed based on sequence similarity to the genome sequences of other species, and contigs small than 100 kbp were discarded as previously described [63, 68]. To ensure accurate representation of the Col-0 wild type genome structure, the assembly was checked for the chromosome fusion events reported for GK-082G09, GK-433E06, and GK-654A12 as well as for the chloroplast DNA integration of GK-909H04. The Sanger reads of the validation amplicons generated for these loci were subjected to a search via BLASTn [69] using default settings. BLASTn was also used to validate the absence of any T-DNA or plasmid sequences in this assembly using pSKI015 (AF187951), pAC161 (AJ537514), and pROK2 [6] as query.

Analyses of the Col-0_GKat-wt assembly and comparison to TAIR9

To identify differences to the TAIR9 reference genome sequence, the Col-0_GK-wt contigs were sorted and orientated using pseudogenetic markers derived from TAIR9. The TAIR9 sequence was split into 500 bp long sequence chunks which were searched against the Col-0_GK-wt contigs via BLAST. Unique hits with at least 80% of the maximal possible BLAST score were considered as genetic markers. The following analysis with ALLMAPS [70] revealed additional and thus unmatched sequences of Col-0_GK-wt around the centromeres.

ONT reads were aligned to the TAIR9 reference genome sequence via Minimap2 v2.1-r761 [71]. Mappings were converted into BED files with bedtools v2.26 [72]. The alignments were evaluated for the ends of mapping reads, and these ends were quantified in genomic bins of 100 bp using a dedicated tool designated Assembly Error Finder (AEF) v0.12 (Table 2) with default parameters. Neighboring regions with high numbers of alignment ends were grouped if their distance was

667 smaller than 30 kbp. Regions with outstanding high numbers of alignment ends indicate potential
668 errors in the targeted assembly. A selection of these regions from TAIR9 was compared against
669 Col-0_GKat-wt through dot plots [30].
670
671 **Analysis of T-DNA insertions**
672 The T-DNA insertions of each line were analyzed in a semi-automatic way. A tool was developed,
673 written in Python and designated "loreta" (Table 2), that needs as input: reads in FASTQ format, T-
674 DNA sequences in FASTA format, a reference file containing sequences for assembly annotation in
675 FASTA format (for this study: sequences of T-DNA and vector backbone, the *A. thaliana* nuclear
676 genome, plastome, chondrome, and the *A. tumefaciens genome*), and – if available – precomputed
677 *de novo* assemblies (as described above). The results are HTML pages with annotated images
678 displaying models of T-DNA insertions and their neighborhood. Partial assemblies of reads
679 containing T-DNA sequences are computed, and the parts of the *de novo* assemblies containing T-
680 DNA sequences are extracted. All resulting sequences as well as all individual reads containing T-
681 DNA sequences are annotated using the reference file. If run on a local machine, a list of tools that
682 need to be installed is given in the github repository (Table 2). For easier access to the tool on a
683 local machine, there is also a Docker file available in the github repository that can be used to build
684 a Docker image.
685 Reads containing T-DNA sequences were identified by BLASTn [69, 73] using an identity cutoff of
686 80% and an e-value cutoff of 1e-50. All identified reads were then assembled using Canu v1.8 with
687 the same parameters as for the *de novo* assemblies (see above) and in addition some parameters
688 to facilitate assemblies with low coverage: correctedErrorRate=0.17, corOutCoverage=200,
689 stopOnLowCoverage=5 and an expected genome size of 10 kbp. From the precomputed *de novo*
690 assemblies, fragments were extracted that contain the T-DNA insertion and 50 kbp up- and
691 downstream sequence. The resulting fragments, contigs from the Canu assembly, the contigs
692 marked as "unassembled" by Canu as well as all individual reads (converted to FASTA using
693 Seqtk-1.3-r106 [74]) were annotated using the reference sequences. For this purpose, a BLASTn

search was performed (contig versus reference sequences) with the same parameters as for the identification of T-DNA reads. These BLAST results were mapped to the sequence (contig/read) as follows: BLAST hits were annotated one after another, sorted by decreasing score. If the overlap of a BLAST hit with a previously annotated one exceeds 10 bp, the second BLAST hit was discarded. For further analysis, reads were mapped back to the assembly. All reads were mapped back to the fragments of the *de novo* assembly, reads containing T-DNA were mapped back to (1) the Canu contigs, (2) "unassembled" contigs and (3) individual reads containing T-DNA sequences. Mapping was performed using Minimap2 [71] with the default options for mapping of ONT sequencing data. To further inspect the chromosome(s) sequences prior to T-DNA insertion, the same analysis was performed using the *A. thaliana* sequences neighboring the T-DNA insertion. A FASTA file was generated that contains these flanking sequences using bedtools [72], reads containing this part of *A. thaliana* sequence (and no T-DNA) are again identified using BLAST, assembled and annotated as described above. Infoseq from the EMBOSS package [75] was used to calculate the length of different sequences in the pipeline.

All information is summarized in HTML files containing images; these images display all annotated sequences along with mapped reads and details on the BLAST results. These pages were used for manual inspection and final determination of the insertion structures. For canonical insertions, such as 050B11-At5g64610, the assembled and annotated contig of the partial assembly was sufficient. In more complex cases like GK-038B07, the exact insertion structure was not clear based on the assembled contigs or based on sections of the *de novo* assembly. If, based on the read mappings shown in the visualization, the partial assembly looked erroneous (many partial mappings), individual reads were used for the determination of the insertion structure. These individual reads were also considered for clearer determination of exact T-DNA positions, if these positions were not clear from contigs / sections. This was often the case for head-to-tail configurations of T-DNA arrays, where one of the T-DNAs was represented by sequence of low quality and could not be annotated (and led to misassemblies in assembled contigs). These cases were resolved by identification of reads orientated in the other direction, because then the sequence derived from the

721 other T-DNA was of low quality and by combining the annotated results, a clear picture could be
722 derived. If different reads contradicted each other in exact positions of the T-DNA, we chose the
723 "largest possible T-DNA" that could be explained by individual reads.
724
725 **Mapping of ONT reads for detection of copy number variation**
726 ONT reads from each line were aligned against the TAIR9 Col-0 reference genome sequence using
727 Minimap v2.10-r761 [76] with the options '-ax map-ont --secondary=no'. The resulting mappings
728 were converted into BAM files via samtools v1.8 [77] and used for the construction of coverage files
729 with a previously developed Python script [78]. Coverage plots (see Additional file 5) were
730 constructed as previously described [66] and manually inspected for the identification of copy
731 number variations.
732
733 **Sequence read quality assessment**
734 Reads containing T-DNA sequence were annotated based on sequence similarity to other known
735 sequences based on BLASTn [69, 73] usually matching parts of the Ti-plasmid or *A. thaliana*
736 genome sequence. Reads associated with complex T-DNA insertions were considered for
737 downstream analysis if substantial parts (>1 kbp) of the read sequence were not matched to any
738 database sequences via BLASTn. Per base quality (Phred score) of such reads was assessed
739 based on a sliding window of 200 nt with a step size of 100 nt.
740
741 **Chromosome fusion and cpDNA insertion validation via PCR and Sanger sequencing**
742 Chromosomal fusions without a connecting T-DNA were analyzed via PCR using manually
743 designed flanking primers (Additional file 4). Amplicons were generated using genomic DNA
744 extracted from plants of the respective line as template with Q5 High-Fidelity DNA polymerase
745 (NEB) following supplier's instructions. PCR products were separated on a 1% agarose gel and
746 visualized using ethidiumbromide and UV light. Amplicons were purified with Exo-CIP Rapid PCR
747 Cleanup Kit (NEB) following supplier's instructions. Sanger sequencing was performed at the

Sequencing Core Facility of the Center for Biotechnology (Bielefeld University, Bielefeld, Germany) using BigDye terminator v3.1 chemistry (Thermo Fisher) on a 3730XL sequencer. The resulting Sanger sequences were merged using tools from the EMBOSS package [75]. After transferring reverse reads to their reverse complement using revseq, a multiple alignment was generated using MAFFT [79]. One consensus sequence for each amplicon was extracted from the alignments using em_cons with option -plurality 1, and the resulting sequences were submitted to ENA (see Additional file 1 for accession numbers).

Analyses of T-DNA free chromosome fusion junctions

The five junction sequences were analyzed by BLAST essentially as described [13]. Briefly, searches were performed against all possible target sequences (*A. tumefaciens*; *A. thaliana* nucleome, plastome and chondrome; T-DNA and vector backbone) using BLASTn default parameters. If the complete query was not covered, the unmatched part of the query sequence was classified as filler and extracted. Subsequently, this sequence was used in a BLAST search with an e-value cutoff of 10, a word-size of 5 and the ‘-task "blastn-short"’ option activated to detect smaller and lower quality hits. If this was not successful (as in GK-909H04), the filler sequence was extended by 10 bases up- and downstream and the procedure described above was repeated.

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

Sequence read datasets generated and analyzed during this study were made available at ENA under the accession PRJEB35658. Individual run IDs are included in Additional file 1. The Col-0 genome sequence assembly of the GABI-Kat Col-0 genetic background (Col-0_GKat-wt) is available at ENA under the accession GCA_905067165.

Table 2: Availability of scripts.

Description	URLs
Previously developed scripts for general tasks	https://github.com/bpucker/script_collection
Tool for the analysis of ONT datasets ("loreta")	https://github.com/nkleinbo/loreta
Scripts for the analyses presented in this study	https://github.com/bpucker/GKseq

Competing interest

The authors declare that they have no competing interests.

Funding

We acknowledge support for the Article Processing Charge by the Open Access Publication Fund of Bielefeld University.

Authors' contribution

BP performed DNA extraction and sequencing. BP and NK performed bioinformatic analyses. BP, NK, and BW interpreted the results and wrote the manuscript.

Acknowledgements

We are very grateful to the Bioinformatics Resource Facility support team of the CeBiTec and to de.NBI for providing computing infrastructure and excellent technical support. We also thank the Sequencing Core Facility of the CeBiTec for granting access to the ONT infrastructure. Many thanks to Tobias Busche, Christian Rückert, and Jörn Kalinowski for general support related to the ONT sequencing, and to Prisva Viehöver for high quality Sanger sequencing. We thank Andrea

Voigt for excellent technical support. The bioinformatic work was supported in part by grants from the German Federal Ministry of Education and Research (BMBF) for the project "Bielefeld-Gießen Center for Microbial Bioinformatics–BiGi" (grant no. 031A533) within the German Network for Bioinformatics Infrastructure (de.NBI).

Supplementary information

Additional file 1: Summary of GABI-Kat insertion line data, segregation data for the F2 families after selection for sulfadiazine resistance, sequencing data including run IDs from submission to ENA/SRA, and accession numbers of T-DNA free junction sequences.

Additional file 2: Extended version of Table 1 covering all 14 lines.

Additional file 3: Overview of the T-DNA insertions and associated structural variants in the investigated GABI-Kat lines.

Additional file 4: Sequences of oligonucleotides used for the validation of fusion points of chromosomal translocations and other large structural variants.

Additional file 5: Read coverage of all analyzed lines in relation to the TAIR9 reference genome sequence.

Additional file 6: Structure of genomic locus around one insertion in GK-909H04.

Additional file 7: Analysis results of genomic fusion junction sequences without T-DNA insertion.

Additional file 8: Visual overview over all insertions detected.

Additional file 9: Assembly statistics of Col-0_GK-wt.

Additional file 10: Potential errors in the TAIR9 reference genome sequence of Col-0.

Additional file 11: Dot plots between TAIR9 and Col-0_GK-wt for potential errors in the reference sequence.

Additional file 12: Protocol for the extraction of genomic DNA from *A. thaliana* for ONT sequencing.

References

1. Ulker B, Peiter E, Dixon DP, Moffat C, Capper R, Bouche N, Edwards R, Sanders D, Knight H, Knight MR: **Getting the most out of publicly available T-DNA insertion lines.** *The Plant Journal* 2008, **56:**665-677.
2. OMalley RC, Ecker JR: **Linking genotype to phenotype using the Arabidopsis unimutant collection.** *The Plant Journal* 2010, **61:**928-940.
3. Fauser F, Roth N, Pacher M, Ilg G, Sanchez-Fernandez R, Biesgen C, Puchta H: **In planta gene targeting.** *Proceedings of the National Academy of Sciences of the United States of America* 2012, **109:**7535-7540.
4. OMalley RC, Barragan CC, Ecker JR: **A user's guide to the Arabidopsis T-DNA insertion mutant collections.** In *Methods in Molecular Biology. Volume* 1284. 2015/03/12 edition. Edited by Alonso J, Stepanova A. New York, NY: Humana Press; 2015: 323-342
5. Strizhov N, Li Y, Rosso MG, Viehoever P, Dekker KA, Weisshaar B: **High-throughput generation of sequence indexes from T-DNA mutagenized Arabidopsis thaliana lines.** *BioTechniques* 2003, **35:**1164-1168.
6. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK, Zimmerman J, Barajas P, Cheuk R, et al: **Genome-wide insertional mutagenesis of Arabidopsis thaliana.** *Science* 2003, **301:**653-657.
7. Puchta H: **Applying CRISPR/Cas for genome engineering in plants: the best is yet to come.** *Current Opinion in Plant Biology* 2017, **36:**1-8.
8. Smith EF, Townsend CO: **A Plant-Tumor of Bacterial Origin.** *Science* 1907, **25:**671-673.
9. Gelvin SB: **Integration of Agrobacterium T-DNA into the Plant Genome.** *Annual Review of Genetics* 2017, **51:**195-217.
10. Zambryski P, Holsters M, Kruger K, Depicker A, Schell J, Van Montagu M, Goodman H: **Tumor DNA structure in plant cells transformed by A. tumefaciens.** *Science* 1980, **209:**1385-1391.
11. Hernalsteens J, Van Vliet F, De Beuckeleer M, Depicker A, Engler G, Lemmers M, Holsters M, Van Montagu M, Schell J: **The Agrobacterium tumefaciens Ti plasmid as a host vector system for introducing foreign DNA in plant cells.** *Nature* 1980, **287:**654-656.
12. Clough SJ, Bent AF: **Floral dip: a simplified method for Agrobacterium-mediated transformation of *Arabidopsis thaliana*.** *The Plant Journal* 1998, **16:**735-743.

13. Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B: **The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism.** *Molecular Plant* 2015, **8:**1651-1664.
14. van Kregten M, de Pater S, Romeijn R, van Schendel R, Hooykaas PJ, Tijsterman M: **T-DNA integration in plants results from polymerase-θ-mediated DNA repair.** *Nature Plants* 2016, **2:**16164.
15. Castle LA, Errampalli D, Atherton TL, Franzmann LH, Yoon ES, Meinke DW: **Genetic and molecular characterization of embryonic mutants identified following seed transformation in Arabidopsis.** *Molecular Genetics and Genomics* 1993, **241:**504-514.
16. Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R: **A comprehensive characterization of single-copy T-DNA insertions in the Arabidopsis thaliana genome.** *Plant Molecular Biology* 2003, **52:**161-176.
17. Lafleuriel J, Degroote F, Depeiges A, Picard G: **A reciprocal translocation, induced by a canonical integration of a single T-DNA, interrupts the HMG-I/Y Arabidopsis thaliana gene.** *Plant Physiology and Biochemistry* 2004, **42:**171-179.
18. Clark KA, Krysan PJ: **Chromosomal translocations are a common phenomenon in Arabidopsis thaliana T-DNA insertion lines.** *The Plant Journal* 2010, **64:**990-1001.
19. Min Y, Frost JM, Choi Y: **Gametophytic Abortion in Heterozygotes but Not in Homozygotes: Implied Chromosome Rearrangement during T-DNA Insertion at the ASF1 Locus in Arabidopsis.** *Molecules and Cells* 2020, **43:**448-458.
20. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B: **An *Arabidopsis thaliana* T-DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse genetics.** *Plant Molecular Biology* 2003, **53:**247-259.
21. Kleinboelting N, Huep G, Kloetgen A, Viehoever P, Weisshaar B: **GABI-Kat SimpleSearch: new features of the Arabidopsis thaliana T-DNA mutant database.** *Nucleic Acids Research* 2012, **40:**D1211-D1215.
22. Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden J, Ko C, et al: **A High-Throughput Arabidopsis Reverse Genetics System.** *The Plant Cell* 2002, **14:**2985-2994.
23. Sussman MR, Amasino RM, Young JC, Krysan PJ, Austin-Phillips S: **The Arabidopsis knockout facility at the University of Wisconsin-Madison.** *Plant Physiology* 2000, **124:**1465-1467.
24. Li Y, Rosso MG, Viehoever P, Weisshaar B: **GABI-Kat SimpleSearch: an Arabidopsis thaliana T-DNA mutant database with detailed information for confirmed insertions.** *Nucleic Acids Research* 2007, **35:**D874-D878.

25. Kleinboelting N, Huep G, Weisshaar B: **Enhancing the GABI-Kat Arabidopsis thaliana T-DNA Insertion Mutant Database by Incorporating Araport11 Annotation.** *Plant and Cell Physiology* 2017, **58:**e7.
26. Cheng CY, Krishnakumar V, Chan A, Thibaud-Nissen F, Schobel S, Town CD: **Araport11: a complete reannotation of the Arabidopsis thaliana reference genome.** *The Plant Journal* 2017, **89:**789-804.
27. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, et al: **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools.** *Nucleic Acids Research* 2012, **40:**D1202-D1210.
28. Huep G, Kleinboelting N, Weisshaar B: **An easy-to-use primer design tool to address paralogous loci and T-DNA insertion sites in the genome of Arabidopsis thaliana.** *Plant Methods* 2014, **10:**28.
29. Vukasinovic N, Cvrckova F, Elias M, Cole R, Fowler JE, Zarsky V, Synek L: **Dissecting a hidden gene duplication: the Arabidopsis thaliana SEC10 locus.** *PLoS ONE* 2014, **9:**e94077.
30. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B: **A chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-1 genome and its gene set.** *PLoS One* 2019, **14:**e0216233.
31. Krispil R, Tannenbaum M, Sarusi-Portuguez A, Loza O, Raskina O, Hakim O: **The Position and Complex Genomic Architecture of Plant T-DNA Insertions Revealed by 4SEE.** *International Journal of Molecular Sciences* 2020, **21:**ijms21072373.
32. Jupe F, Rivkin AC, Michael TP, Zander M, Motley ST, Sandoval JP, Slotkin RK, Chen H, Castanon R, Nery JR, Ecker JR: **The complex architecture and epigenomic impact of plant T-DNA insertions.** *PLoS Genetics* 2019, **15:**e1007819.
33. Nacry P, Camilleri C, Courtial B, Caboche M, Bouchez D: **Major chromosomal rearrangements induced by T-DNA transformation in Arabidopsis.** *Genetics* 1998, **149:**641-650.
34. Tax FE, Vernon DM: **T-DNA-associated duplication/translocations in Arabidopsis. Implications for mutant analysis and functional genomics.** *Plant Physiology* 2001, **126:**1527-1538.
35. Krizkova L, Hrouda M: **Direct repeats of T-DNA integrated in tobacco chromosome: characterization of junction regions.** *The Plant Journal* 1998, **16:**673-680.
36. Ulker B, Li Y, Rosso MG, Logemann E, Somssich IE, Weisshaar B: **T-DNA-mediated transfer of Agrobacterium tumefaciens chromosomal DNA into plants.** *Nature Biotechnology* 2008, **26:**1015-1017.

931 37. Seagrist JF, Su SH, Krysan PJ: **Recombination between T-DNA insertions to cause**
932 **chromosomal deletions in Arabidopsis is a rare phenomenon.** *PeerJ* 2018, **6:**e5076.
933 38. Wendel JF, Jackson SA, Meyers BC, Wing RA: **Evolution of plant genome architecture.**
934 *Genome Biology* 2016, **17:**37.
935 39. Huang K, Rieseberg LH: **Frequency, Origins, and Evolutionary Role of Chromosomal**
936 **Inversions in Plants.** *Frontiers in Plant Science* 2020, **11:**296.
937 40. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA,
938 Grimwood J, Gundlach H, et al: **The Arabidopsis lyrata genome sequence and the basis**
939 **of rapid genome size change.** *Nature Genetics* 2011, **43:**476-481.
940 41. Chaney L, Sharp AR, Evans CR, Udall J: **Genome Mapping in Plant Comparative**
941 **Genomics.** *Trends in Plant Science* 2016, **21:**770-780.
942 42. Jiao WB, Schneeberger K: **Chromosome-level assemblies of multiple Arabidopsis**
943 **genomes reveal hotspots of rearrangements with altered evolutionary dynamics.**
944 *Nature Communications* 2020, **11:**989.
945 43. Schmidt C, Schindele P, Puchta H: **From gene editing to genome engineering:**
946 **restructuring plant chromosomes via CRISPR/Cas.** *aBIOTECH* 2020, **1:**21-31.
947 44. Pellestor F, Gatinois V: **Chromoanagenesis: a piece of the macroevolution scenario.**
948 *Molecular Cytogenetics* 2020, **13:**3.
949 45. Li X, Zhang R, Patena W, Gang SS, Blum SR, Ivanova N, Yue R, Robertson JM, Lefebvre
950 PA, Fitz-Gibbon ST, et al: **An Indexed, Mapped Mutant Library Enables Reverse**
951 **Genetics Studies of Biological Processes in Chlamydomonas reinhardtii.** *The Plant*
952 *Cell* 2016, **28:**367-387.
953 46. Inagaki S, Henry IM, Lieberman MC, Comai L: **High-Throughput Analysis of T-DNA**
954 **Location and Structure Using Sequence Capture.** *PLoS One* 2015, **10:**e0139672.
955 47. Jiang N, Lee YS, Mukundi E, Gomez-Cano F, Rivero L, Grotewold E: **Diversity of genetic**
956 **lesions characterizes new Arabidopsis flavonoid pigment mutant alleles from T-DNA**
957 **collections.** *Plant Science* 2020, **291:**110335.
958 48. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker
959 JR: **High contiguity Arabidopsis thaliana genome assembly with a single nanopore**
960 **flow cell.** *Nature Communications* 2018, **9:**541.
961 49. Jorgensen R, Snyder C, Jones JDG: **T-DNA is organized predominantly in inverted**
962 **repeat structures in plants transformed with Agrobacterium tumefaciens C58**
963 **derivatives.** *Molecular Genetics and Genomics* 1987, **207:**471-477.
964 50. Wu H, Sparks CA, Jones HD: **Characterisation of T-DNA loci and vector backbone**
965 **sequences in transgenic wheat produced by Agrobacterium-mediated transformation.**
966 *Molecular Breeding* 2006, **18:**195-208.

51. Rajapriya V, Kannan P, Sridevi G, Veluthambi K: **A rare transgenic event of rice with Agrobacterium binary vector backbone integration at the right T-DNA border junction.** *Journal of Plant Biochemistry and Biotechnology* 2021.

52. Zambryski P, Depicker A, Kruger K, Goodman HM: **Tumor induction by Agrobacterium tumefaciens: analysis of the boundaries of T-DNA.** *Journal of Molecular and Applied Genetics* 1982, **1:**361-370.

53. Huang CY, Ayliffe MA, Timmis JN: **Direct measurement of the transfer rate of chloroplast DNA into the nucleus.** *Nature* 2003, **422:**72-76.

54. Bock R: **The give-and-take of DNA: horizontal gene transfer in plants.** *Trends in Plant Science* 2009, **15:**11-22.

55. Feldmann KA: **T-DNA insertion mutagenesis in *Arabidopsis*: mutational spektrum.** *The Plant Journal* 1991, **1:**71-82.

56. Wei FJ, Kuang LY, Oung HM, Cheng SY, Wu HP, Huang LT, Tseng YT, Chiou WY, Hsieh-Feng V, Chung CH, et al: **Somaclonal variation does not preclude the use of rice transformants for genetic screening.** *The Plant Journal* 2016, **85:**648-659.

57. Gang H, Liu G, Zhang M, Zhao Y, Jiang J, Chen S: **Comprehensive characterization of T-DNA integration induced chromosomal rearrangement in a birch T-DNA mutant.** *BMC Genomics* 2019, **20:**311.

58. Curtis MJ, Belcram K, Bollmann SR, Tominey CM, Hoffman PD, Mercier R, Hays JB: **Reciprocal chromosome translocation associated with TDNA-insertion mutation in Arabidopsis: genetic and cytological analyses of consequences for gametophyte development and for construction of doubly mutant lines.** *Planta* 2009, **229:**731-745.

59. Bonhomme S, Horlow C, Vezon D, de Laissardiere S, Guyon A, Ferault M, Marchand M, Bechtold N, Pelletier G: **T-DNA mediated disruption of essential gametophytic genes in Arabidopsis is unexpectedly rare and cannot be inferred from segregation distortion alone.** *Molecular Genetics and Genomics* 1998, **260:**444-452.

60. Ruprecht C, Carroll A, Persson S: **T-DNA-induced chromosomal translocations in feronia and anxur2 mutants reveal implications for the mechanism of collapsed pollen due to chromosomal rearrangements.** *Molecular Plant* 2014, **7:**1591-1594.

61. Schmidt C, Fransz P, Ronspies M, Dreissig S, Fuchs J, Heckmann S, Houben A, Puchta H: **Changing local recombination patterns in Arabidopsis by CRISPR/Cas mediated chromosome engineering.** *Nature Communications* 2020, **11:**4418.

62. Spealman P, Burrell J, Gresham D: **Inverted duplicate DNA sequences increase translocation rates through sequencing nanopores resulting in reduced base calling accuracy.** *Nucleic Acids Research* 2020, **48:**4940-4945.

63. Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B: **High Contiguity De Novo Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read Sequencing.** *Genes* 2020, **11:**E274.

64. Belmann P, Fischer B, Kruger J, Prochazka M, Rasche H, Prinz M, Hanussek M, Lang M, Bartusch F, Glassle B, et al: **de.NBI Cloud federation through ELIXIR AAI.** *F1000Research* 2019, **8:**842.

65. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome Research* 2017, **27:**722-736.

66. Pucker B, Ruckert C, Stracke R, Viehover P, Kalinowski J, Weisshaar B: **Twenty-Five Years of Propagation in Suspension Cell Culture Results in Substantial Alterations of the Arabidopsis Thaliana Genome.** *Genes* 2019, **10:**671.

67. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly from long uncorrected reads.** *Genome Research* 2017, **27:**737-746.

68. Pucker B, Holtgräwe D, Rosleff Sörensen T, Stracke R, Viehöver P, Weisshaar B: **A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny.** *PLoS ONE* 2016, **11:**e0164321.

69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of Molecular Biology* 1990, **215:**403-410.

70. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J: **ALLMAPS: robust scaffold ordering based on multiple maps.** *Genome Biology* 2015, **16:**3.

71. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018, **34:**3094-3100.

72. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26:**841-842.

73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL: **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10:**421.

74. **Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences** [https://github.com/lh3/seqtk]

75. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software Suite.** *Trends in Genetics* 2000, **16:**276-277.

76. Li H: **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.** *Bioinformatics* 2016, **32:**2103-2110.

77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-2079.

78. Pucker B, Brockington SF: **Genome-wide analyses supported by RNA-Seq reveal non-canonical splice sites in plant genomes.** *BMC Genomics* 2018, **19:**980.

79. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7: improvements in performance and usability.** *Molecular Biology and Evolution* 2013, **30:**772-780.