bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1
2
3
Large scale genomic rearrangements in selected *Arabidopsis thaliana* T-DNA lines are caused by T-DNA insertion mutagenesis

4

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

6 1 Genetics and Genomics of Plants, Center for Biotechnology (CeBiTec), Bielefeld University,
7 Sequenz 1, 33615 Bielefeld, Germany

8 2 Evolution and Diversity, Department of Plant Sciences, University of Cambridge, Cambridge,
9 United Kingdom

10 3 Bioinformatics Resource Facility, Center for Biotechnology (CeBiTec, Bielefeld University,
11 Sequenz 1, 33615 Bielefeld, Germany

12 + authors contributed equally

13 * corresponding author: Bernd Weisshaar

14

15 BP: bpucker@cebitec.uni-bielefeld.de, ORCID: 0000-0002-3321-7471
16 NK: nkleinbo@cebitec.uni-bielefeld.de, ORCID: 0000-0001-9124-5203
17 BW: bernd.weisshaar@uni-bielefeld.de, ORCID: 0000-0002-7635-3473
18
19
20
21
22
23
24
25
26
27
28
29
30

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

31

32 # Abstract

33 ## Background

34 Experimental prove of gene function assignments in plants is heavily based on mutant analyses. T-
35 DNA insertion lines provided an invaluable resource of mutants and enabled systematic reverse
36 genetics-based investigation of the functions of *Arabidopsis thaliana* genes during the last decades.
37

38 ## Results

39 We sequenced the genomes of 14 *A. thaliana* GABI-Kat T-DNA insertion lines, which eluded
40 flanking sequence tag-based attempts to characterize their insertion loci, with Oxford Nanopore
41 Technologies (ONT) long reads. Complex T-DNA insertions were resolved and 11 previously
42 unknown T-DNA loci identified, suggesting that the number of T-DNA insertions per line was
43 underestimated. T-DNA mutagenesis caused fusions of chromosomes along with compensating
44 translocations to keep the gene set complete throughout meiosis. Also, an inverted duplication of
45 800 kbp was detected. About 10% of GABI-Kat lines might be affected by chromosomal
46 rearrangements, some of which do not involve T-DNA. Local assembly of selected reads was
47 shown to be a computationally effective method to resolve the structure of T-DNA insertion loci. We
48 developed an automated workflow to support investigation of long read data from T-DNA insertion
49 lines. All steps from DNA extraction to assembly of T-DNA loci can be completed within days.
50

51 ## Conclusion

52 Long read sequencing was demonstrated to be a very effective way to resolve complex T-DNA
53 insertions and chromosome fusions. Many T-DNA insertions comprise not just a single T-DNA, but
54 complex arrays of multiple T-DNAs. It is becoming obvious that T-DNA insertion alleles must be
55 characterized by exact identification of both T-DNA::genome junctions to generate clear genotype-
56 to-phenotype relations.
57

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

85 from optimized binary plasmids are transformed into *A. thaliana* plants via floral dip to generate
86 stable lines [12]. T-DNA transfer into the nucleus of a plant cell is supported by several VIR proteins
87 which are, in the biotechnologically optimized system, encoded on a separate helper plasmid. It is
88 assumed that host proteins are responsible for integration of the T-DNA into the genome, most
89 likely as a DNA double strand into a double strand break (DSB) using host DNA repair pathways
90 and DNA polymerase theta [9, 13, 14]. T-DNA integration resembles DNA break repair through
91 non-homologous end-joining (NHEJ) or microhomology-mediated end-joining (MMEJ) and is often
92 accompanied by the presence of filler DNA or microhomology at both T-DNA::genome junctions [9,
93 13]. Chromosomal inversions and translocations are commonly associated with T-DNA insertions
94 [15-19], suggesting that often more than just one DSB is associated with T-DNA integration [9].
95 The most important collections of T-DNA lines for the model plant *Arabidopsis thaliana* are SALK
96 (150,000 lines) [6], GABI-Kat (92,000 lines) [20, 21], SAIL (54,000 lines) [22], and WISC (60,000
97 lines) [23]. In total, over 700,000 insertion lines have been constructed [4]. GABI-Kat lines were
98 generated through the integration of a T-DNA harboring a sulfadiazine resistance gene for selection
99 of transformed lines [20]. Additionally, the T-DNA contains a 35S promoter at RB causing
100 transcriptional up-regulation of genes next to the integration site if the right part of the T-DNA next
101 to RB stays intact during integration [1]. Integration sites were predicted based on FSTs and
102 allowed access to knock-out alleles of numerous genes. At GABI-Kat, T-DNA insertion alleles were
103 confirmed by an additional "confirmation PCR" using DNA from the T2 generation [24] prior to the
104 release of a mutant line and submission of the line to the Nottingham Arabidopsis Stock Centre
105 (NASC). Researchers could identify suitable and available T-DNA insertion lines via SimpleSearch
106 on the GABI-Kat website [25]. Since 2017, SimpleSearch uses Araport11 annotation data [26].
107 Araport11 is based on the *A. thaliana* Col-0 reference genome sequence from TAIR9 which
108 includes about 96 annotated gaps filled with Ns [27], among them the centromers and several gaps
109 in the pericentromeric regions.
110 The prediction of integration sites based on bioinformatic evaluations using FST data does often
111 not reveal the complete picture. Insertions might be masked from FST predictions due to truncated

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

112 borders [13], because of repetitive sequences or paralogous regions in the genome [28], or even
113 lack of the true insertion site in the reference sequence used for FST mapping [29, 30]. Also,
114 confirmation by sequencing an amplicon that spans the predicted insertion site at one of the two
115 expected T-DNA::genome junctions is not fully informative. Deletions and target site duplications at
116 the integration site can occur and can only be detected by examining both borders of the inserted
117 T-DNA [13]. In addition, more complex insertions have been reported by several studies that
118 include large deletions, insertions, inversions or even chromosomal translocations [13, 18, 31-34].
119 Also, binary vector backbone (BVB) sequences have been detected at insertion sites [35] as well
120 as fragments of *A. tumefaciens* chromosomal DNA [36]. In addition, recombination between two T-
121 DNA loci was described as a mechanism for deletion of an enclosed genomic fragment [37].
122 Plant genomes are dynamic and often show whole genome doubling followed by purging processes
123 [38, 39]. Transposable elements (TE) play an important role in restructuring genomes [38], but
124 chromosomal rearrangement events not involving TEs also lead to large structural variation [33, 40,
125 41]. The karyotype of *A. thaliana* is the result of chromosome fusion events which reduced the
126 chromosome number from the ancestral eight to five [40]. Recent advances in long read
127 sequencing pave the way for comprehensive synteny analyses with Brassica species related to *A.*
128 *thaliana*. A recent study reported 13-17 Mbp of rearranged sequence between pairs of
129 geographically diverse *A. thaliana* accessions [42]. Also, structural variants have the potential to
130 contribute to speciation [39]. Chromosomal rearrangements can occur during the repair of DSBs,
131 e.g. via microhomology-mediated end joining or non-allelic homologous recombination [reviewed by
132 43, 44]. Evidently, regions with high sequence similarity like duplications are especially prone to
133 chromosomal rearrangements [43].
134 While usually one T-DNA locus per line was identified by FSTs, the number of T-DNA insertion loci
135 per line is usually higher. For GABI-Kat, it was estimated that about 50% of all lines (12,018 of
136 21,049 tested, according to numbers from the end of 2019) display a single insertion locus. This
137 estimation is based on segregation analyses using sulfadiazine resistance as a selection marker
138 [20]. Other insertion mutant collections report similar numbers [4]. The average number of T-DNA

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

139 insertions per line was reported to be about 1.5, but this is probably a significant underestimation
140 since the kanamycin and BASTA selection marker genes applied to determine the numbers are
141 known to be silenced quite often [4]. For these reasons, it is required that insertion mutants (similar
142 to mutants created by e.g. chemical mutagenesis) are backcrossed to wild type prior to
143 phenotyping a homozygous line.
144 The FSTs produced for the different mutant populations by individual PCR and Sanger-sequencing
145 allowed usually access to a single T-DNA insertion locus per line, although for GABI-Kat there are
146 several examples with up to three confirmed insertion loci based on FST data (e.g. line GK-011F01,
147 see [25]). This leaves a significant potential of undiscovered T-DNA insertions in lines already
148 available at the stock centers, which has been exploited by the group of Joe Ecker by applying
149 TDNA-Seq (Illumina technology) to the SALK and a part of the GABI-Kat mutant populations.
150 Essentially the same technology has later also been used to set up a sequence indexed insertion
151 mutant library of *Chlamydomonas reinhardtii* [45]. With the fast development of new DNA
152 sequencing technologies, the comprehensive characterization of T-DNA insertion lines comes into
153 reach.
154 Several studies already harnessed high throughput sequencing technologies to investigate T-DNA
155 insertion and other mutant lines [46, 47]. Oxford Nanopore Technologies (ONT) provides a cost-
156 effective and fast approach to study *A. thaliana* genomes, since a single MinION/GridION Flow Cell
157 delivers sufficient data to assemble one genotype [48]. Here, we present a method to fully
158 characterize T-DNA insertion loci and additional genomic changes of T-DNA insertion lines through
159 ONT long read sequencing. We selected 14 lines that contain confirmed T-DNA insertion alleles
160 (first border or first T-DNA::genome junction confirmed by sequencing an amplicon across one
161 junction), but which escaped characterization of the second T-DNA::genome junction (we refer to
162 the T-DNA::genome junction that is expected to exist after confirmation of one T-DNA::genome
163 junction as "2nd border"). Within this biased set of lines, we detected several chromosome fragment
164 or chromosome arm translocations, a duplication of 800 kbp and also an insertion of DNA from the
165 chloroplast (plastome), all related to T-DNA insertion events. The results clearly demonstrate the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

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

191

192

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

199 1, Additional file 2 and 3). In case of GK-038B07, the lack of re-detection of the expected insertion
200 allele of At4g19510 was explained by a PCR template contamination during the initial confirmation,
201 the line that contains the real insertion (source of the contamination) is most probably GK-159D11.
202 A similar explanation is true in case of GK-040A12 where the expected insertion allele of
203 At1g52720 was also not found in the ONT data. At least, the error detected fits to the selection
204 criteria, because the 2nd border or 2nd T-DNA::genome junction can obviously not be detected if the
205 T-DNA insertion as such is not present in the line.
206 In six of the 14 lines studied by ONT whole genome sequencing, chromosomal rearrangements
207 were found. To visualize these results, we created ideograms of the five *A. thaliana* chromosomes
208 with a color code for each of the chromosomes. The colors allow to visually perceive information on
209 chromosome arm translocations, and the changeover points indicate presence or absence of T-
210 DNA sequences.
211

212 **Chromosome fusions**

213 In four lines, fusions of different chromosomes were detected. These fusions result from
214 chromosome arm translocations which were, in all four cases, compensated within the line by
215 reciprocal translocations. The T-DNA insertion on chromosome 5 (Chr5) of GK-038B07 is part of a
216 complex chromosome arm translocation (Fig. 1). A part of Chr5 is fused to Chr3, the replaced part
217 of Chr3 is fused to an inversion of 2 Mbp on Chr5. This inversion contains T-DNAs at both ends,
218 one of which is the insertion predicted by FSTs.
219

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

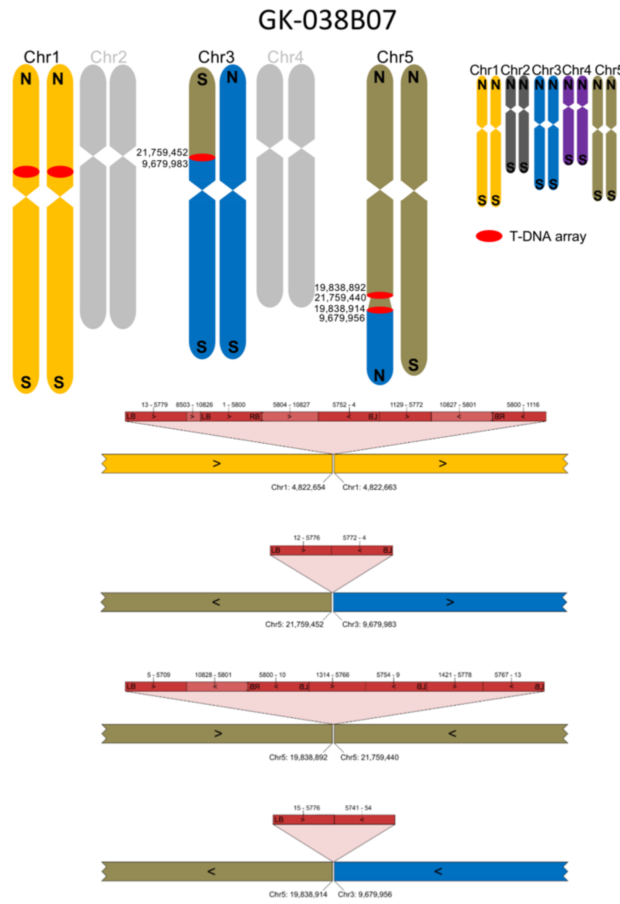
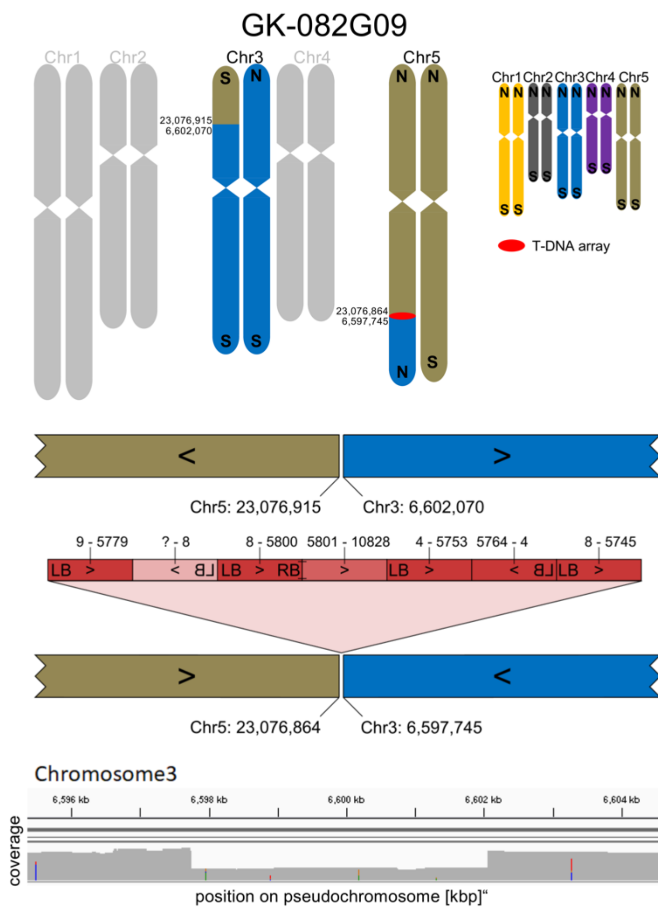


Fig. 1: Structure of the nuclear genome of GK-038B07 with a focus on translocations, inversions and T-DNA structures. Upper right: color codes used for the five chromosomes; N, northern end of chromosome; S, southern end of chromosome. Upper left: ideograms of the chromosomes that display the reciprocal 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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

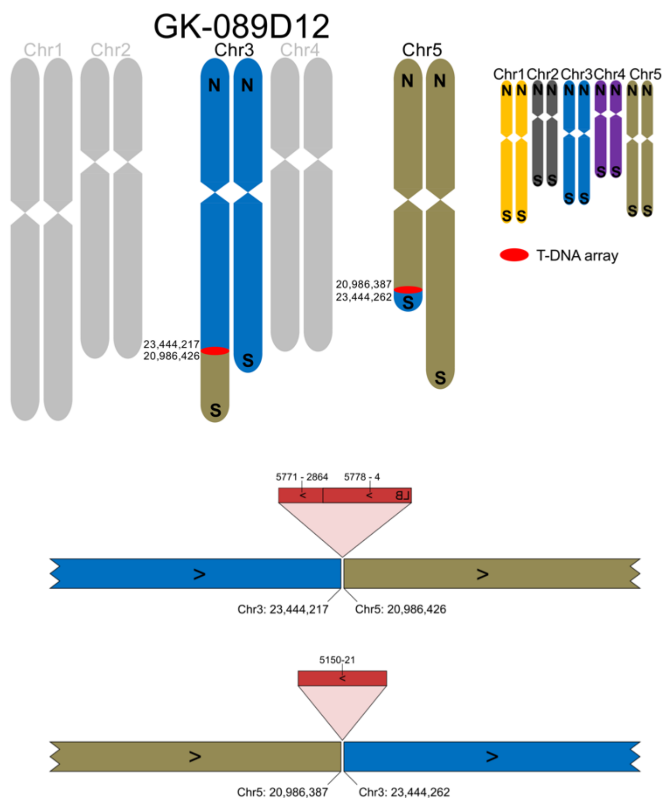


245

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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



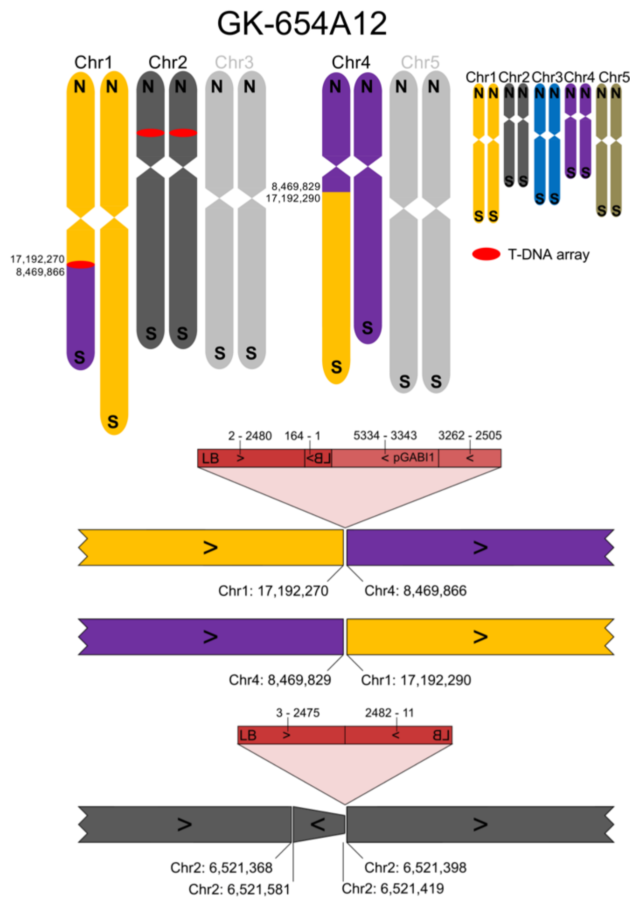
260

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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



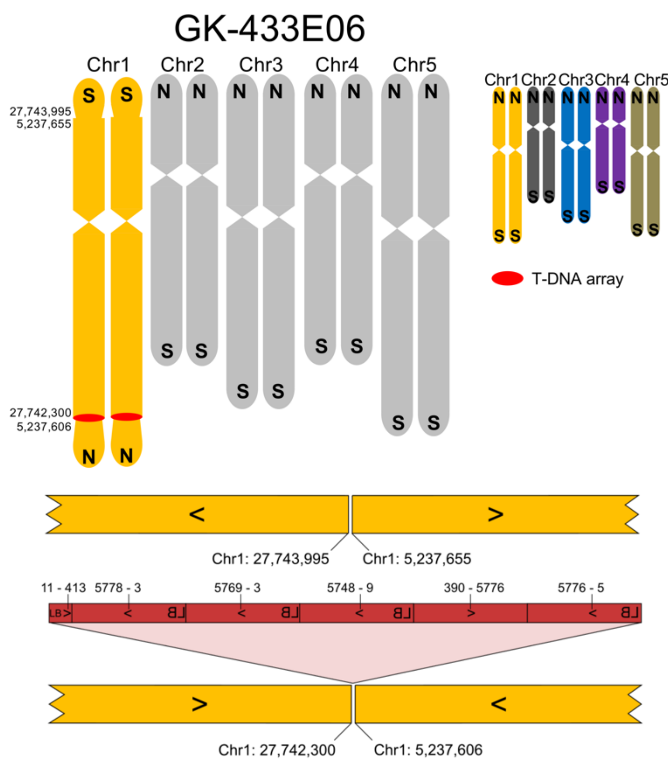
275

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**

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

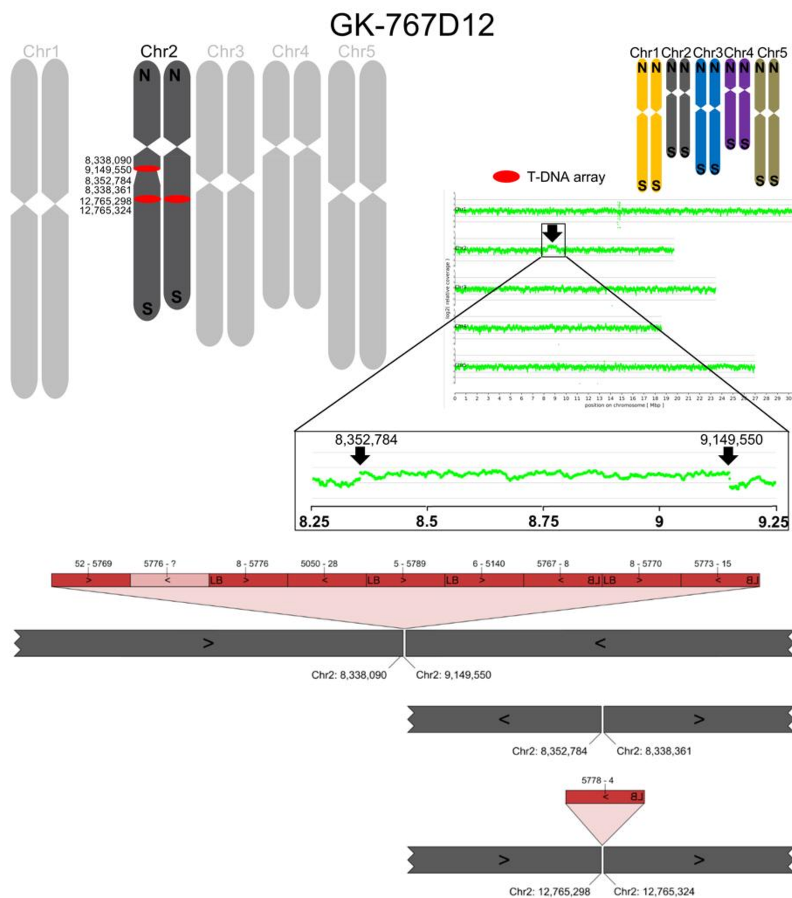
292 **Fig. 5**: Structure of the nuclear genome of GK-433E06 with a focus on translocations, inversions
293 and T-DNA structures. For a description of the figure elements see legend to Fig. 1.
294

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-

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

307



308

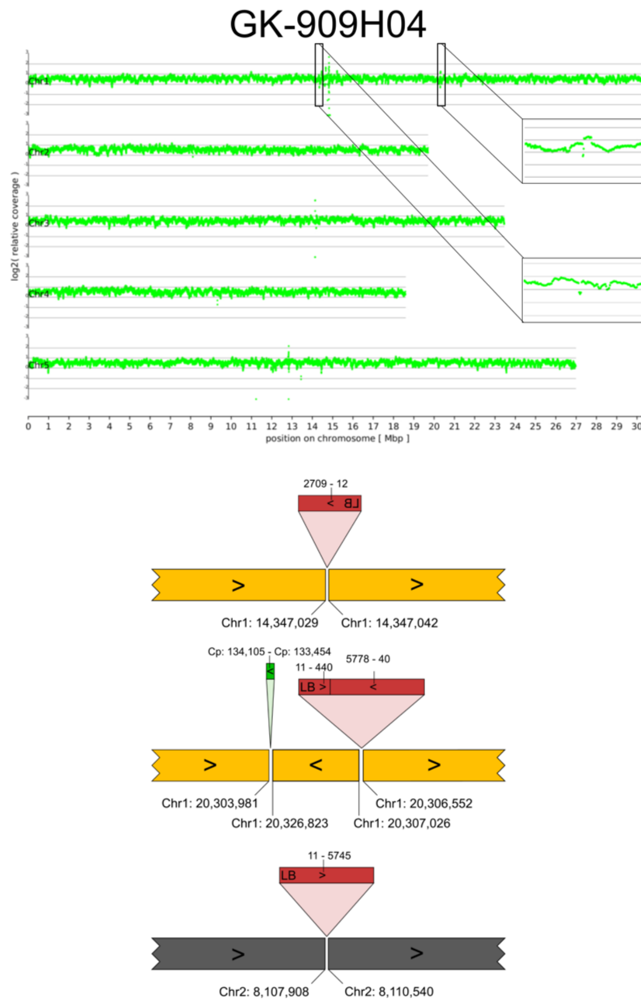
309 **Fig. 6:** Structure of the nuclear genome of GK-767D12 with a focus on translocations, inversions
310 and T-DNA structures. For a description of the figure elements see legend to Fig. 1. On the right,
311 results from a read coverage depth analysis are depicted that revealed a large duplication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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

313 for the selection of the zygosity of the insertions and rearrangements displayed for Chr2 in the
314 ideograms.
315

316 FST analyses detected only one T-DNA insertion in line GK-909H04. This insertion, designated
317 909H04-At1g54390, had been confirmed by PCR but failed for the 2nd border. ONT sequencing
318 revealed an inverted duplication of about 20 kbp next to the T-DNA insertion site (Fig. 7). The
319 fusion between this inverted duplication and the remaining part of Chr1 does not contain T-DNA
320 sequences, but a 652 bp fragment derived from the plastome. The cpDNA insertion was validated
321 by generating and sequencing a PCR amplicon spanning the insertion and both junctions to the
322 genome (Additional files 1 and 4). ONT sequencing also revealed an additional insertion of a
323 truncated T-DNA (909H04-At1g38212 at about 14.3 Mbp of Chr1) which is in the pericentromeric
324 region not far from CEN1 (CEN1 is located at 15,086,046 to 15,087,045 and marked in the
325 reference sequence by a gap of 1,000 Ns). Initial analyses indicated that this insertion might be
326 associated with a deletion of about 45 kbp. However, the predicted deletion was less obvious in the
327 read coverage depth analyses and the region is rich in TEs (also At1g38212 is annotated as
328 "transposable element gene"). We assembled a new genome sequence of the Col-0 wild type used
329 at GABI-Kat (assembly designated Col-0_GKat-wt, see below) and studied the structure of
330 909H04-At1g38212 on the basis of this assembly. The results indicated that the deletion predicted
331 on the basis of the TAIR9 assembly is a tandemly repeated sequence region in TAIR9 which is
332 differently represented in Col-0_GKat-wt (Additional file 6). The 3'-end of an example read from line
333 GK-909H04 maps continuously to Col-0_GKat-wt and also to a sequence further downstream in
334 TAIR9. The evidence collected clearly shows that there are only 13 bp deleted at the T-DNA
335 insertion at 14.3 Mbp of Chr1 (Fig. 7), and that the initially predicted deletion is caused by errors in
336 the TAIR9 assembly in this pericentromeric region.
337

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



338

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

346

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

352 structures of all 14 lines, including those not displaying chromosomal rearrangements, is presented
353 in Additional file 8.
354

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

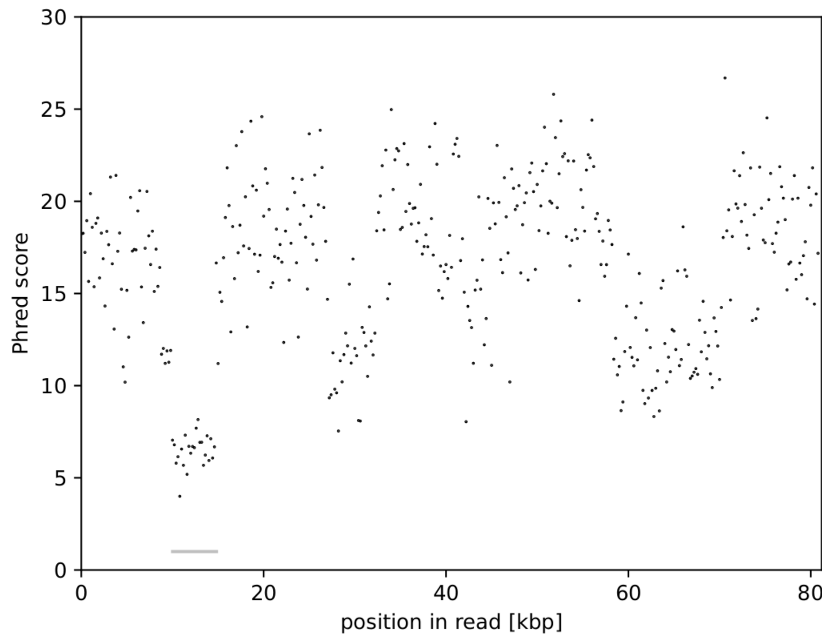
356 As mentioned above, 11 T-DNA insertion loci were newly detected in 7 of 14 lines studied,
357 indicating that these were missed by FST-based studies (Table 1, Additional file 2 and 3). The
358 primer annealing sites for FST generation at LB seem to be present in all 11 T-DNA insertions only
359 found by ONT sequencing. Analysis of the data on T-DNA::genome junctions summarized in
360 Additional file 3 revealed that a majority of the T-DNA structures have LB sequences at both T-
361 DNA::genome junctions (14 of 26). The bias for the T-DNA::genome junctions involving LB is
362 increased by the fact that several of the RB junctions were truncated, and also by some other
363 junctions which involve BVB sequences. True T-DNA::genome junctions involving intact RB were
364 not in the dataset, and in 14 out of 26 cases an internal RB::RB fusion was detected.
365 While about 40% (8 to 10 of 26, depending on judgement of small discontinuous parts) of the
366 insertions contain a single T-DNA copy (here referred to as "canonical" insertions), some of which
367 even further truncated and shortened, there are often cases of complex arrays of T-DNA copies
368 inserted as T-DNA arrays. We observed a wide variety of configurations of the individual T-DNA
369 copies within the complex arrays. In six out the 26 cases BVB sequences were detected, in the
370 case of 038B07-At1g14080, 082G09-At5g57020-At3g19080, 430F05-At4g23850 and 947B06-
371 T7M7 even almost complete vector sequences.
372

373 **Sequence read quality decreased in T-DNA arrays**

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

398 genome sequence. The assembly, designated Col-0_GKat-wt, comprises 35 contigs after polishing
399 and displays an N50 of 14.3 Mbp (GCA_905067165, see Additional file 9). The Col-0_GKat-wt
400 assembly is about 4 Mbp longer than TAIR9 but still does not reach through any of the centromers.
401 Comparison to the TAIR9 sequence indicated that the main gain in assembly length was reached in
402 the pericentromeric regions.
403 Our collection of ONT sequencing datasets from the GABI-Kat lines provides a combined coverage
404 of over 500x for the TAIR9 reference genome sequence of Col-0. In addition to using very long
405 reads for generating an assembly, the reads were also used for identification of potentially
406 problematic regions in the reference sequence. We identified conflicting regions by evaluating read
407 alignments to assemblies and obtained a list of 383 candidate regions (Additional file 10). We
408 compared selected regions against our *de novo* genome assembly and focused first on the locus
409 At1g38212 (at about 14.3 Mbp of Chr1, see Fig. 7). The differences in this region of the TAIR9
410 assembly, which were detected when analyzing the T-DNA insertion allele 909H04-At1g38212
411 (Additional file 6), did show up again. Together with nine other examples selected across all
412 chromosomes, Additional file 11 displays regional comparisons of TAIR9 to Col-0_GKat-wt. Not
413 surprisingly, the 96 gaps containing various numbers of Ns which are reported for TAIR9 are
414 frequently detected (Additional files 10 and 11).
415
416

417
Discussion

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

452 marker silencing, which could mechanistically be realized via siRNA [32]. The comparably reduced
453 complexity of T-DNA arrays derived from pAC161 (the binary vectors mostly used at GABI-Kat)
454 could thus explain why the sulfadiazine selection marker stays mostly active for many generations.
455 Inclusion of BVB sequences in T-DNA array structures has been reported repeatedly for various
456 species [35, 50, 51]. For the studied GABI-Kat lines, BVB sequences were structurally resolved as
457 internal components of T-DNA arrays as well as at the junction to genomic sequences. A total of six
458 T-DNA arrays with BVB sequences were detected among 26 cases, indicating that about 20% of all
459 insertions, and an even higher percentage of lines, contain inserted BVB sequences.
460 We detected only few intact right border sequences in contrast to left border sequences, which fits
461 to the empirical observation that FST-generation for characterization of insertion populations is
462 much more productive at LB than at RB [4-6]. In turn, the lines studied here are selected from
463 insertions detected by using LB for FST generation, which introduces a bias. When insertions
464 accessed via FSTs from RB were studied, RB is found to be more precisely cut than LB [13, 52],
465 which is explained by protection of RB by VirD2 [9]. Nevertheless, within the longer T-DNA arrays
466 and also in the insertions newly detected by ONT sequencing in the lines studied, most of the RBs
467 are lost. This does not fit well to current models for the integration mechanism and explanations for
468 the observed internal "right end to right end" (without RB) fusions in T-DNA arrays and requires
469 further investigation.
470

471 **Changes in the genome sequence at the insertion site**

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

479 one exception, generally into the range already described for DSB-based integration and repair by
480 NHEJ, with filler sequences and microhomology at the insertion site [9]. The exception is 909H04-
481 At1g54385-cp-At1g54440, an insertion allele that displays an about 20 kbp long duplicated
482 inversion and in addition 652 bp derived from the plastome at the additional breakpoint that links
483 the inversion back to the chromosome. It seems that during repair of the initial DSBs and in parallel
484 to T-DNA integration, also cpDNA is used to join broken ends of DNA at the insertion locus.
485 Inversions obviously require more than one repaired DSB in the DNA at the insertion site to be
486 explained, and that cpDNA is available in the nucleus has been demonstrated experimentally [53]
487 and in the context of horizontal gene transfer [54].
488

489 **Genome level changes and translocations related to T-DNA integration**

490 Our analyses revealed five lines with chromosome arm translocations, either exchanged within one
491 chromosome (GK-433E06, Fig. 5) or moved to another chromosome (Figures 1 to 4). In addition,
492 line GK-767D12 displayed a chromosomal rearrangement that resulted in an inverted duplication of
493 0.8 Mbp. In general, this aligns well with previous reports of interchromosomal structural variations,
494 translocations, and chromosome fusions in T-DNA insertion lines [18, 32-34]. Because of the bias
495 for complex cases in the criteria we used for selection of the lines investigated, we cannot deduce a
496 reliable value for the frequency of chromosomal rearrangements in the GABI-Kat population.
497 However, an approximation taking into account that the 6 cases are from 14 lines sequenced, and
498 the 14 lines sequenced are a subset of 342 out of 1,818 lines with attempted confirmation of both
499 borders but failure at the 2nd T-DNA::genome junction, ends up with about one of 10 GABI-Kat lines
500 that may display chromosome-level rearrangements (~10%). It remains to be determined if this
501 rough estimation holds true, but the approximation fits somehow to the percentage of T-DNA
502 insertion lines that show Mendelian inheritance of mutant phenotypes (88%) while 12% do not [55].
503 For the SALK T-DNA population, 19% lines with chromosomal translocations have been reported
504 [18], based on genetic markers and lack of linkage between markers from upstream and
505 downstream of an insertion locus.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

506 Although the number of investigated lines with chromosome arm translocations is small, the high
507 proportion of fusions between Chr3 and Chr5 in our dataset is conspicuous (3 out of 5, see
508 Additional file 3). Also for the line SAIL_232 a fusion of Chr3 and Chr5 was reported [32]. This work
509 addressed 4 T-DNA insertion lines (two SALK, one WISC and SAIL_232) by ONT sequencing and
510 Bionano Genomics (BNG) optical genome maps. Translocations involving chromosomes other than
511 Chr3 and Chr5 were observed in our study and have also been reported before [16-18, 33], but it is
512 possible that translocations between Chr3 and Chr5 occur with a higher rate than others. Full
513 sequence characterization of the genomes of (many) more T-DNA insertion lines by long read
514 sequencing have the potential to reveal hot spots of translocations and chromosome fusions, if
515 these exist. It is worth nothing that T-DNA insertion related chromosomal translocations have also
516 been reported for transgenic rice (*Oryza sativa*) [56] and transgenic birch (*Betula platyphylla* x *B.*
517 *pendula*) plants [57].
518
519 **Compensating translocations**
520 The chromosome arm translocations detected are all "reciprocal" translocations, which involve two
521 breakpoints and exchange parts of chromosomes. Both rearranged chromosomes are equally
522 detected in the sequenced DNA of the line. Most probably, the combination of both rearranged
523 chromosomes in the offspring is maintained because homozygosity of only one of the two
524 rearranged chromosomes is lethal due to imbalance of gene dose for large chromosomal regions.
525 However, if both rearranged chromosomes can be transmitted together in one gametophyte, both
526 rearranged chromosomes might exist in offspring in homozygous state [58]. The fact that T-DNA
527 insertion mutagenesis is accompanied by chromosome mutations, chromosomal rearrangements
528 and chromosome arm translocations has since a long time received attention in *A. thaliana* genetic
529 studies. One reason is that these types of mutations cause distorted segregation among offspring
530 that are also indicative of genes essential for gametophyte development [58-60]. With regard to the
531 chromosome arm translocations we detected, it is important to note that some of the arms are
532 fused without integrated T-DNA. The case of line GK-082G09 (one T-DNA insertion locus, still

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

533 reciprocal chromosome arm translocation) is relevant in this context, because the presence of a
534 single T-DNA insertion per line was used as a criterion to select valid candidates for gametophyte
535 development mutants.
536 Our analyses of the sequences of T-DNA free chromosomal junctions (e.g. 082G09-At5g57020-0-
537 At3g19080) did not result in the detection of specialties that make these junctions different from T-
538 DNA::genome junctions. We cannot fully exclude that the T-DNA free junctions are the result of
539 recombination of two loci that initially both contained T-DNA, and that one T-DNA got lost during
540 recombination at one of two loci. However, it is also possible that the translocations are the direct
541 result of DSB repair, similar to what has been realized by targeted introduction of DSBs [61]. We
542 speculate that both, T-DNA containing and T-DNA free junction cases, result from
543 DSB/integration/repair events that involve genome regions which happen to be in close contact,
544 even if different chromosomes are involved. It is evident that several DSB breaks are required, and
545 repair of these DSB can happen with the DNA that is locally available, might it be cpDNA (see
546 above), T-DNA that must have been delivered to DSB repair sites, or different chromosomes that
547 serve as template for fillers [13] or as target for fusion after a DSB happened to occur.
548 ONT sequencing of mutants offers relatively easy access to data on presence or absence of
549 translocations. For example, the investigation of T-DNA insertion alleles/lines that display deformed
550 pollen phenotypes, which was impacted by chromosome fusions and uncharacterized T-DNA
551 insertions [60], can now be realized by long read sequencing to reveal all insertion and structural
552 variation events with high resolution. Clearly, comprehensive characterization of T-DNA insertion
553 lines, independent from the population from which the mutant originates, as well as other lines used
554 for forward and reverse genetic experiments, can prevent unnecessary work and questionable
555 results. While growing plants for DNA extraction can take a few weeks, the entire workflow from
556 DNA extraction to the final genome sequence can be completed in less than a week. The
557 application of "loreta" supports the inspection of T-DNA insertions as soon as the read data are
558 generated.
559

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

560 **Analyses of inverted duplicated DNA sequences by ONT sequencing**

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

573

574

575 Conclusions

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

587 ONT sequence data from this study allowed to detect and correct many non-centromeric
588 misassemblies in the current reference sequence.
589
590
591 # Methods
592
593 **Plant material**
594 The lines subjected to ONT sequencing were chosen from a collection of GABI-Kat lines which
595 were studied initially to collect statistically meaningful data about the structure of T-DNA insertion
596 sites at both ends of the T-DNA insertions [13]. In this context and also after 2015, confirmation
597 amplicon sequence data from both T-DNA::genome junctions of individual T-DNA insertions were
598 created at GABI-Kat, which was successful for 1,481 cases from 1,476 lines by the end of 2019
599 (1,319 cases were successfully completed for both junctions in the beginning of 2015). To generate
600 this dataset, 1,835 individual T-DNA insertions from 1,818 lines with one T-DNA::genome junction
601 already confirmed were addressed, meaning that there were 354 cases from 342 lines which failed
602 at the 2nd T-DNA::genome junction. From these 354 cases, we randomly selected the 14 insertions
603 (in 14 different lines) that were studied here, with good germination as additional criterion for
604 effective handling (Additional file 1). Since the focus of interest in insertion alleles was always NULL
605 alleles of genes, all 14 insertions addressed are CDSi insertions (insertions in the CDS or enclosed
606 introns). A total of 100 T2 seeds of each line were plated with sulfadiazine selection as described
607 [24]. Surviving T2 plantlets should contain at least one integrated T-DNA, either in hemizygous or in
608 homozygous state. Sulfadiazine-resistant plantlets were transferred to soil, grown to about 8-leaf
609 stage and pooled for DNA extraction. For a single locus with normal heritability, statistically 66% of
610 the chromosomes in the pool should contain the T-DNA. The T-DNA in GK-654A12 is from pGABI1
611 [36], the other lines contain T-DNA from pAC161 [20].
612
613 **DNA extraction, size enrichment, and quality assessment**

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

614 Genomic DNA was extracted from young plantlets or young leaves through a CTAB-based protocol
615 (Additional File 12) modified from [20, 63]. We observed like others before [4] that the quality of
616 extracted DNA decreased with the age of the leaf material processed, with very young leaves
617 leading to best results in our hands. The cause might be increasing cell and vacuole size containing
618 more harmful metabolites which might be responsible for reduced quality and yield in DNA
619 extractions. As DNA quality for ONT sequencing decreases with storage time, we processed the
620 DNA as soon as possible after extraction. DNA quantity and quality was initially assessed based on
621 NanoDrop (Thermo Scientific) measurement, and on an agarose gel for DNA fragment size
622 distribution. Precise DNA quantification was performed via Qubit (Thermo Fisher) measurement
623 using the broad range buffer following the supplier's instructions. Up to 9 µg of genomic DNA were
624 subjected to an enrichment of long fragments via Short Read Eliminator kit (Circulomics) according
625 to the suppliers' instructions.
626
627 **Library preparation and ONT sequencing**
628 DNA solutions enriched for long fragments were quantified via Qubit again. One µg DNA (R9.4.1
629 flow cells) or two µg (R10 flow cells) were subjected to library preparation following the LSK109
630 protocol provided by ONT. Sequencing was performed on R9.4.1 and R10 flow cells on a GridION.
631 Real time base calling was performed using Guppy v3.0 on the GridION (R9.4.1 flow cells) and on
632 graphic cards in the de.NBI cloud [64] (R10 flow cells), respectively.
633
634 ***De novo* genome sequence assemblies**
635 Reads of each GK line were assembled separately to allow validation of other analysis methods
636 (see below). Canu v1.8 [65] was deployed with previously optimized parameters [66]. Assembly
637 quality was assessed based on a previously developed Python script (Table 2). No polishing was
638 performed for assemblies of individual GABI-Kat lines as these assemblies were only used to
639 analyze large structural variants and specifically T-DNA insertions.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

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

656 To identify differences to the TAIR9 reference genome sequence, the Col-0_GK-wt contigs were
657 sorted and orientated using pseudogenetic markers derived from TAIR9. The TAIR9 sequence was
658 split into 500 bp long sequence chunks which were searched against the Col-0_GK-wt contigs via
659 BLAST. Unique hits with at least 80% of the maximal possible BLAST score were considered as
660 genetic markers. The following analysis with ALLMAPS [70] revealed additional and thus
661 unmatched sequences of Col-0_GK-wt around the centromeres.
662 ONT reads were aligned to the TAIR9 reference genome sequence via Minimap2 v2.1-r761 [71].
663 Mappings were converted into BED files with bedtools v2.26 [72]. The alignments were evaluated
664 for the ends of mapping reads, and these ends were quantified in genomic bins of 100 bp using a
665 dedicated tool designated Assembly Error Finder (AEF) v0.12 (Table 2) with default parameters.
666 Neighboring regions with high numbers of alignment ends were grouped if their distance was

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

694 search was performed (contig versus reference sequences) with the same parameters as for the
695 identification of T-DNA reads. These BLAST results were mapped to the sequence (contig/read) as
696 follows: BLAST hits were annotated one after another, sorted by decreasing score. If the overlap of
697 a BLAST hit with a previously annotated one exceeds 10 bp, the second BLAST hit was discarded.
698 For further analysis, reads were mapped back to the assembly. All reads were mapped back to the
699 fragments of the *de novo* assembly, reads containing T-DNA were mapped back to (1) the Canu
700 contigs, (2) "unassembled" contigs and (3) individual reads containing T-DNA sequences. Mapping
701 was performed using Minimap2 [71] with the default options for mapping of ONT sequencing data.
702 To further inspect the chromosome(s) sequences prior to T-DNA insertion, the same analysis was
703 performed using the *A. thaliana* sequences neighboring the T-DNA insertion. A FASTA file was
704 generated that contains these flanking sequences using bedtools [72], reads containing this part of
705 *A. thaliana* sequence (and no T-DNA) are again identified using BLAST, assembled and annotated
706 as described above. Infoseq from the EMBOSS package [75] was used to calculate the length of
707 different sequences in the pipeline.
708 All information is summarized in HTML files containing images; these images display all annotated
709 sequences along with mapped reads and details on the BLAST results. These pages were used for
710 manual inspection and final determination of the insertion structures. For canonical insertions, such
711 as 050B11-At5g64610, the assembled and annotated contig of the partial assembly was sufficient.
712 In more complex cases like GK-038B07, the exact insertion structure was not clear based on the
713 assembled contigs or based on sections of the *de novo* assembly. If, based on the read mappings
714 shown in the visualization, the partial assembly looked erroneous (many partial mappings),
715 individual reads were used for the determination of the insertion structure. These individual reads
716 were also considered for clearer determination of exact T-DNA positions, if these positions were not
717 clear from contigs / sections. This was often the case for head-to-tail configurations of T-DNA
718 arrays, where one of the T-DNAs was represented by sequence of low quality and could not be
719 annotated (and led to misassemblies in assembled contigs). These cases were resolved by
720 identification of reads orientated in the other direction, because then the sequence derived from the

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

748 Sequencing Core Facility of the Center for Biotechnology (Bielefeld University, Bielefeld, Germany)
749 using BigDye terminator v3.1 chemistry (Thermo Fisher) on a 3730XL sequencer. The resulting
750 Sanger sequences were merged using tools from the EMBOSS package [75]. After transferring
751 reverse reads to their reverse complement using revseq, a multiple alignment was generated using
752 MAFFT [79]. One consensus sequence for each amplicon was extracted from the alignments using
753 em_cons with option -plurality 1, and the resulting sequences were submitted to ENA (see
754 Additional file 1 for accession numbers).
755

756 **Analyses of T-DNA free chromosome fusion junctions**

757 The five junction sequences were analyzed by BLAST essentially as described [13]. Briefly,
758 searches were performed against all possible target sequences (*A. tumefaciens*; *A. thaliana*
759 nucleome, plastome and chondrome; T-DNA and vector backbone) using BLASTn default
760 parameters. If the complete query was not covered, the unmatched part of the query sequence was
761 classified as filler and extracted. Subsequently, this sequence was used in a BLAST search with an
762 e-value cutoff of 10, a word-size of 5 and the '-task "blastn-short"' option activated to detect smaller
763 and lower quality hits. If this was not successful (as in GK-909H04), the filler sequence was
764 extended by 10 bases up- and downstream and the procedure described above was repeated.
765
766

767 ## Declarations

768 **Ethics approval and consent to participate**

769 Not applicable
770

771 **Consent for publication**

772 Not applicable
773

774 **Availability of data and materials**

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

775 Sequence read datasets generated and analyzed during this study were made available at ENA
776 under the accession PRJEB35658. Individual run IDs are included in Additional file 1. The Col-0
777 genome sequence assembly of the GABI-Kat Col-0 genetic background (Col-0_GKat-wt) is
778 available at ENA under the accession GCA_905067165.
779
780 **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

781
782
783 **Competing interest**
784 The authors declare that they have no competing interests.
785
786 **Funding**
787 We acknowledge support for the Article Processing Charge by the Open Access Publication Fund
788 of Bielefeld University.
789
790 **Authors' contribution**
791 BP performed DNA extraction and sequencing. BP and NK performed bioinformatic analyses. BP,
792 NK, and BW interpreted the results and wrote the manuscript.
793
794 **Acknowledgements**
795 We are very grateful to the Bioinformatics Resource Facility support team of the CeBiTec and to
796 de.NBI for providing computing infrastructure and excellent technical support. We also thank the
797 Sequencing Core Facility of the CeBiTec for granting access to the ONT infrastructure. Many
798 thanks to Tobias Busche, Christian Rückert, and Jörn Kalinowski for general support related to the
799 ONT sequencing, and to Prisva Viehöver for high quality Sanger sequencing. We thank Andrea

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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

804

805 **Supplementary information**

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

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

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

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

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

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

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

818 **Additional file 8**: Visual overview over all insertions detected.

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

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

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

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

825

826

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

827

828 ## References

829

830 1. Ulker B, Peiter E, Dixon DP, Moffat C, Capper R, Bouche N, Edwards R, Sanders D, Knight
831 H, Knight MR: **Getting the most out of publicly available T-DNA insertion lines.** *The*
832 *Plant Journal* 2008, **56:**665-677.
833 2. OMalley RC, Ecker JR: **Linking genotype to phenotype using the Arabidopsis**
834 **unimutant collection.** *The Plant Journal* 2010, **61:**928-940.
835 3. Fauser F, Roth N, Pacher M, Ilg G, Sanchez-Fernandez R, Biesgen C, Puchta H: **In planta**
836 **gene targeting.** *Proceedings of the National Academy of Sciences of the United States of*
837 *America* 2012, **109:**7535-7540.
838 4. OMalley RC, Barragan CC, Ecker JR: **A user's guide to the Arabidopsis T-DNA insertion**
839 **mutant collections.** In *Methods in Molecular Biology. Volume* 1284. 2015/03/12 edition.
840 Edited by Alonso J, Stepanova A. New York, NY: Humana Press; 2015: 323-342
841 5. Strizhov N, Li Y, Rosso MG, Viehoever P, Dekker KA, Weisshaar B: **High-throughput**
842 **generation of sequence indexes from T-DNA mutagenized Arabidopsis thaliana lines.**
843 *BioTechniques* 2003, **35:**1164-1168.
844 6. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK,
845 Zimmerman J, Barajas P, Cheuk R, et al: **Genome-wide insertional mutagenesis of**
846 **Arabidopsis thaliana.** *Science* 2003, **301:**653-657.
847 7. Puchta H: **Applying CRISPR/Cas for genome engineering in plants: the best is yet to**
848 **come.** *Current Opinion in Plant Biology* 2017, **36:**1-8.
849 8. Smith EF, Townsend CO: **A Plant-Tumor of Bacterial Origin.** *Science* 1907, **25:**671-673.
850 9. Gelvin SB: **Integration of Agrobacterium T-DNA into the Plant Genome.** *Annual Review*
851 *of Genetics* 2017, **51:**195-217.
852 10. Zambryski P, Holsters M, Kruger K, Depicker A, Schell J, Van Montagu M, Goodman H:
853 **Tumor DNA structure in plant cells transformed by A. tumefaciens.** *Science* 1980,
854 **209:**1385-1391.
855 11. Hernalsteens J, Van Vliet F, De Beuckeleer M, Depicker A, Engler G, Lemmers M, Holsters
856 M, Van Montagu M, Schell J: **The Agrobacterium tumefaciens Ti plasmid as a host**
857 **vector system for introducing foreign DNA in plant cells.** *Nature* 1980, **287:**654-656.
858 12. Clough SJ, Bent AF: **Floral dip: a simplified method for Agrobacterium-mediated**
859 **transformation of *Arabidopsis thaliana*.** *The Plant Journal* 1998, **16:**735-743.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

860 13. Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B: **The Structural**
861 **Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand**
862 **Break Repair-Based Insertion Mechanism.** *Molecular Plant* 2015, **8:**1651-1664.
863 14. van Kregten M, de Pater S, Romeijn R, van Schendel R, Hooykaas PJ, Tijsterman M: **T-**
864 **DNA integration in plants results from polymerase-θ-mediated DNA repair.** *Nature*
865 *Plants* 2016, **2:**16164.
866 15. Castle LA, Errampalli D, Atherton TL, Franzmann LH, Yoon ES, Meinke DW: **Genetic and**
867 **molecular characterization of embryonic mutants identified following seed**
868 **transformation in Arabidopsis.** *Molecular Genetics and Genomics* 1993, **241:**504-514.
869 16. Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R: **A comprehensive**
870 **characterization of single-copy T-DNA insertions in the Arabidopsis thaliana genome.**
871 *Plant Molecular Biology* 2003, **52:**161-176.
872 17. Lafleuriel J, Degroote F, Depeiges A, Picard G: **A reciprocal translocation, induced by a**
873 **canonical integration of a single T-DNA, interrupts the HMG-I/Y Arabidopsis thaliana**
874 **gene.** *Plant Physiology and Biochemistry* 2004, **42:**171-179.
875 18. Clark KA, Krysan PJ: **Chromosomal translocations are a common phenomenon in**
876 **Arabidopsis thaliana T-DNA insertion lines.** *The Plant Journal* 2010, **64:**990-1001.
877 19. Min Y, Frost JM, Choi Y: **Gametophytic Abortion in Heterozygotes but Not in**
878 **Homozygotes: Implied Chromosome Rearrangement during T-DNA Insertion at the**
879 **ASF1 Locus in Arabidopsis.** *Molecules and Cells* 2020, **43:**448-458.
880 20. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B: **An *Arabidopsis thaliana* T-**
881 **DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse**
882 **genetics.** *Plant Molecular Biology* 2003, **53:**247-259.
883 21. Kleinboelting N, Huep G, Kloetgen A, Viehoever P, Weisshaar B: **GABI-Kat SimpleSearch:**
884 **new features of the Arabidopsis thaliana T-DNA mutant database.** *Nucleic Acids*
885 *Research* 2012, **40:**D1211-D1215.
886 22. Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden
887 J, Ko C, et al: **A High-Throughput Arabidopsis Reverse Genetics System.** *The Plant Cell*
888 2002, **14:**2985-2994.
889 23. Sussman MR, Amasino RM, Young JC, Krysan PJ, Austin-Phillips S: **The Arabidopsis**
890 **knockout facility at the University of Wisconsin-Madison.** *Plant Physiology* 2000,
891 **124:**1465-1467.
892 24. Li Y, Rosso MG, Viehoever P, Weisshaar B: **GABI-Kat SimpleSearch: an Arabidopsis**
893 **thaliana T-DNA mutant database with detailed information for confirmed insertions.**
894 *Nucleic Acids Research* 2007, **35:**D874-D878.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

895 25. Kleinboelting N, Huep G, Weisshaar B: **Enhancing the GABI-Kat Arabidopsis thaliana T-**
896 **DNA Insertion Mutant Database by Incorporating Araport11 Annotation.** *Plant and Cell*
897 *Physiology* 2017, **58:**e7.
898 26. Cheng CY, Krishnakumar V, Chan A, Thibaud-Nissen F, Schobel S, Town CD: **Araport11:**
899 **a complete reannotation of the Arabidopsis thaliana reference genome.** *The Plant*
900 *Journal* 2017, **89:**789-804.
901 27. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K,
902 Alexander DL, Garcia-Hernandez M, et al: **The Arabidopsis Information Resource**
903 **(TAIR): improved gene annotation and new tools.** *Nucleic Acids Research* 2012,
904 **40:**D1202-D1210.
905 28. Huep G, Kleinboelting N, Weisshaar B: **An easy-to-use primer design tool to address**
906 **paralogous loci and T-DNA insertion sites in the genome of Arabidopsis thaliana.**
907 *Plant Methods* 2014, **10:**28.
908 29. Vukasinovic N, Cvrckova F, Elias M, Cole R, Fowler JE, Zarsky V, Synek L: **Dissecting a**
909 **hidden gene duplication: the Arabidopsis thaliana SEC10 locus.** *PLoS ONE* 2014,
910 **9:**e94077.
911 30. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B: **A**
912 **chromosome-level sequence assembly reveals the structure of the Arabidopsis**
913 **thaliana Nd-1 genome and its gene set.** *PLoS One* 2019, **14:**e0216233.
914 31. Krispil R, Tannenbaum M, Sarusi-Portuguez A, Loza O, Raskina O, Hakim O: **The Position**
915 **and Complex Genomic Architecture of Plant T-DNA Insertions Revealed by 4SEE.**
916 *International Journal of Molecular Sciences* 2020, **21:**ijms21072373.
917 32. Jupe F, Rivkin AC, Michael TP, Zander M, Motley ST, Sandoval JP, Slotkin RK, Chen H,
918 Castanon R, Nery JR, Ecker JR: **The complex architecture and epigenomic impact of**
919 **plant T-DNA insertions.** *PLoS Genetics* 2019, **15:**e1007819.
920 33. Nacry P, Camilleri C, Courtial B, Caboche M, Bouchez D: **Major chromosomal**
921 **rearrangements induced by T-DNA transformation in Arabidopsis.** *Genetics* 1998,
922 **149:**641-650.
923 34. Tax FE, Vernon DM: **T-DNA-associated duplication/translocations in Arabidopsis.**
924 **Implications for mutant analysis and functional genomics.** *Plant Physiology* 2001,
925 **126:**1527-1538.
926 35. Krizkova L, Hrouda M: **Direct repeats of T-DNA integrated in tobacco chromosome:**
927 **characterization of junction regions.** *The Plant Journal* 1998, **16:**673-680.
928 36. Ulker B, Li Y, Rosso MG, Logemann E, Somssich IE, Weisshaar B: **T-DNA-mediated**
929 **transfer of Agrobacterium tumefaciens chromosomal DNA into plants.** *Nature*
930 *Biotechnology* 2008, **26:**1015-1017.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

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.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

967 51. Rajapriya V, Kannan P, Sridevi G, Veluthambi K: **A rare transgenic event of rice with**
968 **Agrobacterium binary vector backbone integration at the right T-DNA border junction.**
969 *Journal of Plant Biochemistry and Biotechnology* 2021.
970 52. Zambryski P, Depicker A, Kruger K, Goodman HM: **Tumor induction by Agrobacterium**
971 **tumefaciens: analysis of the boundaries of T-DNA.** *Journal of Molecular and Applied*
972 *Genetics* 1982, **1:**361-370.
973 53. Huang CY, Ayliffe MA, Timmis JN: **Direct measurement of the transfer rate of**
974 **chloroplast DNA into the nucleus.** *Nature* 2003, **422:**72-76.
975 54. Bock R: **The give-and-take of DNA: horizontal gene transfer in plants.** *Trends in Plant*
976 *Science* 2009, **15:**11-22.
977 55. Feldmann KA: **T-DNA insertion mutagenesis in *Arabidopsis*: mutational spektrum.** *The*
978 *Plant Journal* 1991, **1:**71-82.
979 56. Wei FJ, Kuang LY, Oung HM, Cheng SY, Wu HP, Huang LT, Tseng YT, Chiou WY, Hsieh-
980 Feng V, Chung CH, et al: **Somaclonal variation does not preclude the use of rice**
981 **transformants for genetic screening.** *The Plant Journal* 2016, **85:**648-659.
982 57. Gang H, Liu G, Zhang M, Zhao Y, Jiang J, Chen S: **Comprehensive characterization of T-**
983 **DNA integration induced chromosomal rearrangement in a birch T-DNA mutant.** *BMC*
984 *Genomics* 2019, **20:**311.
985 58. Curtis MJ, Belcram K, Bollmann SR, Tominey CM, Hoffman PD, Mercier R, Hays JB:
986 **Reciprocal chromosome translocation associated with TDNA-insertion mutation in**
987 **Arabidopsis: genetic and cytological analyses of consequences for gametophyte**
988 **development and for construction of doubly mutant lines.** *Planta* 2009, **229:**731-745.
989 59. Bonhomme S, Horlow C, Vezon D, de Laissardiere S, Guyon A, Ferault M, Marchand M,
990 Bechtold N, Pelletier G: **T-DNA mediated disruption of essential gametophytic genes in**
991 **Arabidopsis is unexpectedly rare and cannot be inferred from segregation distortion**
992 **alone.** *Molecular Genetics and Genomics* 1998, **260:**444-452.
993 60. Ruprecht C, Carroll A, Persson S: **T-DNA-induced chromosomal translocations in**
994 **feronia and anxur2 mutants reveal implications for the mechanism of collapsed**
995 **pollen due to chromosomal rearrangements.** *Molecular Plant* 2014, **7:**1591-1594.
996 61. Schmidt C, Fransz P, Ronspies M, Dreissig S, Fuchs J, Heckmann S, Houben A, Puchta H:
997 **Changing local recombination patterns in Arabidopsis by CRISPR/Cas mediated**
998 **chromosome engineering.** *Nature Communications* 2020, **11:**4418.
999 62. Spealman P, Burrell J, Gresham D: **Inverted duplicate DNA sequences increase**
1000 **translocation rates through sequencing nanopores resulting in reduced base calling**
1001 **accuracy.** *Nucleic Acids Research* 2020, **48:**4940-4945.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1002 63. Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B: **High Contiguity De Novo**
1003 **Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long**
1004 **Read Sequencing.** *Genes* 2020, **11:**E274.
1005 64. Belmann P, Fischer B, Kruger J, Prochazka M, Rasche H, Prinz M, Hanussek M, Lang M,
1006 Bartusch F, Glassle B, et al: **de.NBI Cloud federation through ELIXIR AAI.**
1007 *F1000Research* 2019, **8:**842.
1008 65. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and**
1009 **accurate long-read assembly via adaptive k-mer weighting and repeat separation.**
1010 *Genome Research* 2017, **27:**722-736.
1011 66. Pucker B, Ruckert C, Stracke R, Viehover P, Kalinowski J, Weisshaar B: **Twenty-Five**
1012 **Years of Propagation in Suspension Cell Culture Results in Substantial Alterations of**
1013 **the Arabidopsis Thaliana Genome.** *Genes* 2019, **10:**671.
1014 67. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly**
1015 **from long uncorrected reads.** *Genome Research* 2017, **27:**737-746.
1016 68. Pucker B, Holtgräwe D, Rosleff Sörensen T, Stracke R, Viehöver P, Weisshaar B: **A De**
1017 **Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession**
1018 **Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny.** *PLoS ONE*
1019 2016, **11:**e0164321.
1020 69. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.**
1021 *Journal of Molecular Biology* 1990, **215:**403-410.
1022 70. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J:
1023 **ALLMAPS: robust scaffold ordering based on multiple maps.** *Genome Biology* 2015,
1024 **16:**3.
1025 71. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018,
1026 **34:**3094-3100.
1027 72. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic**
1028 **features.** *Bioinformatics* 2010, **26:**841-842.
1029 73. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL:
1030 **BLAST+: architecture and applications.** *BMC Bioinformatics* 2009, **10:**421.
1031 74. **Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences**
1032 [https://github.com/lh3/seqtk]
1033 75. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open**
1034 **Software Suite.** *Trends in Genetics* 2000, **16:**276-277.
1035 76. Li H: **Minimap and miniasm: fast mapping and de novo assembly for noisy long**
1036 **sequences.** *Bioinformatics* 2016, **32:**2103-2110.

bioRxiv preprint doi: https://doi.org/10.1101/2021.03.03.433755; this version posted March 4, 2021. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1037 77. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
1038 R: **The Sequence Alignment/Map format and SAMtools.** *Bioinformatics* 2009, **25:**2078-
1039 2079.
1040 78. Pucker B, Brockington SF: **Genome-wide analyses supported by RNA-Seq reveal non-**
1041 **canonical splice sites in plant genomes.** *BMC Genomics* 2018, **19:**980.
1042 79. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7:**
1043 **improvements in performance and usability.** *Molecular Biology and Evolution* 2013,
1044 **30:**772-780.

1045