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

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

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

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

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

+ authors contributed equally

* corresponding author: Bernd Weisshaar

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

Abstract

Background

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

Results

We sequenced the genomes of 14 *A. thaliana* GABI-Kat T-DNA insertion lines, which eluded flanking sequence tag-based attempts to characterize their insertion loci, with Oxford Nanopore Technologies (ONT) long reads. Complex T-DNA insertions were resolved and 11 previously unknown T-DNA loci identified, resulting in about 2 T-DNA insertions per line and suggesting that this number was previously underestimated. T-DNA mutagenesis caused fusions of chromosomes along with compensating translocations to keep the gene set complete throughout meiosis. Also, an inverted duplication of 800 kbp was detected. About 10% of GABI-Kat lines might be affected by chromosomal rearrangements, some of which do not involve T-DNA. Local assembly of selected reads was shown to be a computationally effective method to resolve the structure of T-DNA insertion loci. We developed an automated workflow to support investigation of long read data from

T-DNA insertion lines. All steps from DNA extraction to assembly of T-DNA loci can be completed within days.

Conclusion

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

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

Background

T-DNA insertion lines contributed substantially to the high-value knowledge about the functions of plant genes. This knowledge has been produced by the plant research community on the basis of gene structures predicted from genome sequences. T-DNA insertional mutagenesis emerged as an effective mechanism for the generation of knock-out alleles for use in reverse genetics and targeted

54 gene function search [1, 2]. Since targeted integration of DNA into plant genomes via homologous
55 recombination was at least technically challenging [3], large collections of sequence-indexed T-
56 DNA integration lines with random insertion sites were used to provide knock-out alleles for the
57 majority of genes [4]. Knowledge about the inserted sequences is an advantage over other
58 mutagenesis methods, because localization of the insertion within the mutagenized genome based
59 on the generation of flanking sequence tags (FSTs) is possible [5, 6]. CRISPR/Cas technology now
60 offers technically feasible alternatives for access to mutant alleles for reverse genetics [7].
61 However, thousands of T-DNA insertion mutants have been characterized and represent today the
62 main or reference mutant allele for (lack of) a given gene function.
63 *Agrobacterium tumefaciens* is a Gram-negative soil bacterium with the ability to transfer DNA into
64 plant cells and to integrate this T-DNA stably and at random positions into the nuclear genome [8,
65 9]. The tumor inducing (Ti) plasmid, which is naturally occurring in *Agrobacteria* and enables them
66 to induce the formation of crown galls in plants, contains the T-DNA that is transferred into plant
67 cells [10]. The T-DNA is enclosed by 25 bp long imperfect repeats that were designated left (LB)
68 and right border (RB) [9]. The T-DNA sequence between LB and RB can be modified to contain
69 resistance genes for selection of successfully transformed plants [11]. T-DNAs from optimized
70 binary plasmids are transformed into *A. thaliana* plants via floral dip to generate stable lines [12]. T-
71 DNA transfer into the nucleus of a plant cell is supported by several VIR proteins which are, in the

72 biotechnologically optimized system, encoded on a separate helper plasmid. It is assumed that host
73 proteins are responsible for integration of the T-DNA into the genome, most likely as a DNA double
74 strand into a double strand break (DSB) using host DNA repair pathways and DNA polymerase
75 theta [9, 13, 14]. T-DNA integration resembles DNA break repair through non-homologous end-
76 joining (NHEJ) or microhomology-mediated end-joining (MMEJ) and is often accompanied by the
77 presence of filler DNA or microhomology at both T-DNA::genome junctions [9, 13]. Chromosomal
78 inversions and translocations are commonly associated with T-DNA insertions in numerous plant
79 species, not only in *A. thaliana* [15-21], suggesting that often more than just one DSB is associated
80 with T-DNA integration [9].
81 The most important collections of T-DNA lines for the model plant *Arabidopsis thaliana* are SALK
82 (150,000 lines) [6], GABI-Kat (92,000 lines) [22, 23], SAIL (54,000 lines) [24], and WISC (60,000
83 lines) [25]. In total, over 700,000 insertion lines have been constructed [4]. GABI-Kat lines were
84 generated through the integration of a T-DNA harboring a sulfadiazine resistance gene for selection
85 of transformed lines [22]. Additionally, the T-DNA contains a 35S promoter at RB causing
86 transcriptional up-regulation of plant genes next to the integration site if the right part of the T-DNA
87 next to RB stays intact during integration [1]. Integration sites were predicted based on FSTs and
88 allowed access to knock-out alleles of numerous genes. At GABI-Kat, T-DNA insertion alleles were
89 confirmed by an additional "confirmation PCR" using DNA from the T2 generation [26] prior to the

90 release of a mutant line and donation of the line to the Nottingham Arabidopsis Stock Centre
91 (NASC). Researchers could identify available T-DNA insertion lines via SimpleSearch on the GABI-
92 Kat website [27]. Since 2017, SimpleSearch uses Araport11 annotation data [28]. Araport11 is
93 based on the *A. thaliana* Col-0 reference genome sequence from TAIR9 which includes about 96
94 annotated gaps filled with Ns as a sign for unknown sequences [29], among them the centromers
95 and several gaps in the pericentromeric regions.
96 The prediction of integration sites based on bioinformatic evaluations using FST data does often
97 not reveal the complete picture. Insertions might be masked from FST predictions due to truncated
98 borders [13], because of repetitive sequences or paralogous regions in the genome [30], or even
99 lack of the true insertion site in the reference sequence used for FST mapping [31, 32]. Also,
100 confirmation by sequencing an amplicon that spans the predicted insertion site at one of the two
101 expected T-DNA::genome junctions is not fully informative. Deletions and target site duplications at
102 the integration site can occur and can only be detected by examining both borders of the inserted
103 T-DNA [13]. In addition, several studies have reported more complex T-DNA insertions in
104 association with large deletions, insertions, inversions or even chromosomal translocations [13, 18,
105 19, 21, 33, 34]. A relation of T-DNA insertion mutagenesis to chromosomal translocations has not
106 only been reported for *A. thaliana*. Also for transgenic rice (*Oryza sativa*) [35] and transgenic birch
107 (*Betula platyphylla* x *B. pendula*) [36] changes at the genome level have been reported. Binary

108 vector backbone (BVB) sequences have been detected at insertion sites [19, 37] as well as
109 fragments of *A. tumefaciens* chromosomal DNA [38]. Recombination between two T-DNA loci was
110 described as a mechanism for deletion of an enclosed genomic fragment [39].
111 Chromosomal rearrangements can occur during the repair of DSBs, e.g. via microhomology-
112 mediated end joining or non-allelic homologous recombination [reviewed by 40, 41]. Evidently,
113 regions with high sequence similarity like duplications are especially prone to chromosomal
114 rearrangements [40].
115 While usually one T-DNA locus per line was identified by FSTs, the number of T-DNA insertion loci
116 per line is usually higher. For GABI-Kat, it was estimated that about 50% of all lines (12,018 of
117 21,049 tested, according to numbers from the end of 2019) display a single insertion locus. This
118 estimation is based on segregation analyses using sulfadiazine resistance as a selection marker
119 [22]. Other insertion mutant collections report similar numbers [4]. The average number of T-DNA
120 insertions per line was reported to be about 1.5, but this is probably an underestimation since the
121 kanamycin and BASTA selection marker genes applied to determine the numbers are known to be
122 silenced quite often [4]. For these reasons, it is required that insertion mutants (similar to mutants
123 created by e.g. chemical mutagenesis) are backcrossed to wild type prior to phenotyping a
124 homozygous line.

125 The FSTs produced for the different mutant populations by individual PCR and Sanger-sequencing
126 allowed usually access to a single T-DNA insertion locus per line, although for GABI-Kat there are
127 several examples with up to three confirmed insertion loci based on FST data (e.g. line GK-011F01,
128 see [27]). This leaves a significant potential of undiscovered T-DNA insertions in lines already
129 available at the stock centers, which has been exploited by the group of Joe Ecker by applying
130 TDNA-Seq (Illumina technology) to the SALK and a part of the GABI-Kat mutant populations
131 (unpublished, all data available from GenBank/ENA). Essentially the same technology has
132 subsequently been used to set up a sequence indexed insertion mutant library of *Chlamydomonas*
133 *reinhardtii* [42]. With the fast development of new DNA sequencing technologies, the
134 comprehensive characterization of T-DNA insertion lines comes into reach.
135 Several studies already harnessed high throughput sequencing technologies to investigate T-DNA
136 insertion and other mutant lines [43, 44]. Oxford Nanopore Technologies (ONT) provides a cost-
137 effective and fast approach to study *A. thaliana* genomes, since a single MinION/GridION Flow Cell
138 delivers sufficient data to assemble one genotype [45]. Here, we present a method to fully
139 characterize T-DNA insertion loci and additional genomic changes of T-DNA insertion lines through
140 ONT long read sequencing. We selected 14 lines that contain confirmed T-DNA insertion alleles.
141 The first border or first T-DNA::genome junction has been confirmed by sequencing an amplicon
142 across one junction, but the insertion eluded characterization of the second T-DNA::genome

143 junction. We refer to the additional T-DNA::genome junction that is expected to exist after
144 confirmation of one T-DNA::genome junction as "2nd border". Within this set of lines that is enriched
145 for unusual relations between the two expected T-DNA::genome junctions, we detected several
146 chromosome fragment or chromosome arm translocations, a duplication of 800 kbp and also an
147 insertion of DNA from the chloroplast (plastome), all related to T-DNA insertion events. The results
148 clearly demonstrate the importance of characterizing both T-DNA::genome junctions for reliable
149 selection of suitable alleles for setting up genotype/phenotype relations for gene function search. In
150 parallel to data evaluation, we created an automated workflow to support long-read-based analyses
151 of T-DNA insertion lines and alleles.
152
153

154 Results

155 In total, 14 GABI-Kat T-DNA insertion lines (Table 1, Additional file 1) were selected for genomic
156 analysis via ONT long read sequencing. This set of lines was selected based on prior knowledge
157 which indicated that the insertion locus addressed in the respective line was potentially unusual.
158 The specific feature used for selection was the (negative) observation that creation of confirmation
159 amplicons which span the T-DNA::genome junction failed for one of the two junctions, operationally

160 that means that the 2nd border could not be confirmed. T-DNA insertion loci in the selected lines

161 were assessed by *de novo* assembly of the 14 individual genome sequences, and by a

162 computationally more effective local assembly of selected reads.

163

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

Line ID[a]	number of insertions				Summary of observation
	FST pred.	PCR-conf.	total found	ONT found	
GK-038B07	2	0 (1)[b]	4[c]	3	FST-predicted insertion in Chr5 is part of a fusion of Chr3 and Chr5, 2 T-DNA arrays detected at translocation fusion points of which one contains in addition an inversion of ~2 Mbp fused with another T-DNA array, additional insertion of a complex T-DNA array in Chr1.
GK-089D12	1	1	2	1	Fusion of Chr3 and Chr5, FSTs are derived from the single DNA::genome junction that contains LB, both translocation fusions contain mostly canonical T-DNAs, failure to confirm 2nd T-DNA::genome junction explained by shortened T-DNA.
GK-430F05	1	1	3	2	Three T-DNA insertion sites, two complex T-DNA arrays with one containing more than 5 kbp BVB[d], the confirmed T-DNA insertion in Chr3 is complex, no theoretical explanation for failure of confirmation PCR at 2nd DNA::genome junction, one insertion in the pericentromeric region of (probably) Chr4 containing a rearranged RB region.
GK-654A12[d]	1	1	2[c]	1	Fusion of Chr1 and Chr4 with a complex T-DNA array that also contains BVB, Chr1-part of the fusion predicted by FST, the compensating fusion of Chr4 and Chr1 does not contain T-DNA at the translocation fusion point, translocation explains failure to confirm 2nd T-DNA::genome junction, additional insertion of a T-DNA array in Chr2 with a 162 bp duplicated inversion at the integration site.

GK-767D12	1	1	2	1	Large segmental duplication and inversion at the predicted insertion site, a long T-DNA array also containing BVB present at the southern end of the segmental duplication, inversion explains failure to confirm 2nd T-DNA::genome junction, the northern fusion point of the inverted segmental duplication does not contain T-DNA, another canonical T-DNA insertion in Chr2 but with 4 Mbp distance.
GK-909H04	1	1	3	2	The predicted T-DNA insertion contains an inverted duplication of 20 kbp at the integration site explains why the 2nd T-DNA::genome junction could not be confirmed, integration of 652 bp derived from the plastome at the northern fusion point of the duplication but no T-DNA, two additional canonical T-DNA insertions in Chr1 and Chr2.
GK-947B06	1	1	2	1	Complex T-DNA array insertion of 3 T-DNAs at the predicted insertion site in Chr1, no theoretical explanation for failure of the confirmation PCR at the 2nd DNA::genome junction, additional complex insertion containing 2 T-DNAs and BVB on Chr2.

165 [a] Lines with newly detected insertions, all lines are listed in Additional file 2.

166 [b] The T-DNA insertion used for line selection was a false positive case not detected by ONT seq.

167 [c] One locus in a line with a chromosome translocation causes the FSTs from this single locus to
168 map to two places in the reference sequence. These two places might also be the source of
169 potentially existing FST from the compensating chromosome fusion. If the compensating chro-
170 mosome fusion does not contain a T-DNA, the number of insertions is lower than expected.

171 [d] BVB, binary vector backbone

172

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

178 were not revealed by previous FST based attempts to detect T-DNA insertions (Table 1, Additional
179 file 2 and 3). In case of GK-038B07, the lack of re-detection of the expected insertion allele of
180 At4g19510 was explained by a PCR template contamination during the initial confirmation, the line
181 that contains the real insertion (source of the contamination) is most probably GK-159D11. A DNA
182 template contamination is also likely in case of GK-040A12 where the insertion allele of At1g52720
183 expected from the confirmation was also not found in the ONT data. At least, the error detected fits
184 to the selection criteria, because the 2nd T-DNA::genome junction can obviously not be detected if
185 the T-DNA insertion allele as such is not present in the line.
186 Six of the 14 lines studied by ONT whole genome sequencing displayed chromosomal
187 rearrangements. To visualize these results, we created ideograms of the five *A. thaliana*
188 chromosomes with a color code for each chromosome. The colors allow to visually perceive
189 information on chromosome arm translocations, and the changeover points indicate presence or
190 absence of T-DNA sequences.
191

192 **Chromosome fusions**

193 In four lines, fusions of different chromosomes were detected. These fusions result from
194 chromosome arm translocations which were, in all four cases, compensated within the line by
195 reciprocal translocations.

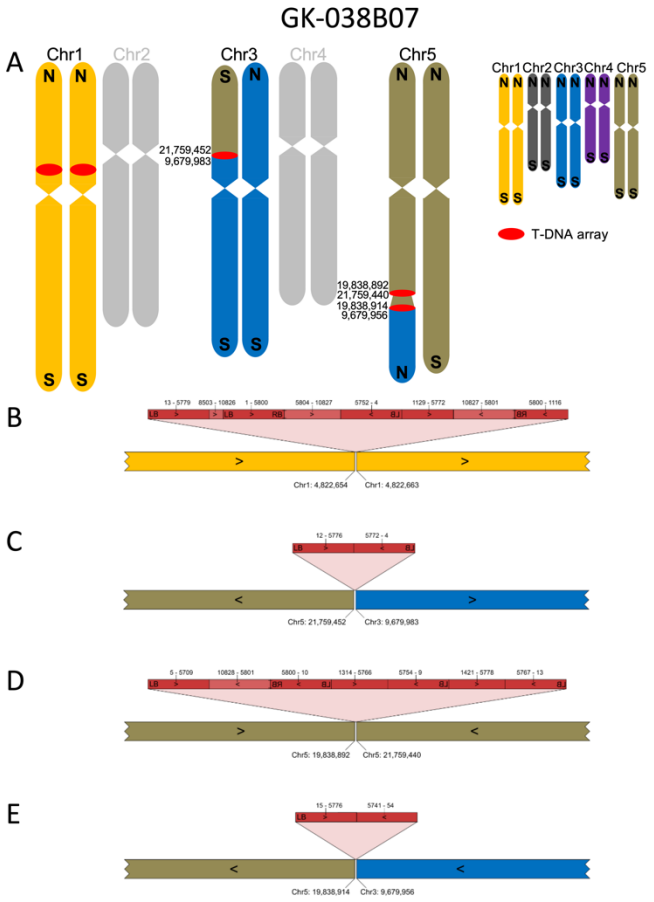
196 The T-DNA insertion on chromosome 5 (Chr5) of GK-038B07 is part of a complex chromosome

197 arm translocation (Fig. 1). A part of Chr5 is fused to Chr3, the replaced part of Chr3 is fused to an

198 inversion of 2 Mbp on Chr5. This inversion contains T-DNAs at both ends, one of which is the

199 insertion predicted by FSTs.

200



201

202 **Fig. 1**: Organization of the nuclear genome of GK-038B07 with a focus on translocations, inversions

203 and T-DNA structures.

204 A: Ideograms of the chromosomes display the reciprocal fusion of Chr3 and Chr5 as well as a 2

205 Mbp inversion between two T-DNA arrays at the fusion sites; numbers indicate end points of

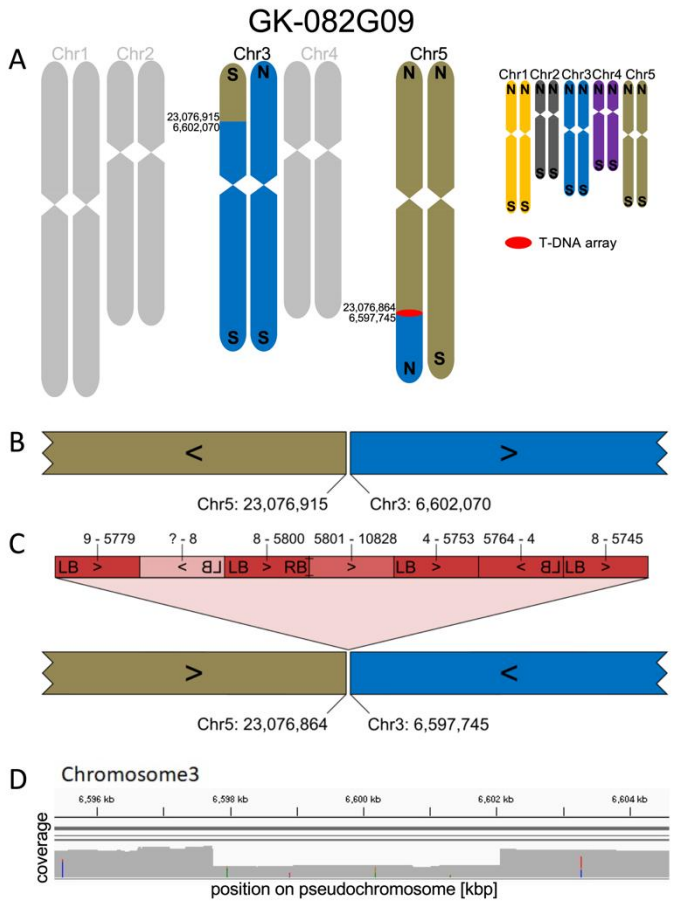
206 pseudochromosome fragments according to TAIR9. Chromosomes depicted in grey match the
207 Col-0 sequence. The insert displays the color codes used for all five chromosomes; N, northern end
208 of chromosome; S, southern end of chromosome. B-E: Visualization of the four T-DNA insertion loci
209 of GK-038B07 resolved by local assembly. LB and RB, T-DNA left and right border; dark red, *bona*
210 *fide* T-DNA sequences located between the borders; light red, sequence parts from the binary
211 vector backbone (BVB); numbers above the red bar indicate nucleotide positions with position 1
212 placed at the left end of LB in the binary vector which makes position 4 the start of the transferred
213 DNA [13]; numbers below the colored bars indicate pseudochromosome positions according to
214 TAIR9.
215

216 For line GK-082G09, two FST predictions had been generated and one FST lead to the prediction
217 of an insertion at Chr3:6,597,745 which was confirmed by PCR. Confirmation of the expected
218 corresponding 2nd border failed. Another FST-based prediction at Chr5:23,076,864 was not
219 addressed by PCR. ONT sequencing confirmed both predictions (Fig. 2). There is only one
220 complex insertion consisting of multiple T-DNA copies and BVB in GK-082G09 that fuses the south
221 of Chr5 to an about 6.6 Mbp long fragment from the north of Chr3 (Fig. 2C). This translocation is
222 compensated by a fusion of the corresponding parts of both chromosomes without a T-DNA at the
223 fusion point. The second fusion point of Chr3 and Chr5, that was detected in the *de novo* assembly
224 of the genome sequence of GK-082G09, was validated by generating and sequencing a PCR

225 amplicon spanning the translocation fusion point (Fig. 2B, see Additional file 1 for

226 sequences/accession numbers and Additional file 4 for the primer sequences).

227



228

229 **Fig. 2**: Organization of the nuclear genome of GK-082G09 with a focus on translocations,

230 inversions and T-DNA structures. A-C: For a description of the figure elements see legend to Fig. 1.

231 D: Read coverage depth analyses of the region of Chr3 that is involved in the fusions which

232 confirms a deletion of about 4 kbp from Chr3. The reads that cover the deleted part were derived

233 from the wild type allele present in the segregating population (see methods).

234

235 Line GK-089D12 harbors two T-DNA insertions (Fig. 3) and both were predicted by FSTs, one in

236 Chr3 and one in Chr5. Since fragments of Chr3 and Chr5 are exchanged in a reciprocal way with

237 no change in sequence direction (southern telomeres stay at the southern ends of the

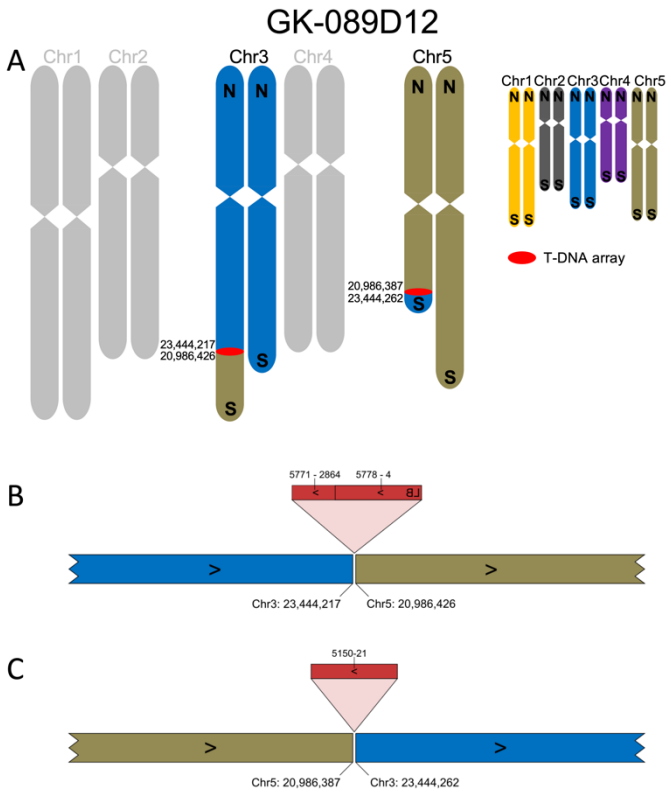
238 chromosomes), PCR confirmation would have usually resulted in "fully confirmed" insertion alleles.

239 Only long read sequencing allowed to determine the involvement of translocations. The line was

240 studied because the shortened T-DNA at 089D12-At5g51660-At3g63490 (Fig. 3B, see Additional

241 file 3 for designations of insertions) caused failure of formation of the confirmation amplicon.

242



243

244 **Fig. 3**: Organization of the nuclear genome of GK-089D12 with a focus on translocations,

245 inversions and T-DNA structures. A-C: For a description of the figure elements see legend to Fig. 1.

246

247 FSTs from line GK-654A12 indicated a T-DNA insertion on Chr1. ONT sequencing revealed a

248 translocation between Chr1 and Chr4 that explained failure to generate the confirmation amplicon

249 at the 2nd border (Fig. 4). The southern arms of Chr1 and Chr4 are exchanged, with a T-DNA array

250 inserted at the fusion point of the new chromosome that contains CEN1 (centromere of Chr1). The

251 fusion point of the new chromosome that contains CEN4 does not contain T-DNA sequences. Also

252 this T-DNA-free fusion point (designated 654A12-FCAALL-0-At1g45688) was validated by

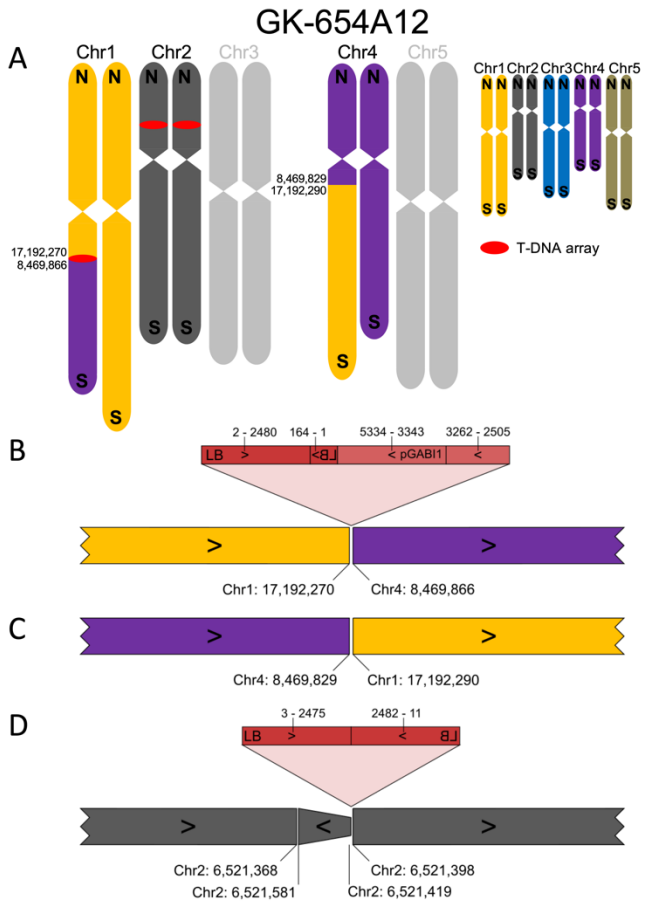
253 generating and sequencing a PCR amplicon which spanned the fusion site (Fig. 4C, Additional files

254 1 and 4). The T-DNA array at 654A12-At1g45688-FCAALL contains BVB sequences, interestingly

255 as an independent fragment and not in an arrangement that is similar to the binary plasmid

256 construction which provided the T-DNA (Fig. 4B).

257



258

259 **Fig. 4**: Organization of the nuclear genome of GK-654A12 with a focus on translocations, inversions

260 and T-DNA structures. A-C: For a description of the figure elements see legend to Fig. 1. See

261 Additional file 2 for an explanation of pGABI1. D: The T-DNA insertion in Chr2 is associated with a

262 small duplicated inversion of about 160 bp as already described for a fraction of all T-DNA::genome

263 junctions [13].

264

265 **Intrachromosomal rearrangements and a large duplication**

266 For line GK-433E06 the FST data indicated four insertions, one T-DNA insertion (433E06-

267 At1g73770-F9L1) at Chr1:27,742,275 has been confirmed by amplicon sequencing. ONT

268 sequencing revealed an intrachromosomal translocation that exchanged the two telomeres of Chr1

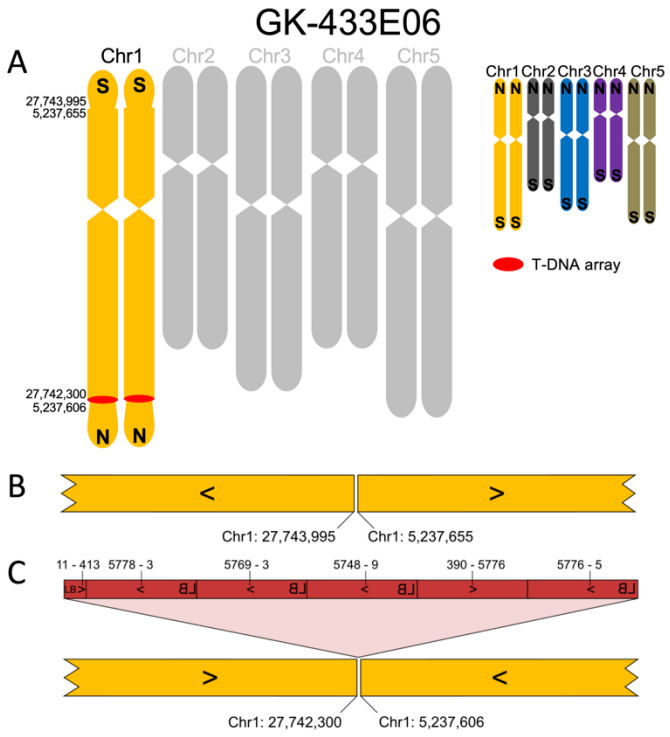
269 together with about 5 Mbp DNA. The FSTs that indicated two T-DNA insertions in chromosome 1

270 were derived from one T-DNA array (Fig. 5). Once more, the compensating fusion point,

271 designated 433E06-F9L1-0-At1g73770, does not contain T-DNA sequences which was validated

272 by amplicon sequencing (Fig. 5C, Additional files 1 and 4).

273



274

275 **Fig. 5**: Organization of the nuclear genome of GK-433E06 with a focus on translocations, inversions

276 and T-DNA structures. A-C: For a description of the figure elements see legend to Fig. 1.

277

278 In line GK-767D12, a large duplication of a part of Chr2 that covers about 800 kbp was detected

279 (Fig. 6). The duplication is apparent from read coverage analyses (Fig. 6B) based on read mapping

280 against the TAIR9 reference genome sequence (Col-0) which was performed for all lines studied

281 (Additional file 5). The duplicated region is inserted in reverted orientation (inversion) next to the T-

282 DNA insertion 767D12-At2g19210-At2g21385. This insertion was predicted by an FST at

283 Chr2:8,338,072 and has been confirmed by PCR, the 2nd border confirmation for the T-DNA

284 insertion failed because of reversed orientation. The other end of the duplicated inversion of Chr2 is

285 fused to Chr2:8,338,361 (designated 767D12-At2g21385-0-At2g19210) without T-DNA sequences

286 (Fig. 6D). Also this T-DNA-free fusion point was validated by sequencing a PCR amplicon which

287 spanned the fusion site (Additional files 1 and 4). The T-DNA array at 767D12-At2g19210-

288 At2g21385 is the largest we detected in this study. It consists of 8 almost complete T-DNA copies

289 arranged in diversified configurations and includes a large BVB fragment (Fig. 6C).

290

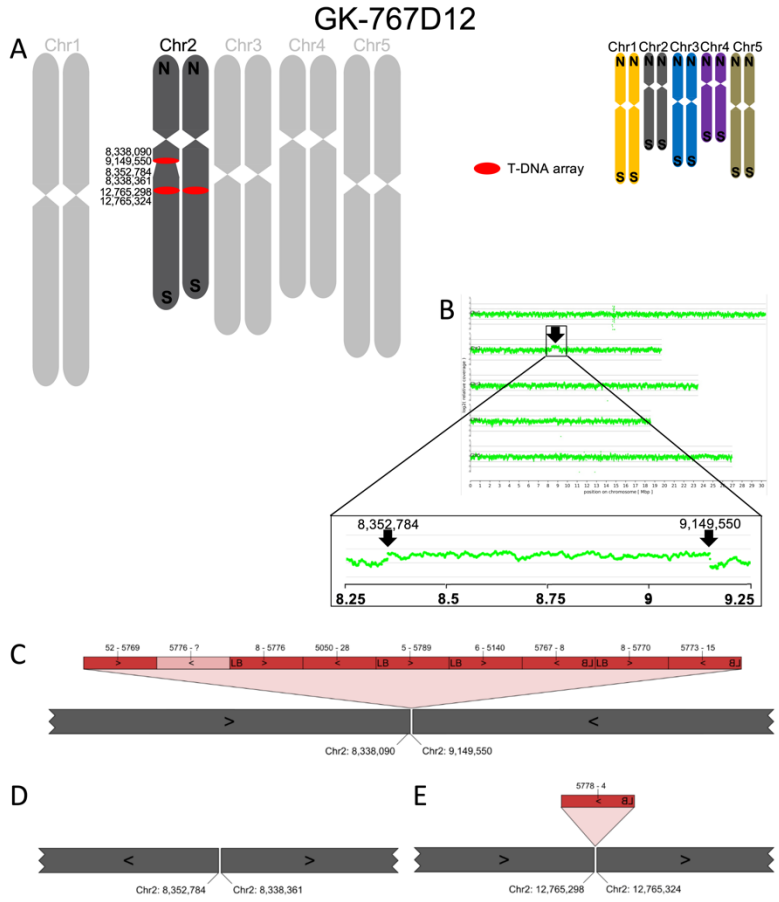
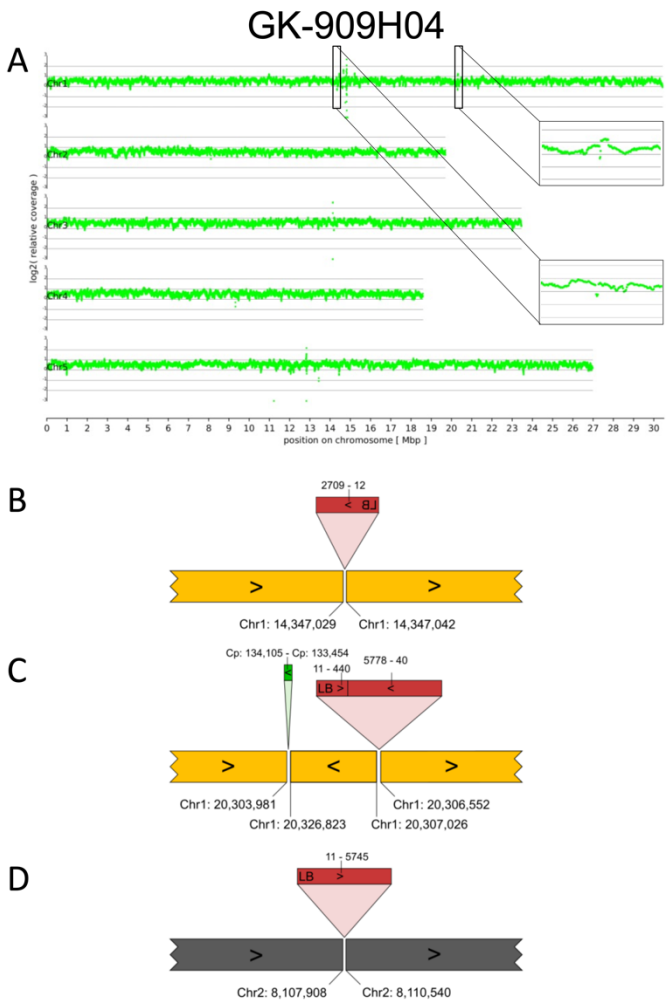


Fig. 6: Organization of the nuclear genome of GK-767D12 with a focus on translocations, inversions and T-DNA structures. A-E: For a description of the figure elements see legend to Fig. 1. B: Results from a read coverage depth analysis are depicted that revealed a large duplication compared to the TAIR9 Col-0 reference sequence. We used read coverage depth data to decide for the selection of the zygosity of the insertions and rearrangements displayed for Chr2 in the ideograms in panel A.

FST analyses detected only one T-DNA insertion in line GK-909H04. This insertion, designated 909H04-At1g54390, had been confirmed by PCR but failed for the 2nd border. ONT sequencing revealed an inverted duplication of about 20 kbp next to the T-DNA insertion site (Fig. 7). The

302 fusion between this inverted duplication and the remaining part of Chr1 does not contain T-DNA
303 sequences, but a 652 bp fragment derived from the plastome (Fig. 7C). The cpDNA insertion was
304 validated by generating and sequencing a PCR amplicon spanning the insertion and both junctions
305 to the genome (Additional files 1 and 4). ONT sequencing also revealed an additional insertion of a
306 truncated T-DNA (909H04-At1g38212 at about 14.3 Mbp of Chr1) which is in the pericentromeric
307 region not far from CEN1 (CEN1 is located at 15,086,046 to 15,087,045 and marked in the TAIR9
308 reference sequence by a gap of 1,000 Ns). Initial analyses indicated that this insertion might be
309 associated with a deletion of about 45 kbp. However, the predicted deletion was not indicated by
310 the read coverage depth analyses, and the region is rich in TEs (also At1g38212 is annotated as
311 "transposable element gene"). We assembled a new genome sequence of the Col-0 wild type used
312 at GABI-Kat (assembly designated Col-0_GKat-wt, see below) and studied the structure of 909H04-
313 At1g38212 on the basis of this assembly. The results indicated that the deletion predicted on the
314 basis of the TAIR9 assembly is a tandemly repeated sequence in TAIR9 which is differently
315 represented in Col-0_GKat-wt (Additional file 6). The 3'-end of an example read from line GK-
316 909H04 maps continuously to Col-0_GKat-wt and also to a sequence further downstream in TAIR9.
317 The evidence collected clearly shows that there are only 13 bp deleted at the T-DNA insertion at
318 14.3 Mbp of Chr1 (Fig. 7D), and that the initially predicted deletion is caused by errors in the TAIR9
319 assembly in this pericentromeric region.

320



321

322 **Fig. 7**: Organization of the nuclear genome of GK-909H04 with a focus on insertions and T-DNA
323 structures. A-D: For a description of the figure elements see legend to Fig. 6. A: The read coverage
324 depth plot includes two zoom-in enlargements of the regions at 14.3 and 20.3 Mbp of Chr1. These
325 two display variable coverage in the region of the truncated T-DNA insertion 909H04-At1g38212
326 (see text), and increased coverage next to the T-DNA insertion 909H04-At1g54390. C: Duplicated
327 inversion detected in the local assembly of GK-909H04 which is also supported by read coverage
328 depicted in Panel A. Green block, sequence part from the plastome (cpDNA).

329

330 The six junctions (five sequences) that contained no T-DNA, three from compensating chromosome
331 fusions, one from the 800 kbp inversion and two at both ends of the cpDNA insertion (see
332 Additional file 3), were analyzed for specific features at the junctions. The observations made were
333 fully in line with what has already been described for T-DNA insertion junctions: some short filler
334 DNA and microhomology was found (Additional file 7). A visual overview over the T-DNA insertion
335 structures of all 14 lines, including those not displaying chromosomal rearrangements, is presented
336 in Additional file 8.
337

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

339 As mentioned above, 11 T-DNA insertion loci were newly detected in 7 of 14 lines studied,
340 indicating that these were missed by FST-based studies (Table 1, Additional file 2 and 3). The
341 primer annealing sites for FST generation at LB seem to be present in all 11 T-DNA insertions only
342 found by ONT sequencing. Analysis of the data on T-DNA::genome junctions summarized in
343 Additional file 3 revealed that a majority of the T-DNA structures have LB sequences at both T-
344 DNA::genome junctions (14 of 26). The bias for the T-DNA::genome junctions involving LB is
345 increased by the fact that several of the RB junctions were truncated, and also by some other
346 junctions which involve BVB sequences. True T-DNA::genome junctions involving intact RB were
347 not in the dataset, and in 14 out of 26 cases an internal RB::RB fusion was detected.

348 While about 40% (8 to 10 of 26, depending on judgement of small discontinuous parts) of the
349 insertions contain a single T-DNA copy (referred to here as "canonical" insertions), some of which
350 even further truncated and shortened, there are often cases of complex arrays of T-DNA copies
351 inserted as T-DNA arrays. We observed a wide variety of configurations of the individual T-DNA
352 copies within the complex arrays. In six out the 26 cases BVB sequences were detected, in the
353 case of 038B07-At1g14080, 082G09-At5g57020-At3g19080, 430F05-At4g23850 and 947B06-
354 T7M7 even almost complete vector sequences.
355

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

357 During the analyses of T-DNA insertion sequences, we frequently faced regions without sequence
358 similarity to any sequence in the *A. thaliana* genome sequence, the sequence of the Ti-plasmid (T-
359 DNA and BVB), or the *A. tumefaciens* genome sequence. Analysis of the read quality (Phred score)
360 in these regions compared to other regions on the same read revealed a quality drop (Fig. 8).
361 Consequently, the number of miscalled bases in these regions is especially high. These miscalls
362 prevent matches in BLAST searches where a perfect match of several consecutive bases is
363 required as seed for a larger alignment. In some cases, entire reads displayed an extremely low
364 quality. Reads that display such locally increased error rates were found in the context of T-DNA
365 array structures which involve head-to-head or tail-to-tail configurations that have the ability to form

366 foldback structures. It is important to note that the second T-DNA unit was sequenced with lower

367 quality independent of the read direction (Fig. 8).

368

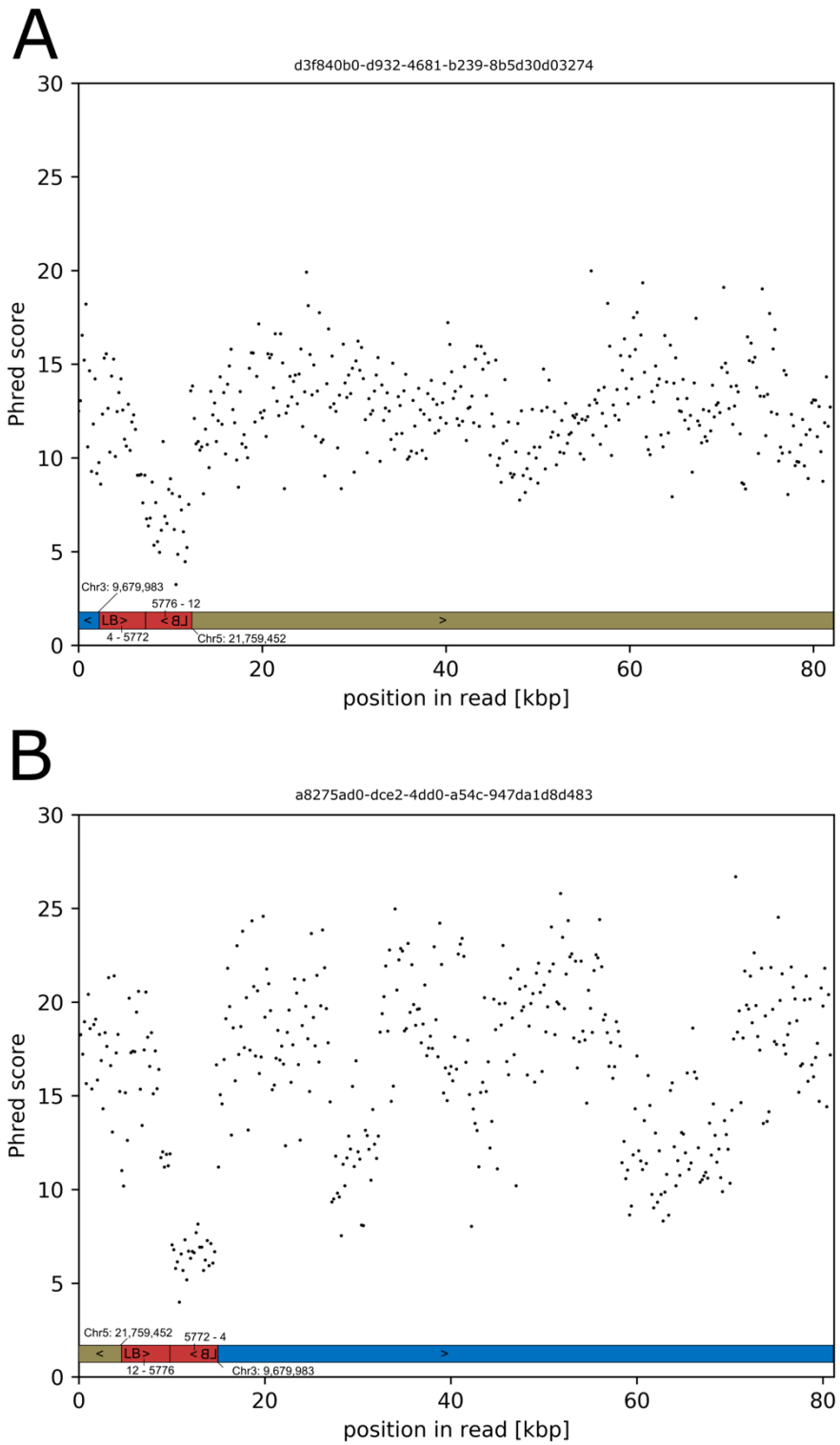


Fig. 8: Decrease of Phred score in ONT reads when running from genomic sequence into a T-DNA array. The insertion used as example is from line GK-038B07 and is shown in Fig. 1C. A/B: Read quality data from two ONT reads with opposite direction. The reads are displayed in 5' to 3'

373 orientation. Green and blue bars indicate the read parts matching Chr3 and Chr5, respectively. The
374 red bar indicates the position of the T-DNA array which is arranged as inverted repeat. Read IDs of
375 the two examples are shown above each panel. The read quality drop to a Phred score of about 7
376 occurs when the second part of the inverted repeat is sequenced.
377

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

379 As mentioned above for the insertion allele 909H04-At1g38212, the detection of rearrangements in
380 the insertion lines is not only dependent on the quality of the reads from or assemblies of the
381 genomes of the lines studied. The results are also dependent on the correctness of the reference
382 sequence. While the quality of the Col-0 reference sequence (the sequence from TAIR9 is still the
383 most recent, see Introduction) is generally of high quality, there are some sequence regions that
384 are not fully resolved. We used a subset of our ONT data, namely ultra-long (> 100 kbp, see
385 Methods) T-DNA free reads, to *de novo* assemble the Col-0 genome sequence. The assembly,
386 designated Col-0_GKat-wt, comprises 35 contigs after polishing and displays an N50 of 14.3 Mbp
387 (GCA_905067165, see Additional file 9). The Col-0_GKat-wt assembly is about 4 Mbp longer than
388 TAIR9 but still does not reach through any of the centromers. Comparison to the TAIR9 sequence
389 indicated that the main gain in assembly length was reached in the pericentromeric regions.
390 Our collection of ONT sequencing datasets from the GABI-Kat lines provides a combined coverage
391 of over 500x for the TAIR9 reference genome sequence of Col-0. In addition to using ultra-long

392 reads for generating an assembly, the reads were also used for identification of potentially
393 problematic regions in the reference sequence. We identified conflicting regions by evaluating read
394 alignments to assemblies and obtained a list of 383 candidate regions (Additional file 10). We
395 compared selected regions against our *de novo* genome assembly and focused first on the locus
396 At1g38212 (at about 14.3 Mbp of Chr1, see Fig. 7). The differences in this region of the TAIR9
397 assembly, which were detected when analyzing the T-DNA insertion allele 909H04-At1g38212
398 (Additional file 6), did show up again. Together with nine other examples selected across all
399 chromosomes, Additional file 11 displays regional comparisons of TAIR9 to Col-0_GKat-wt. The 96
400 gaps containing various numbers of Ns which are reported for TAIR9 are frequently detected
401 (Additional files 10 and 11).
402
403

404 Discussion

405 By sequencing GABI-Kat T-DNA insertion lines with ONT technology, we demonstrate the power of
406 long read sequencing for the characterization of complex T-DNA insertion lines. The complexity of
407 these lines has, at least, four aspects: (i) the number of different insertion loci present in a given
408 line in different regions of the nuclear genome, (ii) the variance of the structures of one or several

409 T-DNA copies appearing at a given insertion locus, (iii) the changes in the genome sequence in the

410 direct vicinity of the T-DNA, and (iv) the changes of chromosome topology related to T-DNA

411 integration.

412

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

414 The average number of T-DNA insertions per *A. thaliana* T-DNA insertion line is assumed to be

415 about 1.5 [4]. However, the insertion lines available at the stock centers like NASC or ABRC list in

416 almost all cases only one insertion per line. In our limited dataset of 14 lines, 11 new insertions

417 were detected among a total of 26, indicating that one should expect an average of about 2

418 insertions per line. The 11 new insertions all contain sufficiently intact LB sequences that should

419 have allowed generation of FSTs. The reason for the lack of detection is probably that the FST data

420 generated at GABI-Kat in total have not reached the saturation level, although several insertions

421 are predicted per line at GABI-Kat [23]. The potential of existing T-DNA insertion lines for finding

422 additional knock-out alleles in existing T-DNA insertion lines is also indicated by the fact that TDNA-

423 Seq revealed additional insertion loci in established and FST-indexed lines (see Introduction).

424 Clearly, analysis by ONT sequencing can effectively reveal additional insertions and can

425 successfully be used to fully characterize the genomes of T-DNA insertion lines. This approach is

426 faster, less laborious, more comprehensive and compared to the level of reliability also significantly

427 cheaper than PCR- or short-read based methods.

428

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

430 The variance of the T-DNA structures that were resolved by ONT sequencing spans a really wide

431 range of configurations and lengths. Tandem repeats as well as inverted repeats [46] are occurring.

432 Insertion length starts with 2.7 kbp for 909H04-At1g38212 (Fig. 7B) and reaches up to about 50 kbp

433 for 767D12-At2g19210-At2g21385 (Fig. 6C). The lines were selected to contain a T-DNA by

434 checking for resistance to sulfadiazine [22] which is provided by the T-DNA used at GABI-Kat.

435 However, since there are regularly several insertions per line, also T-DNA fragments with a

436 truncated selection marker gene are to be expected - given that resistance is provided in trans. For

437 SALK, SAIL and WISC insertion lines, T-DNA arrays sizes of up to 236 kbp have been reported

438 [19]. We hypothesize that the complexity of T-DNA arrays might correlate with the tendency of

439 selection marker silencing, which could mechanistically be implemented via siRNA [19]. The

440 comparably reduced complexity of T-DNA arrays derived from pAC161 (the binary vectors mostly

441 used at GABI-Kat) could thus explain why the sulfadiazine selection marker stays active for many

442 generations.

443 Inclusion of BVB sequences in T-DNA array structures has been reported repeatedly for various
444 species [37, 47, 48]. For the studied GABI-Kat lines, BVB sequences were structurally resolved as
445 internal components of T-DNA arrays as well as at the junction to genomic sequences. A total of six
446 T-DNA arrays with BVB sequences were detected among 26 cases, indicating that about 20% of all
447 insertions, and an even higher percentage of lines, contain inserted BVB sequences.
448 We detected only few intact right border sequences in contrast to left border sequences, which fits
449 the empirical observation that FST-generation for characterization of insertion populations is much
450 more productive at LB than at RB [4-6]. In turn, the lines studied here are selected from insertions
451 detected by using LB for FST generation, which introduces a bias. When insertions accessed via
452 FSTs from RB were studied, RB is found to be more precisely cut than LB [13, 49], which is
453 explained by protection of RB by VirD2 [9]. Nevertheless, within the longer T-DNA arrays and also
454 in the insertions newly detected by ONT sequencing, most of the RBs are lost. This does not fit well
455 to current models for the integration mechanism and explanations for the observed internal "right
456 end to right end" (without RB) fusions in T-DNA arrays and requires further investigation.
457

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

459 Changes in the genome sequence in the direct vicinity of the T-DNA insertion site have already
460 been described in detail [13]. However, that study relied on data from PCR amplicon sequences

461 and could, therefore, not detect events that affect distances longer than the length of an average
462 amplimer of about 2 kbp. In addition, amplicons from both T-DNA::genome junctions were required.
463 Here, we addressed insertions that failed to fulfill the "amplicon sequences from both junctions
464 available" criterion. This allowed to focus on a set of GABI-Kat lines that has a higher chance of
465 showing genomic events (Table 1). The T-DNA::genome junctions studied here fall, with one
466 exception, into the range already described for DSB-based integration and repair by NHEJ, with
467 filler sequences and microhomology at the insertion site [9]. The exception is 909H04-At1g54385-
468 cp-At1g54440 (Fig. 7C), an insertion allele that displays an about 20 kbp long duplicated inversion
469 and in addition 652 bp derived from the plastome at the additional breakpoint that links the
470 inversion back to the chromosome. During repair of the initial DSBs and in parallel to T-DNA
471 integration, also cpDNA is used to join broken ends of DNA at insertion loci. Inversions require
472 more than one repaired DSB in the DNA at the insertion site to be explained, and that cpDNA is
473 available in the nucleus has been demonstrated experimentally [50] and also in the context of
474 horizontal gene transfer [51].
475

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

477 Our analyses revealed five lines with chromosome arm translocations, either exchanged within one
478 chromosome (GK-433E06, Fig. 5) or exchanged between two different chromosomes (Figures 1 to

479 4). In addition, line GK-767D12 displayed a chromosomal rearrangement that resulted in an
480 inverted duplication of 0.8 Mbp. This aligns well with previous reports of interchromosomal
481 structural variations, translocations, and chromosome fusions in T-DNA insertion lines [18, 19, 33,
482 34]. Because of the bias for complex cases in the criteria we used for selection of the lines
483 investigated, we cannot deduce a reliable value for the frequency of chromosomal rearrangements
484 in the GABI-Kat population. However, an approximation taking into account that the 6 cases are
485 from 14 lines sequenced, and the 14 lines sequenced are a subset of 342 out of 1,818 lines with
486 attempted confirmation of both borders but failure at the 2nd T-DNA::genome junction, ends up with
487 about one of 10 GABI-Kat lines that may display chromosome-level rearrangements (~10%). It
488 remains to be determined if this estimation holds true, but the approximation fits somehow to the
489 percentage of T-DNA insertion lines that show Mendelian inheritance of mutant phenotypes (88%)
490 while 12% do not [52]. For the SALK T-DNA population, 19% lines with chromosomal translocations
491 have been reported [18], based on genetic markers and lack of linkage between markers from
492 upstream and downstream of an insertion locus.
493 Although the number of investigated lines with chromosome arm translocations is small, the high
494 proportion of fusions between Chr3 and Chr5 in our dataset is conspicuous (3 out of 5, see
495 Additional file 3). Also for the line SAIL_232 a fusion of Chr3 and Chr5 was reported [19]. This work
496 addressed 4 T-DNA insertion lines (two SALK, one WISC and SAIL_232) by ONT sequencing and

497 Bionano Genomics (BNG) optical genome maps. Translocations involving chromosomes other than
498 Chr3 and Chr5 were observed in our study and have also been reported before [16-18, 33], but it is
499 possible that translocations between Chr3 and Chr5 occur with a higher rate than others. Full
500 sequence characterization of the genomes of (many) more T-DNA insertion lines by long read
501 sequencing have the potential to reveal hot spots of translocations and chromosome fusions, if
502 these exist.
503 ONT reads are best suited for the identification of large scale structural variants, for this reason we
504 did not address single nucleotide variants (SNV) or small insertions/deletions (InDels). T-DNA-
505 independent structural variants that were identified, except chromosome fusions, are mostly
506 explained by potential errors in the TAIR9 reference sequence as discussed below. One may
507 conclude that the remaining 8 out of 14 lines for which we have not detected large structural
508 variation are identical in genome organization to the Col-0 wild type. However, this conclusion is
509 based on negative results and should be taken with care.
510

511 **Compensating translocations**

512 The chromosome arm translocations detected are all "reciprocal" translocations, which involve two
513 breakpoints and exchange parts of chromosomes. Both rearranged chromosomes are equally
514 detected in the sequenced DNA of the line. The combination of both rearranged chromosomes in

515 the offspring is maintained because homozygosity of only one of the two rearranged chromosomes
516 is expected to be lethal due to imbalance of gene dose for large chromosomal regions. However, if
517 both rearranged chromosomes can be transmitted together in one gametophyte, both rearranged
518 chromosomes might exist in offspring in homozygous state [53]. The induction of chromosome
519 mutations, chromosomal rearrangements and chromosome arm translocations by T-DNA insertion
520 mutagenesis has since a long time received attention in *A. thaliana* genetics. One reason is that
521 these types of mutations cause distorted segregation among offspring since similar segregation
522 results are also indicative of genes essential for gametophyte development [53-55]. With regard to
523 the chromosome arm translocations we detected, it is important to note that some of the arms are
524 fused without integrated T-DNA. The case of line GK-082G09 (one T-DNA insertion locus, still
525 reciprocal chromosome arm translocation) is relevant in this context, because the presence of a
526 single T-DNA insertion per line was used as a criterion to select valid candidates for gametophyte
527 development mutants.
528 Our analyses of the sequences of T-DNA free chromosomal junctions (e.g. 082G09-At5g57020-0-
529 At3g19080) did not result in the detection of specialties that make these junctions different from T-
530 DNA::genome junctions. We cannot fully exclude that the T-DNA free junctions are the result of
531 recombination of two loci that initially both contained T-DNA, and that one T-DNA got lost during
532 recombination at one of two loci. However, it is also possible that the translocations are the direct

533 result of DSB repair, similar to what has been found after targeted introduction of DSBs [56]. We
534 speculate that both, T-DNA containing and T-DNA free junction cases, result from
535 DSB/integration/repair events that involve genome regions which happen to be in close contact,
536 even if different chromosomes are involved. It is evident that several DSB breaks are required, and
537 repair of these DSB can happen with the DNA that is locally available, might it be cpDNA (see
538 above), T-DNA that must have been delivered to DSB repair sites, or different chromosomes that
539 serve as template for fillers [13] or as target for fusion after a DSB happened to occur.
540 ONT sequencing of mutants offers easy access to data on presence or absence of translocations
541 and other structural variants. We demonstrated that long read sequencing is a suitable method for
542 characterizing lines that failed in previous characterization attempts based on FSTs. For example,
543 the investigation of T-DNA insertion alleles/lines that display deformed pollen phenotypes, which
544 was impacted by chromosome fusions and uncharacterized T-DNA insertions [55], can now be
545 combined with long read sequencing to reveal all insertion and structural variation events with high
546 resolution. Clearly, comprehensive characterization of T-DNA insertion lines, independent from the
547 population from which the mutant originates, as well as other lines used for forward and reverse
548 genetic experiments, can prevent unnecessary work and questionable results. The entire workflow
549 from DNA extraction to the final genome sequence can be completed in less than a week.
550 Estimated costs per line for ONT sequencing are $200-300 if multiple samples are consecutively

551 analyzed on one GridION flow cell or if barcoding and multiplexing are applied. The application of
552 "loreta" supports the inspection of T-DNA insertions as soon as the read data are generated.
553 Therefore, we recommend to characterize T-DNA insertion alleles by long read sequencing.
554

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

556 Decreased quality (Phred scores) was previously described for ONT sequence reads as
557 consequence of inverted repeats which might form secondary structures and thus interfere with the
558 DNA translocation through the nanopore [57]. Complex T-DNA arrays are a challenge to ONT
559 sequencing and probably all other current sequencing technologies because of the frequently
560 included inverted repeat structures. We observed that the second part of an inverted repeat has low
561 sequence quality, while the first part (probably not engaging in secondary structure) is of good
562 sequence quality. Evidence has been provided that changed translocation speed is responsible for
563 the low Phred scores in inverted repeats [57]. The quality decrease needs to be considered
564 especially when performing analyses at the single read level. These stretches of sequence with bad
565 quality also pose a challenge for the assembly, especially since the orientation of the read
566 determines which part of the inverted repeat is of good or poor quality. However, we were able to
567 solve the problem to a satisfying level by manual consideration of reads from the opposite direction
568 which contain the other part of the inverted repeat in good sequence quality.

569

570

571 ## Conclusions

572 This study presents a comprehensive characterization of multiple GABI-Kat lines by long read
573 sequencing. The results argue strongly for full characterization of mutant alleles to avoid
574 misinterpretation and errors in gene function assignments. If an insertion mutant and the T-DNA
575 insertion allele in question are not characterized well at the level of the genotype, the phenotype
576 observed for the mutant might be due to a complex integration locus, and not causally related to the
577 gene that is expected to be knocked-out by the insertion. Structural changes at the genome level,
578 including chromosome translocations and other large rearrangements with junctions without T-
579 DNA, may have confounding effects when studying the genotype to phenotype relations with T-
580 DNA lines. This conclusion must also consider that during the last 20 to 30 years, many T-DNA
581 alleles have been used in reverse genetic experiments. Finally, and similar to the ONT sequence
582 data that resulted from the analyses of four SALK and SAIL/WISC T-DNA insertion lines [19], the
583 ONT sequence data from this study allowed to detect and correct many non-centromeric
584 misassemblies in the current reference sequence.

585

586

587 # Methods

588

589 **Plant material**

590 The lines subjected to ONT sequencing were chosen from a collection of GABI-Kat lines which
591 were studied initially to collect statistically meaningful data about the structure of T-DNA insertion
592 sites at both ends of the T-DNA insertions [13]. In this context and also after 2015, confirmation
593 amplicon sequence data from both T-DNA::genome junctions of individual T-DNA insertions were
594 created at GABI-Kat, which was successful for 1,481 cases from 1,476 lines by the end of 2019
595 (1,319 cases were successfully completed for both junctions in the beginning of 2015). To generate
596 this dataset, 1,835 individual T-DNA insertions from 1,818 lines with one T-DNA::genome junction
597 already confirmed were addressed, meaning that there were 354 cases from 342 lines which failed
598 at the 2nd T-DNA::genome junction. From these 354 cases, we randomly selected the 14 insertions
599 (in 14 different lines) that were studied here, with good germination as additional criterion for
600 effective handling (Additional file 1). Since the focus of interest in insertion alleles was always to
601 identify null alleles of genes, all 14 insertions addressed are CDSi insertions (insertions in the
602 coding sequence or enclosed introns, that means between ATG and Stop codons on genomic

603 DNA). A total of 100 T2 seeds of each line were plated with sulfadiazine selection as described
604 [26]. Surviving T2 plantlets contain at least one integrated T-DNA, either in hemizygous or in
605 homozygous state. Sulfadiazine-resistant plantlets were transferred to soil, grown to about 8-leaf
606 stage and pooled for DNA extraction. For a single locus with normal heritability, statistically 66% of
607 the chromosomes in the pool contain the T-DNA. The T-DNA in GK-654A12 is from pGABI1 [38],
608 the other lines contain T-DNA from pAC161 [22].
609

610 **DNA extraction, size enrichment, and quality assessment**

611 Genomic DNA was extracted from young plantlets or young leaves through a CTAB-based protocol
612 (Additional file 12) modified from [22, 58]. We observed, like others before [4], that the quality of
613 extracted DNA decreased with the age of the leaf material processed. Young leaves lead to the
614 best results in our hands. The cause might be increasing cell and vacuole size containing more
615 harmful metabolites which may cause reduced quality and yield of DNA. As DNA quality for ONT
616 sequencing also decreases with storage time, we processed the DNA as soon as possible after
617 extraction. DNA quantity and quality was initially assessed based on NanoDrop (Thermo Scientific)
618 measurement, and on an agarose gel for DNA fragment size distribution. Precise DNA
619 quantification was performed via Qubit (Thermo Fisher) measurement using the broad range buffer

620 following the supplier's instructions. Up to 9 µg of genomic DNA were subjected to an enrichment of

621 long fragments via Short Read Eliminator kit (Circulomics) according to the suppliers' instructions.

622

623 **Library preparation and ONT sequencing**

624 DNA solutions enriched for long fragments were quantified via Qubit again. One µg DNA (R9.4.1

625 flow cells) or two µg (R10 flow cells) were subjected to library preparation following the LSK109

626 protocol provided by ONT. Sequencing was performed on R9.4.1 and R10 flow cells on a GridION.

627 Real time base calling was performed using Guppy v3.0 on the GridION (R9.4.1 flow cells) and on

628 graphic cards in the de.NBI cloud [59] (R10 flow cells), respectively.

629

630 ***De novo* genome sequence assemblies**

631 Reads of each GK line were assembled separately to allow validation of other analysis methods

632 (see below). Canu v1.8 [60] was deployed with previously optimized parameters [61]. Assembly

633 quality was assessed based on a previously developed Python script (Table 2). No polishing was

634 performed for assemblies of individual GABI-Kat lines as these assemblies were only used to

635 analyze large structural variants and specifically T-DNA insertions.

636 Through removal of all T-DNA reads from the combined ONT read dataset from all insertion lines

637 (Col-0 background [22]) and size filtering, a comprehensive data set of ultra-long reads (longer than

638 100,000 nt) was generated. This dataset is available from ENA/GenBank with the ID ERS5246674
639 (SAMEA7490021). The assembly of these ultra-long reads was computed as described above.
640 Polishing was performed with Racon v.1.4.7 [62] and medaka v.0.10.0 as previously described [58].
641 Potential contamination sequences were removed based on sequence similarity to the genome
642 sequences of other species, and contigs smaller than 100 kbp were discarded as previously
643 described [58, 63]. To ensure accurate representation of the Col-0 wild type genome structure, the
644 assembly was checked for the chromosome fusion events reported for GK-082G09, GK-433E06,
645 and GK-654A12 as well as for the chloroplast DNA integration of GK-909H04. The Sanger reads of
646 the validation amplicons generated for these loci were subjected to a search via BLASTn [64] using
647 default settings. BLASTn was also used to validate the absence of any T-DNA or plasmid
648 sequences in this assembly using pSKI015 (AF187951), pAC161 (AJ537514), and pROK2 [6] as
649 query.
650

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

652 To identify differences to the TAIR9 reference genome sequence, the Col-0_GK-wt contigs were
653 sorted and orientated using pseudogenetic markers derived from TAIR9. The TAIR9 sequence was
654 split into 500 bp long sequence chunks which were searched against the Col-0_GK-wt contigs via
655 BLAST. Unique hits with at least 80% of the maximal possible BLAST score were considered as

656 genetic markers. The following analysis with ALLMAPS [65] revealed additional and thus

657 unmatched sequences of Col-0_GK-wt around the centromeres.

658 ONT reads were aligned to the TAIR9 reference genome sequence via Minimap2 v2.1-r761 [66].

659 Mappings were converted into BED files with bedtools v2.26 [67]. The alignments were evaluated

660 for the ends of mapping reads, and these ends were quantified in genomic bins of 100 bp using a

661 dedicated tool designated Assembly Error Finder (AEF) v0.12 (Table 2) with default parameters.

662 Neighboring regions with high numbers of alignment ends were grouped if their distance was

663 smaller than 30 kbp. Regions with outstanding high numbers of alignment ends indicate potential

664 errors in the targeted assembly. A selection of these regions from TAIR9 was compared against

665 Col-0_GKat-wt through dot plots [32].

666

667 **Analysis of T-DNA insertions**

668 The T-DNA insertions of each line were analyzed in a semi-automatic way. A tool was developed,

669 written in Python and designated "loreta" (Table 2), that needs as input: reads in FASTQ format, T-

670 DNA sequences in FASTA format, a reference file containing sequences for assembly annotation in

671 FASTA format (for this study: sequences of T-DNA and vector backbone, the *A. thaliana* nuclear

672 genome, plastome, chondrome, and the *A. tumefaciens genome*), and – if available – precomputed

673 *de novo* assemblies (as described above). The output are HTML pages with annotated images

674 displaying models of T-DNA insertions and their neighborhood. Partial assemblies of reads
675 containing T-DNA sequences are computed, and the parts of the *de novo* assemblies containing T-
676 DNA sequences are extracted. All resulting sequences as well as all individual reads containing T-
677 DNA sequences are annotated using the reference file. If run on a local machine, a list of tools that
678 need to be installed is given in the github repository (Table 2). For easier access to the tool on a
679 local machine, there is also a Docker file available that can be used to build a Docker image.
680 Reads containing T-DNA sequences were identified by BLASTn [64, 68] using an identity cutoff of
681 80% and an e-value cutoff of 1e-50. All identified reads were then assembled using Canu v1.8 with
682 the same parameters as for the *de novo* assemblies (see above) and in addition some parameters
683 to facilitate assemblies with low coverage: correctedErrorRate=0.17, corOutCoverage=200,
684 stopOnLowCoverage=5 and an expected genome size of 10 kbp. From the precomputed *de novo*
685 assemblies, fragments were extracted that contain the T-DNA insertion and 50 kbp up- and
686 downstream sequence. The resulting fragments, contigs from the Canu assembly, the contigs
687 marked as "unassembled" by Canu as well as all individual reads (converted to FASTA using
688 Seqtk-1.3-r106 [69]) were annotated using the reference sequences. For this purpose, a BLASTn
689 search was performed (contig versus reference sequences) with the same parameters as for the
690 identification of T-DNA reads. These BLAST results were mapped to the sequence (contig/read) as
691 follows: BLAST hits were annotated one after another, sorted by decreasing score. If the overlap of

692 a BLAST hit with a previously annotated one exceeds 10 bp, the second BLAST hit was discarded.
693 For further analysis, reads were mapped back to the assembly. All reads were mapped back to the
694 fragments of the *de novo* assembly, reads containing T-DNA were mapped back to (1) the Canu
695 contigs, (2) "unassembled" contigs and (3) individual reads containing T-DNA sequences. Mapping
696 was performed using Minimap2 [66] with the default options for mapping of ONT sequencing data.
697 To further inspect the chromosome(s) sequences prior to T-DNA insertion, the same analysis was
698 performed using the *A. thaliana* sequences neighboring the T-DNA insertion. A FASTA file was
699 generated that contains these flanking sequences using bedtools [67], reads containing this part of
700 *A. thaliana* sequence (and no T-DNA) are again identified using BLAST, assembled and annotated
701 as described above. Infoseq from the EMBOSS package [70] was used to calculate the length of
702 different sequences in the pipeline.
703 All information is summarized in HTML files containing images; these images display all annotated
704 sequences along with mapped reads and details on the BLAST results. These pages were used for
705 manual inspection and final determination of the insertion structures. For lines with a single
706 canonical insertion, such as 050B11-At5g64610 in GK-050B11 (see Additional file 3), the
707 assembled and annotated contig of the partial assembly was sufficient. In more complex cases like
708 GK-038B07, the exact insertion structure was not clear based on the assembled contigs or based
709 on sections of the *de novo* assembly. If, based on the read mappings shown in the visualization,

710 the partial assembly looked erroneous (many partial mappings), individual reads were used for the
711 determination of the insertion structure. These individual reads were also considered for
712 determination of exact T-DNA positions, if these positions were not clear from contigs / sections.
713 This was often the case for head-to-head or tail-to-tail configurations of T-DNA arrays, where one of
714 the T-DNAs was represented by sequence of low quality and could not be annotated (and led to
715 misassemblies in assembled contigs). These cases were resolved by identification of reads
716 orientated in the other direction, because then the sequence derived from the other T-DNA was of
717 low quality and by combining the annotated results, a clear picture could be derived. If different
718 reads contradicted each other in exact positions of the T-DNA, we chose the "largest possible T-
719 DNA" that could be explained by individual reads.
720

721 **Mapping of ONT reads for detection of copy number variation**

722 ONT reads from each line were aligned against the TAIR9 Col-0 reference genome sequence using
723 Minimap v2.10-r761 [71] with the options '-ax map-ont --secondary=no'. The resulting mappings
724 were converted into BAM files via samtools v1.8 [72] and used for the construction of coverage files
725 with a previously developed Python script [73]. Coverage plots (see Additional file 5) were
726 constructed as previously described [61] and manually inspected for the identification of copy
727 number variations.

728

729 **Sequence read quality assessment**

730 Reads containing T-DNA sequence were annotated based on sequence similarity to other known
731 sequences based on BLASTn [64, 68] usually matching parts of the Ti-plasmid or *A. thaliana*
732 genome sequence. Reads associated with complex T-DNA insertions were considered for
733 downstream analysis if substantial parts (>1 kbp) of the read sequence were not matched to any
734 database sequences via BLASTn. Per base quality (Phred score) of such reads was assessed
735 based on a sliding window of 200 nt with a step size of 100 nt.

736

737 **Chromosome fusion and cpDNA insertion validation via PCR and Sanger sequencing**

738 Chromosomal fusions without a connecting T-DNA were analyzed via PCR using manually
739 designed flanking primers (Additional file 4). Amplicons were generated using genomic DNA
740 extracted from plants of the respective line as template with Q5 High-Fidelity DNA polymerase
741 (NEB) following supplier's instructions. PCR products were separated on a 1% agarose gel and
742 visualized using ethidiumbromide and UV light. Amplicons were purified with Exo-CIP Rapid PCR
743 Cleanup Kit (NEB) following supplier's instructions. Sanger sequencing was performed at the
744 Sequencing Core Facility of the Center for Biotechnology (Bielefeld University, Bielefeld, Germany)
745 using BigDye terminator v3.1 chemistry (Thermo Fisher) on a 3730XL sequencer. The resulting

746 Sanger sequences were merged using tools from the EMBOSS package [70]. After transferring
747 reverse reads to their reverse complement using revseq, a multiple alignment was generated using
748 MAFFT [74]. One consensus sequence for each amplicon was extracted from the alignments using
749 em_cons with option -plurality 1, and the resulting sequences were submitted to ENA (see
750 Additional file 1 for accession numbers).
751

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

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

763 # Declarations

764 **Ethics approval and consent to participate**

765 Not applicable

766

767 **Consent for publication**

768 Not applicable

769

770 **Availability of data and materials**

771 Sequence read datasets generated and analyzed during this study were made available at ENA

772 under the accession PRJEB35658. Individual run IDs are included in Additional file 1. The Col-0

773 genome sequence assembly of the GABI-Kat Col-0 genetic background (Col-0_GKat-wt) is

774 available at ENA under the accession GCA_905067165.

775

776 **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

777

778

779 **Competing interest**

780 The authors declare that they have no competing interests.

781

782 **Funding**

783 We acknowledge support for the Article Processing Charge by the Open Access Publication Fund

784 of Bielefeld University.

785

786 **Authors' contribution**

787 BP performed DNA extraction and sequencing. BP and NK performed bioinformatic analyses. BP,

788 NK, and BW interpreted the results and wrote the manuscript.

789

790 **Acknowledgements**

791 We are very grateful to the Bioinformatics Resource Facility support team of the CeBiTec and to

792 de.NBI for providing computing infrastructure and excellent technical support. We also thank the

793 Sequencing Core Facility of the CeBiTec for granting access to the ONT infrastructure. Many

794 thanks to Tobias Busche, Christian Rückert, and Jörn Kalinowski for general support related to the

795 ONT sequencing, and to Prisva Viehöver for high quality Sanger sequencing. We thank Andrea

796 Voigt for excellent technical support. The bioinformatic work was supported in part by grants from

797 the German Federal Ministry of Education and Research (BMBF) for the project "Bielefeld-Gießen

798 Center for Microbial Bioinformatics–BiGi" (grant no. 031A533) within the German Network for
799 Bioinformatics Infrastructure (de.NBI).

800

801 **Supplementary information**

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

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

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

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

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

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

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

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

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

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

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

818 sequence.

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

820

821

822 # References

823

824

825

826 1. Ulker B, Peiter E, Dixon DP, Moffat C, Capper R, Bouche N, Edwards R, Sanders D, Knight

827 H, Knight MR: **Getting the most out of publicly available T-DNA insertion lines**. *The Plant*

828 *Journal* 2008, **56**(4):665-677.

829 2. OMalley RC, Ecker JR: **Linking genotype to phenotype using the Arabidopsis unimutant**

830 **collection.** *The Plant Journal* 2010, **61**(6):928-940.

831 3. Fauser F, Roth N, Pacher M, Ilg G, Sanchez-Fernandez R, Biesgen C, Puchta H: **In planta**

832 **gene targeting**. *Proceedings of the National Academy of Sciences of the United States of*

833 *America* 2012, **109**(19):7535-7540.

834 4. OMalley RC, Barragan CC, Ecker JR: **A user's guide to the Arabidopsis T-DNA insertion**
835 **mutant collections**. In: *Methods in Molecular Biology.* Edited by Alonso J, Stepanova A, vol.
836 1284, 2015/03/12 edn. New York, NY: Humana Press; 2015: 323-342.
837 5. Strizhov N, Li Y, Rosso MG, Viehoever P, Dekker KA, Weisshaar B: **High-throughput**
838 **generation of sequence indexes from T-DNA mutagenized Arabidopsis thaliana lines**.
839 *BioTechniques* 2003, **35**(6):1164-1168.
840 6. Alonso JM, Stepanova AN, Leisse TJ, Kim CJ, Chen H, Shinn P, Stevenson DK,
841 Zimmerman J, Barajas P, Cheuk R *et al*: **Genome-wide insertional mutagenesis of**
842 **Arabidopsis thaliana**. *Science* 2003, **301**(5633):653-657.
843 7. Puchta H: **Applying CRISPR/Cas for genome engineering in plants: the best is yet to come**.
844 *Current Opinion in Plant Biology* 2017, **36**:1-8.
845 8. Smith EF, Townsend CO: **A Plant-Tumor of Bacterial Origin**. *Science* 1907, **25**(643):671-
846 673.
847 9. Gelvin SB: **Integration of Agrobacterium T-DNA into the Plant Genome.** *Annual Review of*
848 *Genetics* 2017, **51**:195-217.
849 10. Zambryski P, Holsters M, Kruger K, Depicker A, Schell J, Van Montagu M, Goodman H:
850 **Tumor DNA structure in plant cells transformed by A. tumefaciens.** *Science* 1980,
851 **209**(4463):1385-1391.
852 11. Hernalsteens J, Van Vliet F, De Beuckeleer M, Depicker A, Engler G, Lemmers M, Holsters
853 M, Van Montagu M, Schell J: **The Agrobacterium tumefaciens Ti plasmid as a host vector**
854 **system for introducing foreign DNA in plant cells**. *Nature* 1980, **287**(5783):654-656.
855 12. Clough SJ, Bent AF: **Floral dip: a simplified method for Agrobacterium-mediated**
856 **transformation of *Arabidopsis thaliana***. *The Plant Journal* 1998, **16**(6):735-743.

857 13. Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B: **The Structural**
858 **Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break**
859 **Repair-Based Insertion Mechanism.** *Molecular Plant* 2015, **8**(11):1651-1664.
860 14. van Kregten M, de Pater S, Romeijn R, van Schendel R, Hooykaas PJ, Tijsterman M: **T-**
861 **DNA integration in plants results from polymerase-θ-mediated DNA repair.** *Nature Plants*
862 2016, **2**(11):16164.
863 15. Castle LA, Errampalli D, Atherton TL, Franzmann LH, Yoon ES, Meinke DW: **Genetic and**
864 **molecular characterization of embryonic mutants identified following seed transformation in**
865 **Arabidopsis**. *Molecular Genetics and Genomics* 1993, **241**(5-6):504-514.
866 16. Forsbach A, Schubert D, Lechtenberg B, Gils M, Schmidt R: **A comprehensive**
867 **characterization of single-copy T-DNA insertions in the Arabidopsis thaliana genome.** *Plant*
868 *Molecular Biology* 2003, **52**(1):161-176.
869 17. Lafleuriel J, Degroote F, Depeiges A, Picard G: **A reciprocal translocation, induced by a**
870 **canonical integration of a single T-DNA, interrupts the HMG-I/Y Arabidopsis thaliana gene**.
871 *Plant Physiology and Biochemistry* 2004, **42**(3):171-179.
872 18. Clark KA, Krysan PJ: **Chromosomal translocations are a common phenomenon in**
873 **Arabidopsis thaliana T-DNA insertion lines.** *The Plant Journal* 2010, **64**(6):990-1001.
874 19. Jupe F, Rivkin AC, Michael TP, Zander M, Motley ST, Sandoval JP, Slotkin RK, Chen H,
875 Castanon R, Nery JR *et al*: **The complex architecture and epigenomic impact of plant T-**
876 **DNA insertions.** *PLoS Genetics* 2019, **15**(1):e1007819.
877 20. Min Y, Frost JM, Choi Y: **Gametophytic Abortion in Heterozygotes but Not in Homozygotes:**
878 **Implied Chromosome Rearrangement during T-DNA Insertion at the ASF1 Locus in**
879 **Arabidopsis**. *Molecules and Cells* 2020, **43**(5):448-458.

880 21. Krispil R, Tannenbaum M, Sarusi-Portuguez A, Loza O, Raskina O, Hakim O: **The Position**
881 **and Complex Genomic Architecture of Plant T-DNA Insertions Revealed by 4SEE.**
882 *International Journal of Molecular Sciences* 2020, **21**(7):ijms21072373.
883 22. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B: **An *Arabidopsis thaliana* T-**
884 **DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse genetics.**
885 *Plant Molecular Biology* 2003, **53**(1-2):247-259.
886 23. Kleinboelting N, Huep G, Kloetgen A, Viehoever P, Weisshaar B: **GABI-Kat SimpleSearch:**
887 **new features of the Arabidopsis thaliana T-DNA mutant database.** *Nucleic Acids Research*
888 2012, **40**(Database issue):D1211-D1215.
889 24. Sessions A, Burke E, Presting G, Aux G, McElver J, Patton D, Dietrich B, Ho P, Bacwaden
890 J, Ko C *et al*. **A High-Throughput Arabidopsis Reverse Genetics System**. *The Plant Cell*
891 2002, **14**(12):2985-2994.
892 25. Sussman MR, Amasino RM, Young JC, Krysan PJ, Austin-Phillips S: **The Arabidopsis**
893 **knockout facility at the University of Wisconsin-Madison**. *Plant Physiology* 2000,
894 **124**(4):1465-1467.
895 26. Li Y, Rosso MG, Viehoever P, Weisshaar B: **GABI-Kat SimpleSearch: an Arabidopsis**
896 **thaliana T-DNA mutant database with detailed information for confirmed insertions.** *Nucleic*
897 *Acids Research* 2007, **35**(Database issue):D874-D878.
898 27. Kleinboelting N, Huep G, Weisshaar B: **Enhancing the GABI-Kat Arabidopsis thaliana T-**
899 **DNA Insertion Mutant Database by Incorporating Araport11 Annotation**. *Plant and Cell*
900 *Physiology* 2017, **58**(1):e7.

901 28. Cheng CY, Krishnakumar V, Chan A, Thibaud-Nissen F, Schobel S, Town CD: **Araport11: a**
902 **complete reannotation of the Arabidopsis thaliana reference genome.** *The Plant Journal*
903 2017, **89**(4):789-804.
904 29. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K,
905 Alexander DL, Garcia-Hernandez M *et al*: **The Arabidopsis Information Resource (TAIR):**
906 **improved gene annotation and new tools.** *Nucleic Acids Research* 2012, **40**(Database
907 issue):D1202-D1210.
908 30. Huep G, Kleinboelting N, Weisshaar B: **An easy-to-use primer design tool to address**
909 **paralogous loci and T-DNA insertion sites in the genome of Arabidopsis thaliana.** *Plant*
910 *Methods* 2014, **10**:28.
911 31. Vukasinovic N, Cvrckova F, Elias M, Cole R, Fowler JE, Zarsky V, Synek L: **Dissecting a**
912 **hidden gene duplication: the Arabidopsis thaliana SEC10 locus.** *PLoS ONE* 2014,
913 **9**(4):e94077.
914 32. Pucker B, Holtgräwe D, Stadermann KB, Frey K, Huettel B, Reinhardt R, Weisshaar B: **A**
915 **chromosome-level sequence assembly reveals the structure of the Arabidopsis thaliana Nd-**
916 **1 genome and its gene set.** *PLoS One* 2019, **14**(5):e0216233.
917 33. Nacry P, Camilleri C, Courtial B, Caboche M, Bouchez D: **Major chromosomal**
918 **rearrangements induced by T-DNA transformation in Arabidopsis**. *Genetics* 1998,
919 **149**(2):641-650.
920 34. Tax FE, Vernon DM: **T-DNA-associated duplication/translocations in Arabidopsis.**
921 **Implications for mutant analysis and functional genomics**. *Plant Physiology* 2001,
922 **126**(4):1527-1538.

923 35. Wei FJ, Kuang LY, Oung HM, Cheng SY, Wu HP, Huang LT, Tseng YT, Chiou WY, Hsieh-
924 Feng V, Chung CH *et al*: **Somaclonal variation does not preclude the use of rice**
925 **transformants for genetic screening**. *The Plant Journal* 2016, **85**(5):648-659.
926 36. Gang H, Liu G, Zhang M, Zhao Y, Jiang J, Chen S: **Comprehensive characterization of T-**
927 **DNA integration induced chromosomal rearrangement in a birch T-DNA mutant**. *BMC*
928 *Genomics* 2019, **20**(1):311.
929 37. Krizkova L, Hrouda M: **Direct repeats of T-DNA integrated in tobacco chromosome:**
930 **characterization of junction regions**. *The Plant Journal* 1998, **16**(6):673-680.
931 38. Ulker B, Li Y, Rosso MG, Logemann E, Somssich IE, Weisshaar B: **T-DNA-mediated**
932 **transfer of Agrobacterium tumefaciens chromosomal DNA into plants**. *Nature Biotechnology*
933 2008, **26**(9):1015-1017.
934 39. Seagrist JF, Su SH, Krysan PJ: **Recombination between T-DNA insertions to cause**
935 **chromosomal deletions in Arabidopsis is a rare phenomenon**. *PeerJ* 2018, **6**:e5076.
936 40. Schmidt C, Schindele P, Puchta H: **From gene editing to genome engineering: restructuring**
937 **plant chromosomes via CRISPR/Cas**. *aBIOTECH* 2020, **1**(1):21-31.
938 41. Pellestor F, Gatinois V: **Chromoanagenesis: a piece of the macroevolution scenario**.
939 *Molecular Cytogenetics* 2020, **13**:3.
940 42. Li X, Zhang R, Patena W, Gang SS, Blum SR, Ivanova N, Yue R, Robertson JM, Lefebvre
941 PA, Fitz-Gibbon ST *et al*: **An Indexed, Mapped Mutant Library Enables Reverse Genetics**
942 **Studies of Biological Processes in Chlamydomonas reinhardtii**. *The Plant Cell* 2016,
943 **28**(2):367-387.
944 43. Inagaki S, Henry IM, Lieberman MC, Comai L: **High-Throughput Analysis of T-DNA Location**
945 **and Structure Using Sequence Capture.** *PLoS One* 2015, **10**(10):e0139672.

946 44. Jiang N, Lee YS, Mukundi E, Gomez-Cano F, Rivero L, Grotewold E: **Diversity of genetic**
947 **lesions characterizes new Arabidopsis flavonoid pigment mutant alleles from T-DNA**
948 **collections**. *Plant Science* 2020, **291**:110335.
949 45. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, Loudet O, Weigel D, Ecker
950 JR: **High contiguity Arabidopsis thaliana genome assembly with a single nanopore flow cell.**
951 *Nature Communications* 2018, **9**(1):541.
952 46. Jorgensen R, Snyder C, Jones JDG: **T-DNA is organized predominantly in inverted repeat**
953 **structures in plants transformed with Agrobacterium tumefaciens C58 derivatives**. *Molecular*
954 *Genetics and Genomics* 1987, **207**(2-3):471-477.
955 47. Wu H, Sparks CA, Jones HD: **Characterisation of T-DNA loci and vector backbone**
956 **sequences in transgenic wheat produced by Agrobacterium-mediated transformation**.
957 *Molecular Breeding* 2006, **18**(3):195-208.
958 48. Rajapriya V, Kannan P, Sridevi G, Veluthambi K: **A rare transgenic event of rice with**
959 **Agrobacterium binary vector backbone integration at the right T-DNA border junction**. *J*
960 *Plant Biochem Biot* 2021.
961 49. Zambryski P, Depicker A, Kruger K, Goodman HM: **Tumor induction by Agrobacterium**
962 **tumefaciens: analysis of the boundaries of T-DNA.** *Journal of Molecular and Applied*
963 *Genetics* 1982, **1**(4):361-370.
964 50. Huang CY, Ayliffe MA, Timmis JN: **Direct measurement of the transfer rate of chloroplast**
965 **DNA into the nucleus**. *Nature* 2003, **422**(6927):72-76.
966 51. Bock R: **The give-and-take of DNA: horizontal gene transfer in plants.** *Trends in Plant*
967 *Science* 2009, **15**(1):11-22.

968 52. Feldmann KA: **T-DNA insertion mutagenesis in *Arabidopsis*: mutational spektrum**. *The Plant*
969 *Journal* 1991, **1**(1):71-82.
970 53. Curtis MJ, Belcram K, Bollmann SR, Tominey CM, Hoffman PD, Mercier R, Hays JB:
971 **Reciprocal chromosome translocation associated with TDNA-insertion mutation in**
972 **Arabidopsis: genetic and cytological analyses of consequences for gametophyte**
973 **development and for construction of doubly mutant lines**. *Planta* 2009, **229**(4):731-745.
974 54. Bonhomme S, Horlow C, Vezon D, de Laissardiere S, Guyon A, Ferault M, Marchand M,
975 Bechtold N, Pelletier G: **T-DNA mediated disruption of essential gametophytic genes in**
976 **Arabidopsis is unexpectedly rare and cannot be inferred from segregation distortion alone**.
977 *Molecular Genetics and Genomics* 1998, **260**(5):444-452.
978 55. Ruprecht C, Carroll A, Persson S: **T-DNA-induced chromosomal translocations in feronia**
979 **and anxur2 mutants reveal implications for the mechanism of collapsed pollen due to**
980 **chromosomal rearrangements.** *Molecular Plant* 2014, **7**(10):1591-1594.
981 56. Schmidt C, Fransz P, Ronspies M, Dreissig S, Fuchs J, Heckmann S, Houben A, Puchta H:
982 **Changing local recombination patterns in Arabidopsis by CRISPR/Cas mediated**
983 **chromosome engineering**. *Nature Communications* 2020, **11**(1):4418.
984 57. Spealman P, Burrell J, Gresham D: **Inverted duplicate DNA sequences increase**
985 **translocation rates through sequencing nanopores resulting in reduced base calling**
986 **accuracy**. *Nucleic Acids Research* 2020, **48**(9):4940-4945.
987 58. Siadjeu C, Pucker B, Viehöver P, Albach DC, Weisshaar B: **High Contiguity De Novo**
988 **Genome Sequence Assembly of Trifoliate Yam (Dioscorea dumetorum) Using Long Read**
989 **Sequencing.** *Genes* 2020, **11**(3):E274.

990 59. Belmann P, Fischer B, Kruger J, Prochazka M, Rasche H, Prinz M, Hanussek M, Lang M,
991 Bartusch F, Glassle B *et al*: **de.NBI Cloud federation through ELIXIR AAI**. *F1000Research*
992 2019, **8**:842.
993 60. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and**
994 **accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome*
995 *Research* 2017, **27**(5):722-736.
996 61. Pucker B, Ruckert C, Stracke R, Viehover P, Kalinowski J, Weisshaar B: **Twenty-Five Years**
997 **of Propagation in Suspension Cell Culture Results in Substantial Alterations of the**
998 **Arabidopsis Thaliana Genome**. *Genes* 2019, **10**(9):671.
999 62. Vaser R, Sović I, Nagarajan N, Šikić M: **Fast and accurate de novo genome assembly from**
1000 **long uncorrected reads.** *Genome Research* 2017, **27**(5):737-746.
1001 63. Pucker B, Holtgräwe D, Rosleff Sörensen T, Stracke R, Viehöver P, Weisshaar B: **A De**
1002 **Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1**
1003 **Displays Presence/Absence Variation and Strong Synteny.** *PLoS ONE* 2016,
1004 **11**(10):e0164321.
1005 64. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**.
1006 *Journal of Molecular Biology* 1990, **215**(3):403-410.
1007 65. Tang H, Zhang X, Miao C, Zhang J, Ming R, Schnable JC, Schnable PS, Lyons E, Lu J:
1008 **ALLMAPS: robust scaffold ordering based on multiple maps**. *Genome Biology* 2015, **16**:3.
1009 66. Li H: **Minimap2: pairwise alignment for nucleotide sequences.** *Bioinformatics* 2018,
1010 **34**(18):3094-3100.
1011 67. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.**
1012 *Bioinformatics* 2010, **26**(6):841-842.

1013 68. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL:
1014 **BLAST+: architecture and applications**. *BMC Bioinformatics* 2009, **10**:421.
1015 69. **Seqtk: a fast and lightweight tool for processing FASTA or FASTQ sequences**
1016 [https://github.com/lh3/seqtk]
1017 70. Rice P, Longden I, Bleasby A: **EMBOSS: The European Molecular Biology Open Software**
1018 **Suite**. *Trends in Genetics* 2000, **16**(6):276-277.
1019 71. Li H: **Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences.**
1020 *Bioinformatics* 2016, **32**(14):2103-2110.
1021 72. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin
1022 R: **The Sequence Alignment/Map format and SAMtools**. *Bioinformatics* 2009, **25**(16):2078-
1023 2079.
1024 73. Pucker B, Brockington SF: **Genome-wide analyses supported by RNA-Seq reveal non-**
1025 **canonical splice sites in plant genomes**. *BMC Genomics* 2018, **19**(1):980.
1026 74. Katoh K, Standley DM: **MAFFT multiple sequence alignment software version 7:**
1027 **improvements in performance and usability.** *Molecular Biology and Evolution* 2013,
1028 **30**(4):772-780.

1029