

1 **A Chromosome-level Sequence Assembly Reveals the Structure of the *Arabidopsis***  
2 ***thaliana* Nd-1 Genome and its Gene Set**

3

4 **Boas Pucker<sup>1</sup>, Daniela Holtgräwe<sup>1</sup>, Kai Bernd Stadermann<sup>1</sup>, Katharina Frey<sup>1</sup>, Bruno**  
5 **Huettel<sup>2</sup>, Richard Reinhardt<sup>2</sup> and Bernd Weisshaar<sup>1\*</sup>**

6

7 <sup>1</sup> Bielefeld University, Faculty of Biology & Center for Biotechnology, Bielefeld, Germany

8 <sup>2</sup> Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research,  
9 Cologne, Germany

10 \* Corresponding author (BW)

11

12 Email addresses:

13 BP: [bpucker@cebitec.uni-bielefeld.de](mailto:bpucker@cebitec.uni-bielefeld.de)  
14 DH: [dholtgra@cebitec.uni-bielefeld.de](mailto:dholtgra@cebitec.uni-bielefeld.de)  
15 KS: [kstaderm@cebitec.uni-bielefeld.de](mailto:kstaderm@cebitec.uni-bielefeld.de)  
16 KF: [katharina.frey@uni-bielefeld.de](mailto:katharina.frey@uni-bielefeld.de)  
17 BH: [huettel@mpipz.mpg.de](mailto:huettel@mpipz.mpg.de)  
18 RR: [reinhardt@mpipz.mpg.de](mailto:reinhardt@mpipz.mpg.de)  
19 BW: [bernd.weisshaar@uni-bielefeld.de](mailto:bernd.weisshaar@uni-bielefeld.de)

20

21 ORCIDs

22 BP: <https://orcid.org/0000-0002-3321-7471>  
23 DH: <https://orcid.org/0000-0002-1062-4576>  
24 KS: <https://orcid.org/0000-0002-8036-1492>  
25 BW: <https://orcid.org/0000-0002-7635-3473>

26

27

28 **Abstract**

29 **Background**

30 In addition to the BAC-based reference sequence of the accession Columbia-0 from the year  
31 2000, several short read assemblies of THE plant model organism *Arabidopsis thaliana* were  
32 published during the last years. Also, a SMRT-based assembly of Landsberg *erecta* has  
33 been generated that identified translocation and inversion polymorphisms between two  
34 genotypes of the species.

35 **Results**

36 Here we provide a chromosome-arm level assembly of the *A. thaliana* accession  
37 Niederzenz-1 (AthNd-1\_v2c) based on SMRT sequencing data. The best assembly  
38 comprises 69 nucleome sequences and displays a contig length of up to 16 Mbp. Compared  
39 to an earlier Illumina short read-based NGS assembly (AthNd-1\_v1), a 75 fold increase in  
40 contiguity was observed for AthNd-1\_v2c. To assign contig locations independent from the  
41 Col-0 gold standard reference sequence, we used genetic anchoring to generate a *de novo*  
42 assembly. In addition, we assembled the chondrome and plastome sequences.

43

44 **Conclusions**

45 Detailed analyses of AthNd-1\_v2c allowed reliable identification of large genomic  
46 rearrangements between *A. thaliana* accessions contributing to differences in the gene sets  
47 that distinguish the genotypes. One of the differences detected identified a gene that is  
48 lacking from the Col-0 gold standard sequence. This *de novo* assembly extends the known  
49 proportion of the *A. thaliana* pan-genome.

50

51

52 **Background**

53 **Introduction**

54 *Arabidopsis thaliana* became the most important model for plant biology within decades due  
55 to properties valuable for basic research like short generation time, small footprint, and a  
56 small genome [1]. Shortcomings of the BAC-by-BAC assembled 120 Mbp long Col-0 gold  
57 standard sequence [2] are some missing sequences and gaps in almost inaccessible regions  
58 like repeats in the centromeres [3, 4], at the telomeres and throughout NORs as well as few  
59 mis-assemblies [5, 6]. Information about genomic differences between *A. thaliana*  
60 accessions were mostly derived from short read data [7-9]. Only selected accessions were  
61 sequenced deep enough and with sufficient read length to reach almost reference-size  
62 assemblies [7, 10-15]. While the identification of SNPs can be based on short read  
63 mappings, the identification of structural variants had an upper limit of 40 bp for most of the  
64 investigated accessions [9]. Larger insertions and deletions, which will often result in  
65 presence/absence variations of entire genes, are often missed in short read data sets but are  
66 easily recovered by long read sequencing [14-16]. *De novo* assemblies based on long  
67 sequencing reads are currently emphasized to resolve structural variants without an upper  
68 limit and to facilitate *A. thaliana* pan-genomics. Even a fully complete Col-0 genome  
69 sequence would not reveal the entire diversity of this species, as this accession is assumed  
70 to have a relatively small genome compared to other *A. thaliana* accessions.

71 The strong increase in the length of sequencing reads that was technically realized during  
72 the last years is enabling new assembly approaches [17, 18]. Despite the high error rate of  
73 'Single Molecule, Real Time' (SMRT) sequencing, the long reads significantly improve  
74 contiguity of *de novo* assemblies due to an efficient correction of the almost unbiased errors  
75 [19-21], provided that sufficient read coverage is available. SMRT sequencing offered by  
76 PacBio results routinely in average read lengths above 10 kbp [10, 22, 23]. These long reads  
77 were incorporated into high quality hybrid assemblies involving Illumina short read data [14,  
78 23], but increasing sequencing output supports the potential for so called 'PacBio only  
79 assemblies' [10-12, 15, 20]. Oxford Nanopore Technology's (ONT) sequencing provides

80 even longer reads with recent reports of longest reads over 2 Mbp [24]. However, the error  
81 rate of 5-15% [25] is still an issue in plant genome assembly although it has been shown that  
82 a high contiguity assembly is possible for *A. thaliana* [15].

83

84 Here we provide a SMRT sequencing-based *de novo* genome assembly of Nd-1 comprising  
85 contigs of chromosome-arm size anchored to chromosomes and oriented within  
86 pseudochromosome sequences based on genetic linkage information. The application of  
87 long sequencing reads abolished limitations of short read mapping and short read  
88 assemblies for genome sequence comparison. Based on this genome sequence assembly,  
89 we identified genomic rearrangements between Col-0 and Nd-1 ranging from a few kbp up to  
90 one Mbp. Gene duplications between both accessions as well as lineage specific genes in  
91 Nd-1 and Col-0 were revealed by this high quality sequence. The current assembly version  
92 outperforms the Illumina-based version (AthNd-1\_v1) about 75 fold with respect to assembly  
93 contiguity calculated with respect to number of contigs [13] and is in the same quality range  
94 as the recently released Ler and KBS-Mac-74 genome sequence assembly [14, 15].

95

## 96 **Methods**

### 97 **Plant material**

98 Niederzenz-1 (Nd-1) seeds were obtained from the European Arabidopsis Stock Centre  
99 (NASC; stock number N22619). The DNA source was the same as described earlier [13].

100

### 101 **DNA extraction**

102 The DNA isolation procedure was based on previously published protocols [12, 26] and  
103 started with 5 g of frozen leaves which were homogenized by grinding. Samples were mixed  
104 in a 1:10 ratio with extraction buffer (300mM Tris pH8.0, 25mM EDTA, 2M NaCl, 2%  
105 polyvinylpyrrolidone (PVP), 2% hexadecyltrimethylammonium bromide (CTAB)) and  
106 incubated at 65°C for 30 minutes with six inversions to mix the samples again. After five  
107 minute spinning at 5,000xg, the supernatant was transferred and mixed with one volume of

108 chlorophorm/isoamylalcohol (24:1). Again, the upper phase was transferred after repetition of  
109 the centrifugation step for ten minutes. RNA was removed by adding 30 $\mu$ L RNase A  
110 (10mg/ml) and incubation for 30 minutes at 37°C. Addition of chlorophorm/isoamylalcohol,  
111 centrifugation and transfer of supernatant were repeated. One volume of isopropanol and 0.1  
112 volumes of 3M NaOAc (pH 5.2) were added and mixed. DNA was precipitated by incubating  
113 at -80°C for 30 minutes and spinning for 45 minutes at 5,000g and a final ethanol wash step  
114 was performed. Finally, 500 $\mu$ l of 10mM Tris/HCl (pH 8.0) were added and samples were  
115 incubated over night at 4°C for resuspension.

116

### 117 **Library preparation and sequencing**

118 Sequencing was performed using PacBio RS II (Menlo Park, CA, USA). Five microgram high  
119 molecular weight DNA without further fragmentation was used to prepare a SMRTbell library  
120 with PacBio SMRTbell Template Prep Kit 1 (Pacific Biosciences, Menlo Park, CA, USA)  
121 according to the manufacturer's recommendations. The resulting library was size-selected  
122 using a BluePippin system (Sage Science, Inc. Beverly, MA, USA) to enrich for molecules  
123 larger than 11 kbp. The recovered library was again damage repaired and then sequenced  
124 on a total of 25 SMRT cells with P6-C4v2 chemistry and by MagBead loading on the PacBio  
125 RSII system (Pacific Biosciences, Menlo Park, CA, USA) with 360 min movie length.

126

### 127 **Assembly parameters**

128 A total of 1,972,766 subreads with an N50 read length of 15,244 bp and containing  
129 information about 16,798,450,532 bases were generated. Assuming a genome size of  
130 150 Mbp, the data cover the genome at 112 fold.

131 Read sequences derived from the plastome [GenBank: AP000423.1] or chondrome  
132 [GenBank: Y08501.2] were extracted from the raw data set by mapping to the respective  
133 sequence of Col-0 as previously described [27]. Canu v1.4 [28] was used for the assembly of  
134 the organelle genome sequences. Beside default parameters, a genome size of 0.37 Mbp for  
135 the chondrome and 0.167 Mbp for the plastome were assumed in the assembly process.

136 Scaffolding of initial contigs was performed with SSPACE-LongRead v1.1 [29] (parameters: -  
137 a 0 -i 70 -k 3 -r 0.3 -o 10). The quality of both assemblies was checked by mapping of NGS  
138 reads from Nd-1 [13] and Col-0 [30]. Manual inspection and polishing with Quiver v.2.3.0  
139 (parameters: minSubReadLength=500, readScore>80, minLength=500, maxHits=10,  
140 maxDivergence=30, minAnchorSize=12, --maxHits=1, --minAccuracy=0.75, --minLength=50,  
141 --algorithmOptions="-useQuality", --seed=1, --minAccuracy=0.75, --minLength=50, --  
142 concordant --algorithmOptions="-useQuality", minConfidence=40, minCoverage=5,  
143 diploidMode=False) [12] led to the final sequences. The start of the Nd-1 plastome and  
144 chondrome sequences was set according to the corresponding Col-0 sequences to ease  
145 comparisons. Finally, small assembly errors were corrected via CLC basic variant detection  
146 (ploidy=2; coverage<100,000; minimalCoverage=5; minimalCount=5;  
147 minimalFrequency=0.2) based on mapped Illumina paired-end reads (SRX1683594, [13])  
148 and PacBio reads. Sequence properties like GC content and GC skew were determined and  
149 visualized by CGView [31].

150

151 A total of 166,600 seed reads spanning 4,500,092,354 nt (N50 = 26,295 nt) and covering the  
152 expected 150 Mbp genome sequence were used for the assembly thus leading to a  
153 coverage of 30 fold (see AdditionalFile1 for details). Assemblies were performed with Canu  
154 [28], FALCON [12], Flye [32] and miniasm [33]. These assemblies are named Ath-Nd1\_v2  
155 with an assembler specific suffix (c=Canu, f=FALCON, y=Flye, m=miniasm).

156

157 All available SMRT sequencing reads were subjected to Canu v.1.7.1 [28] for read  
158 correction, trimming, overlap detection, and *de novo* assembly (see AdditionalFile2 for  
159 parameters). Contigs with a length below 50 kb were discarded to avoid artifacts. The  
160 remaining contigs were checked for contaminations with bacterial sequences and organell  
161 genome sequences as previously described [13]. BWA-MEM v0.7.13 [34] was used with  
162 default parameters and -m to map Nd-1 Illumina reads [13] to this assembly. The coverage  
163 depth was extracted from the resulting BAM file as previously described [35] and used to

164 support the identification of plastome and chondrome contigs. Pilon v.1.22 [36] was applied  
165 twice with default parameters for polishing of the assembly.

166

167 Release version 1.7.5 of the FALCON assembler  
168 <https://github.com/PacificBiosciences/FALCON/> [12] was used for a *de novo* assembly (see  
169 AdditionalFile3 for parameters) of the nuclear genome sequence. Resulting contigs were  
170 checked for contaminations with bacterial sequences and organell genome sequences as  
171 previously described [13]. Small fragments with low coverage were removed prior to  
172 polishing and error correction with Quiver [12].

173

174 SMRT sequencing reads were corrected via Canu and subjected to miniasm v0.3-r179 [33]  
175 for *de novo* assembly. Illumina reads were mapped with BWA-MEM for two rounds of  
176 polishing via Pilon as described above. The assembly was reduced to nucleome contigs as  
177 described above for the Canu assembly.

178

179 Flye v2.3.1 [32] was deployed on the subreads with an estimated genome size of 150 Mbp.  
180 The resulting assembly was polished twice via Pilon and reduced to nucleome contigs as  
181 described above for the Canu assembly.

182

### 183 **Construction of pseudochromosomes based on genetic information**

184 All 26 contigs of Ath-Nd1\_v2f were sorted and orientated based on genetic linkage  
185 information derived from 63 genetic markers (AdditionalFile4, AdditionalFile5,  
186 AdditionalFile6), which were analyzed in about 1,000 F2 plants, progeny of a reciprocal  
187 cross of Nd-1xCol-0 and Col-0xNd-1. Genetic markers belong to three different types: (1)  
188 fragment length polymorphisms, which can be distinguished by agarose gel electrophoresis,  
189 (2) small nucleotide polymorphisms which can be distinguished by Sanger sequencing and  
190 (3) small nucleotide polymorphisms, which were identified by high resolution melt analysis.  
191 Design of oligonucleotides was performed manually and using Primer3Plus [37]. DNA for

192 genotyping experiments was extracted from *A. thaliana* leaf tissue using a  
193 cetyltrimethylammonium bromide (CTAB) based method [38]. PCRs were carried out using  
194 GoTaq<sup>®</sup> G2 DNA Polymerase (Promega) generally based on the suppliers' protocol. The  
195 total reaction volume was reduced to 15 µl and only 0.2u of the polymerase were used per  
196 reaction. Sizes of amplicons generated were checked on agarose gels. If required, samples  
197 were purified for sequencing by ExoSAP-IT (78201.1.ML ThermoFisher Scientific) treatment  
198 as previously described [39]. Sanger sequencing on ABI3730XL was applied to identify  
199 allele-specific SNPs for the genotyping. Manual inspection of gel pictures and  
200 electropherograms lead to genotype calling. High resolution melt analysis was performed on  
201 a CFX96 Touch Real-Time PCR Detection System (BioRad) using the Precision Melt  
202 Supermix according to suppliers instructions (BioRad). All data were combined, processed  
203 by customized Python scripts to calculate recombination frequencies between genetic  
204 markers. Linkage of genetic markers provided information about relationships of assembled  
205 sequences. The north-south orientation of the chromosomes was transferred from the  
206 reference sequence based on RBH support. Contigs were joined to pseudochromosome  
207 sequences for Ath-Nd1\_v2f. Subsequently, positioning was transferred to the 69 contigs of  
208 the Canu assembly to create pseudochromosome sequences for AthNd-1\_v2c  
209 (AdditionalFile7).

210 Since the assemblies except Ath-Nd1\_v2f contain smaller contigs in the range of 50-100 kb,  
211 anchoring via genetic linkage information was not feasible. Therefore, these short contigs  
212 were placed based on best BLASTn matches to the TAIR9 sequence (i.e. the Col-0 gold  
213 standard reference) and integrated into pseudochromosomes as previously described [13].  
214 We used the TAIR9 sequence, because this is the sequence basis of the structural and  
215 functional annotation provided in TAIR9, TAIR10 and Araport11 [40]. Since the total length of  
216 the TAIR9 sequence exceeds that of the Ath-Nd1\_v2f assembly, TAIR9 was selected for  
217 anchoring of small contigs.

218 All data generated is stored in the PGP Repository [41] and is accessible via DOI  
219 (<https://doi.org/10.5447/IPK/2019/4>).



220

### 221 **Genome structure investigation**

222 Characteristic elements of the Nd-1 genome sequence were annotated by mapping of known  
223 sequences as previously described [13]. Fragments and one complete 45S rDNA unit were  
224 discovered based on gi|16131:848-4222 and gi|16506:88-1891. AF198222.1 was subjected  
225 to a BLASTn search for the identification of 5S rDNA sequences. Telomeric repeats were  
226 used to validate the assembly completeness at the pseudochromosome end as well as  
227 centromere positions as previously described [13].

228

### 229 **BUSCO analysis**

230 BUSCO v3 [42] was run with default parameters on the Nd-1 pseudochromosomes and on  
231 the TAIR9 reference sequence to produce a gold standard for Arabidopsis. The  
232 'embryophyta\_odb9' was used as reference gene set.

233

### 234 **Genome sequence alignment**

235 Nd-1 pseudochromosome sequences were aligned to the Col-0 gold standard sequence [2]  
236 via nucmer [43] and NucDiff [44] based on default parameters of NucDiff. Customized Python  
237 scripts were deployed to process the results and to assess the genome-wide distribution of  
238 differences. Spearman correlation coefficient was calculated using the implementation in the  
239 Python module scipy to validate the indication of increased numbers of SV around the  
240 centromeres.

241

### 242 **Gene prediction and RBH analysis**

243 AUGUSTUS v.3.3 [45] was applied to all four Nd-1 assemblies with previously optimized  
244 parameters [39]. Afterwards, the identification of RBHs at the protein sequence level  
245 between Nd-1 and Col-0 (Araport11, representative peptide sequences) was carried out with  
246 a custom Python script as previously described [13]. Additionally, gene prediction was run on  
247 the Col-0 gold standard sequence [2] as well as on the *Ler* chromosome sequences [14].

248 Parameters were set as described before to generate two control data sets. Previously  
249 trimmed and filtered ESTs [13, 46] were matched via BLASTn [47] to the predicted mRNAs.  
250 Pairwise global alignments were constructed via MAFFT v.7.299b [48] to validate the  
251 annotation quality. After processing GCA\_000001735.1 (Col-0), Ath-Nd-1\_v2c, and the most  
252 recent *Ler* assembly [14] with RepeatMasker v4 [49] all three were subjected to cactus [50]  
253 for alignment. CAT [51] was run on this alignment with the ENSEMBL v42 annotation of Col-  
254 0.

255 INFERNAL [52] with default parameters and based on Rfam13 [53] was applied to detect  
256 various non-coding RNA genes. In addition, tRNAscan-SE [54] with -G option was deployed  
257 to identify tRNA and rRNA genes. Overlaps between both methods were analysed.

258

#### 259 **Transposable element annotation**

260 All annotated transposable element (TE) sequences of Araport11 (derived from TAIR) [40]  
261 were mapped via BLASTn to the Nd-1 assembly AthNd-1\_v2c and against the Col-0 gold  
262 standard sequence. The top BLAST score for each element in the mapping against the  
263 TAIR9 reference sequence was identified. All hits against Nd-1 with at least 90% of this top  
264 score were considered for further analysis. Overlapping hits were removed to annotate a final  
265 TE set. All predicted Nd-1 genes which overlapped TEs with more than 80% of their gene  
266 space were flagged as putative protein encoding TE genes. In addition, RepeatModeler  
267 v.1.0.11 [49] was deployed to identify novel TE sequences in the Ath-Nd-1\_v2c assembly.

268

#### 269 **Identification of gene space differences**

270 Genes in insertions in Ath-Nd1\_v2c were searched in the Col-0 gold standard sequence and  
271 vice versa. The BLAST results on DNA and peptide level indicated the absence of any truly  
272 novel genes thus indicating that presence/absence variants are the results of gene  
273 duplications. Apparent gene space differences could be caused by regions missing in the  
274 assembly. Therefore, Nd-1 (SRR2919279, SRR3340908, SRR3340909) [13] and Col-0  
275 (SRR1810832, SRR1945757)[9, 30] Illumina sequencing reads were mapped to Ath-

276 Nd1\_v2c via BWA MEM [34] using the -m flag to discard spurious hits. The sequencing read  
277 coverage depth was calculated via bedtools by a customized Python script [35]. Average  
278 coverage per accession was calculated as the median of all coverage values. Per gene  
279 coverage was calculated as the median of all coverage values at positions in the respective  
280 gene and normalized to the average coverage of the respective accession. To correct for  
281 accession specific mapability differences caused e.g. by sequence divergence from Nd-1,  
282 the resulting values were additionally normalized to the median of all per gene ratios. Genes  
283 were considered as duplicated in Col-0 if their relative coverage in Nd-1 was below 50 % of  
284 the Col-0 value. Genes with Nd-1 values above 150 % of the Col-0 value were considered  
285 duplicated in Nd-1. These cutoff values were validated based on experimentally validated  
286 gene differences. Corresponding AGIs were identified via BLASTp to transfer the functional  
287 annotation of Araport11 if possible. Following this initial identification, putative TE genes  
288 were removed based on the annotation or the overlap with annotated TE sequences  
289 (AdditionalFile8), respectively.

290 Sequencing data of 1,135 *A. thaliana* accessions [9] were retrieved from the Sequence Read  
291 Archive, mapped to Ath-Nd1\_v2c, and processed as described above. Accessions displaying  
292 an average coverage below 10x were excluded from the following identification of  
293 presence/absence variations. GeneSet\_Nd-1\_v2.0 genes were classified as dispensable if  
294 their relative coverage was below 0.1 in more than 100 accessions. Remaining genes were  
295 classified as core or TE genes depending on their overlap with TE features (see TE  
296 annotation).

297

### 298 **Validation of rearrangements and duplications**

299 LongAmpTaq (NEB) was used for the generation of large genomic amplicons up to 18 kbp  
300 based on the suppliers' protocol. Sanger sequencing was applied for additional confirmation  
301 of generated amplicons. The amplification of small fragments and the following procedures  
302 were carried out with standard polymerases as previously described [13].

303

## 304 **Investigation of collapsed region**

305 The region around At4g22214 was amplified in five overlapping parts using the Q5 High  
306 Fidelity polymerase (NEB) with genomic DNA from Col-0. Amplicons were checked on  
307 agarose gels and finally cloned into pCR2.1 (Invitrogen) or pMiniT 2.0 (NEB), respectively,  
308 based on the suppliers' recommendations. Cloned amplicons were sequenced on an  
309 ABI3730XL by primer walking. Sequencing reads were assembled using CLC  
310 GenomicsWorkbench (v. 9.5 CLC bio). In addition, 2x250 nt paired-end Illumina reads of  
311 Col-0 [30] were mapped to correct small variants in the assembled contigs and to close a  
312 small gap between cloned amplicons.

313

## 314 **Analysis of gaps in the Col-0 reference sequence**

315 Flanking sequences of gaps in the Col-0 gold standard sequence were submitted to a  
316 BLASTn against the Ath-Nd-1\_v2c genome sequence. Nd-1 sequences enclosed by hits of  
317 pairs of 30 kbp long flanking sequences from Col-0 were extracted. Homotetramer  
318 frequencies were calculated for all sequences and compared against the frequencies in  
319 randomly picked sequences. A Mann-Whitney U test was applied to analyze the difference  
320 between both groups.

321

## 322 **Results**

### 323 **Nd-1 genome**

324 Assemblies generated via Canu (Ath-Nd-1\_v2c), FALCON (Ath-Nd-1\_v2f), Flye (Ath-Nd-  
325 1\_v2y) and miniasm (Ath-Nd-1\_v2m) were compared based on numerous assembly  
326 statistics (Table 1). AthNd-1\_v2f exceeds the previously reported assembly version AthNd-  
327 1\_v1 by 2.5 Mbp, while reducing the number of contigs by a factor of about 200 to 26. AthNd-  
328 1\_v2c is adding 6.9 Mbp to the previous assembly version (AthNd-1\_v1), but at the expense  
329 of a higher number of contigs than AthNd-1\_v2f. We selected AthNd-1\_v2c that contains 69  
330 contigs as the representative assembly which should be used for further comparison.

331 Pseudochromosomes of AthNd-1\_v2f were constructed truly *de novo* from 3-7 contigs based  
332 on genetic linkage information. All 26 contigs were anchored based on 63 genetic markers,  
333 but precise positioning around the centromeres of chromosome 4 and 5 was ambiguous.  
334 Pseudochromosomes reach similar lengths as the corresponding chromosome sequences in  
335 the Col-0 gold standard sequence. Pseudochromosomes for AthNd-1\_v2c were generated  
336 by transferring orientation and position of large contigs from AthNd-1\_v2f. The Nd-1 genome  
337 sequence AthNd-1\_v2c contains complete 45S rDNA units on pseudochromosome 3 as well  
338 as several fragments of additional 45S rDNA units on all other pseudochromosomes (Fig. 1).  
339 Centromeric and telomeric repeat sequences as well as 5S rDNA sequences were detected  
340 at centromere positions. Completeness of most assembled sequences was confirmed by the  
341 occurrence of telomeric repeat sequences (Fig. 1). The high assembly quality and  
342 completeness of AthNd-1\_v2c is supported by the detection of 98.2% of all embryophyta  
343 BUSCO genes – even one more than detected in the Col-0 gold standard sequence  
344 (AdditionalFile9).

345 The plastome and chondrome sequences comprise 154,443 bp and 368,216 bp, respectively  
346 (<https://doi.org/10.5447/IPK/2019/4>). A total of 148 small variants were identified from a  
347 global alignment between the Nd-1 and Col-0 plastome sequences. General sequence  
348 properties like GC content and GC skew (AdditionalFile10, AdditionalFile11) are almost  
349 identical to the plastome and chondrome of Col-0. Nevertheless, there are some  
350 rearrangements between the chondrome sequences of Nd-1 and Col-0.

351

### 352 **Genome structure differences**

353 Sequence comparison between AthNd-1\_v2c and the Col-0 gold standard sequence  
354 revealed a large inversion on chromosome 4 involving about 1 Mbp (Fig. 2). The left break  
355 point is at 1,637,889 bp and the right break point at 2,708,850 bp on chromosome 4  
356 (NdChr4). The inverted sequence is 120 kbp shorter than the corresponding Col-0 sequence.  
357 PCR amplification of both inversion borders (AdditionalFile12) and Sanger sequencing of the  
358 generated amplicons was used to validate this rearrangement.

359 The recombination frequency in this region was analyzed using the marker pair M84/M74.  
360 Only a single recombination was observed between these markers while investigating 60  
361 plants. Moreover, only 8 recombination events in 108 plants were observed between another  
362 pair of markers, spanning a larger region across the inversion (AdditionalFile5). In contrast,  
363 the average recombination frequency per Mbp at the corresponding position on other  
364 chromosomes was between 12%, observed for M31/M32, and 18%, observed for M13/M14.  
365 Statistical analysis revealed a significant difference in the recombination frequencies  
366 between the corresponding positions on different chromosomes ( $p < 0.001$ , `prop.test()` in R)  
367 supporting a reduced recombination rate across the inversion on chromosome 4.  
368 Comparison of a region on Chr2, which is probably of mitochondrial origin (mtDNA), in the  
369 Col-0 gold standard sequence with the Nd-1 genome sequence revealed a 300 kbp highly  
370 divergent region (Fig. 3). Sequences between position 3.20 Mbp and 3.29 Mbp on NdChr2 of  
371 AthNd-1\_v2c display low similarity to the Col-0 gold standard sequence, while there is almost  
372 no similarity between 3.29 Mbp and 3.48 Mbp. However, the length of both regions is roughly  
373 the same. Comparison against the Ler genome sequence assembly revealed the absence of  
374 the entire region between 3.29 Mbp and 3.48 Mbp on chromosome 2. The Nd-1 sequence  
375 from this region lacks continuous similarity to any other region in the Col-0 or Nd-1 genome  
376 sequence. However, the 28 genes encoded in this region in Nd-1 show weak similarity to  
377 other Arabidopsis genes. Comparison of gene space sequences from this region against the  
378 entire Nd-1 assembly revealed some similarity on chromosome 3, 4, and 5  
379 (AdditionalFile13).

380 An inversion on chromosome 3 of 170 kbp which was described between Col-0 and Ler [14]  
381 is not present in Nd-1. The sequence similarity between Col-0 and Nd-1 is high in this region.  
382 In total, 2206 structural variants larger than 1 kbp were identified between Col-0 and Nd-1.  
383 The genome-wide distribution of these variants indicated a clustering around the  
384 centromeres (AdditionalFile14). A Spearman correlation coefficient of  $-0.79$  ( $p = 1.1 \times 10^{-27}$ ) was  
385 calculated for the correlation of the number of SVs in a given interval and the distance of this  
386 interval to the centromere (AdditionalFile15). Therefore, these large structural variants are

387 significantly more frequent in the centromeric and pericentromic regions. A total, of 148 new  
388 regions larger than 1 kbp were identified in Ath-Nd-1\_v2c. These regions are also more  
389 frequent in proximity of the centromere ( $r=-0.43$ ,  $p=3.7*10^{-7}$ ).

390

### 391 **Hint-based gene prediction**

392 Hint-based gene prediction using AUGUSTUS with the *A. thaliana* species parameter set on  
393 the Nd-1 pseudochromosomes resulted in 30,126 nuclear protein coding genes  
394 (GeneSet\_Nd-1\_v2.0) with an average predicted transcript length of 1,798 bp, an average  
395 predicted CDS length of 1,391 bp and an average exon number per predicted transcript of  
396 5.98. The number of predicted genes is reduced compared to the GeneSet\_Nd-1\_v1.1 [39]  
397 by 708 genes. In total, 28,042 (93%) representative peptide sequences were matched to  
398 Araport11 sequences and functionally annotated based on Araport11 information  
399 (AdditionalFile16). As controls we ran the gene prediction with same parameters on Col-0  
400 and Ler pseudochromosome sequences resulting in 30,352 genes and 29,302 genes,  
401 respectively. There were only minor differences concerning the average transcript and CDS  
402 length as well as the number of exons per gene.

403 Based on 35,636 TEs detected in Nd-1 (AdditionalFile8) 2,879 predicted Nd-1 genes were  
404 flagged as putative TE genes (AdditionalFile17, AdditionalFile18). This number matches well  
405 with the difference between the predicted genes in Nd-1 and the annotated protein coding  
406 genes in Araport11, which is supposed to be free of TE genes. The predicted mRNAs were  
407 supported by ESTs, which matched almost perfectly with an average similarity of 98.7%  
408 (AdditionalFile19). Additionally, the assembly was screened for TEs resulting in 613  
409 consensus sequences of TE families (<https://doi.org/10.5447/IPK/2019/4>).

410 The comparative gene prediction with CAT resulted in 26,717 genes in Nd-1 and 26,681  
411 genes in Ler, respectively. Average CDS lengths were 1,292 bp and 1,291 bp, respectively.  
412 Since TE associated genes were not identified in this gene prediction process, the number of  
413 gene models predicted by CAT should be smaller than the number predicted by

414 AUGUSTUS. The difference of 3,409 for Nd-1 is slightly exceeding the number of 2,879 TE  
415 associated genes that were predicted and flagged in the GeneSet\_Nd-1\_v2.0.  
416 Besides protein encoding genes, 557 tRNA genes and 963 rRNA genes were predicted by  
417 INFERNAL. Comparison of these predicted tRNA genes to the result of tRNAscan-SE  
418 revealed an overlap of 552 (96.5%).

419

#### 420 **Detection of gene space differences between Nd-1 and Col-0**

421 A BLASTp-based comparison of all predicted Nd-1 peptide sequences and Col-0 Araport11  
422 representative peptide sequences in both directions revealed 24,453 reciprocal best hits  
423 (RBHs) (AdditionalFile20). In total, 89.1% of all 27,445 nuclear Col-0 genes are represented  
424 in this RBH set. Analysis of the colinearity of the genomic location of all 24,453 RBHs (see  
425 AdditionalFile21 for a list) between Nd-1 and Col-0 showed overall synteny of both genomes  
426 as well as an inversion on chromosome 4 (AdditionalFile22). While most RBHs are properly  
427 flanked by their syntenic homologs and thus lead to a diagonal positioning of points in the  
428 scatter plot, there are 345 outliers (see AdditionalFile23 for a list). In general, outliers occur  
429 frequently in regions around the centromeres. Positional analysis revealed an involvement of  
430 many outliers in the large inversion on chromosome 4. An NGS read mapping at the  
431 positions of randomly selected outliers was manually inspected and indicated  
432 rearrangements between Nd-1 and Col-0. Structural variants, which affect at least three  
433 different genes in a row of RBH pairs, were identified from the RBH analysis. Examples  
434 beside the previously mentioned 1.2 Mbp inversion on chromosome 4 (*At4g03820-*  
435 *At4g05497*) are a translocation on chromosome 3 (*At3g60975-At3g61035*) as well as an  
436 inversion on chromosome 3 around *At3g30845*.

437 As a control we identified 25,556 (91.1%) RBHs between our gene prediction on Col-0 and  
438 the manually curated reference annotation Araport11. In addition, 24,329 (88.6 %) RBHs  
439 were identified between our gene prediction on the Ler assembly and the Col-0 annotation  
440 Araport11.



441 In total, 947 protein encoding genes in Nd-1 (AdditionalFile24) were detected to be copies of  
442 only 421 genes annotated in Araport11. *SEC10* (*At5g12370*) [5] was previously described as  
443 an example for a tandem gene duplication collapsed in the Col-0 gold standard sequence.  
444 However, this region was already properly represented in AthNd-1\_v1 [13]. Gene  
445 duplications of *At2g06555* (unknown protein), *At3g05530* (*RPT5A*) and *At4g11510*  
446 (*RALFL28*) in Nd-1 were confirmed by PCR amplification and Sanger sequencing of the  
447 sequences enclosed by both copies as well as through amplification of the entire event locus.  
448 On the other hand, there are 383 predicted genes in Nd-1 (AdditionalFile25) which appeared  
449 at least duplicated in Col-0.

450

#### 451 **Pan-genomic analyses**

452 Presence/absence variations of genes were inspected across a panel of 964 additional  
453 accessions (<https://doi.org/10.5447/IPK/2019/4>). In total, 25,809 genes were present in  
454 almost all accessions, 1,438 genes were considered dispensable, and the remaining 2,879  
455 genes were flagged as TE or at least TE-associated genes (AdditionalFile26). Dispensable  
456 genes and TE genes are frequently located in proximity of the centromeres while core genes  
457 are less frequent in these regions (AdditionalFile27).

458

#### 459 **Hidden locus in Col-0**

460 *At4g22214* was identified as a gene duplicated in Nd-1 in our analysis. During experimental  
461 validation, we did not detect the expected difference between Col-0 and Nd-1 DNA  
462 concerning the locus around *At4g22214*. However, the PCR results from Col-0 matched the  
463 expectation based on the Nd-1 genome sequence thus suggesting a collapsed gene  
464 sequence in the Col-0 gold standard sequence. This hypothesis was supported by PCR  
465 results with outwards facing primers (Fig. 4). Cloning of the *At4g22214* region of Col-0 in five  
466 overlapping fragments was done to enable Sanger sequencing. The combination of Sanger  
467 and paired-end Illumina sequencing reads revealed a tandem duplication with modification of  
468 the original gene (Fig. 4). The copies were designated *At4g22214a* and *At4g22214b* based

469 on their position in the genome (GenBank: MG720229). While At4g22214b almost perfectly  
470 matches the Araport11 annotation of At4g22214, a significant part of the CDS of At4g22214a  
471 is missing. Therefore, the gene product of this copy is probably functionless. At4g22214 is  
472 annotated as defensin-like family protein [55]. Since the family of defensin proteins contains  
473 about 300 members in *A. thaliana* [55], the functional implications of this duplication are  
474 probably low.

475

#### 476 **Gaps in the Col-0 reference sequence**

477 Despite its very high quality, the Col-0 gold standard sequence contains 92 N stretches of  
478 various sizes representing regions of unknown sequence like the NOR clusters or  
479 centromeres. Ath-Nd1\_v2c enabled the investigation of some of these sequences based on  
480 homology assumptions. A total of 13 Col-0 gold standard sequence gaps were spanned with  
481 high confidence by Ath-Nd1\_v2c and therefore selected for homopolymer frequency  
482 analysis. The corresponding regions in Nd-1 are significantly enriched with homopolymers in  
483 comparison to randomly picked control sequences ( $p=0.000048$ , Mann-Whitney U test)  
484 (AdditionalFile28).

485

486

#### 487 **Discussion**

##### 488 **Genome structure of the *A. thaliana* accession Nd-1**

489 In order to further investigate large variations in the range of several kbp up to several Mbp  
490 between *A. thaliana* accessions, we performed a *de novo* genome assembly for the Nd-1  
491 accession using long sequencing reads and cutting-edge assembly software. Based on  
492 SMRT sequencing reads the assembly contiguity was improved by over 75 fold considering  
493 the number of contigs in the previously released NGS-based assembly [13]. Assembly  
494 statistics were comparable to other projects using similar data [11, 14, 15, 56, 57]. Despite  
495 the very high contiguity, regions like NORs still pose a major challenge. These regions are  
496 not just randomly clustered repeats, but highly controlled and ordered repetitions of

497 sequences [58]. Therefore, the identification of accession-specific sequence differences  
498 could explain phenotypic differences. NOR repeat unit sequences in Ath-Nd-1\_v2c are  
499 located on the distal short arm of NdChr2 and NdChr4. This NOR position matches the  
500 situation in Col-0 where the NOR2 is located distal on the short arm of Chr2 [59] and NOR4  
501 on Chr4, respectively. In addition to NORs, the assembly of chromosome ends remains still  
502 challenging, since the absence of some telomeric sequences in a high quality assembly was  
503 observed before [14]. Despite the absence of challenging repeats, regions close to the  
504 telomeres including the genes At3g01060 (BUSCO ID: EOG09360D4T) and At5g01010  
505 (BUSCO ID: EOG09360DFK) were not assembled by FALCON (Ath-Nd1\_v2f) although  
506 sequence reads covering these regions were present in the input data. Also, this region is  
507 represented in the Canu assembly (Ath-Nd1\_v2c). In our hands, Canu performed best for the  
508 assembly of the Nd-1 genome sequence based on SMRT sequencing reads.

509

#### 510 **Nuclear genome sequence differences**

511 The increased contiguity of this long read assembly was necessary to discover an  
512 approximately 1 Mbp inversion through sequence comparison as well as RBH analysis. An  
513 earlier Illumina short read based assembly [13] lacked sufficient contiguity in the region of  
514 interest to reveal both breakpoints of this variant between Col-0 and Nd-1. The large  
515 inversion at the north of NdChr4 relative to the Col-0 gold standard sequence turned out as a  
516 modification of the allele originally detected in *Ler* [14, 60]. However, the Nd-1 allele is  
517 different from the *Ler* allele. This could explain previous observations in several hundred  
518 *A. thaliana* accessions, which share the left inversion border with *Ler*, but show a different  
519 right inversion border [14].

520 Despite the very much improved contiguity and selection of Canu as the currently best  
521 assembler for our dataset, there are only very small parts of pericentromeric sequences  
522 represented in the Ath-Nd1\_v2c assembly. Centromeric regions with an estimated size of 5  
523 Mbp each [13] were not assembled. The absence of these highly repetitive regions is in  
524 agreement with previous findings that the error rate of long reads is still too high to resolve

525 NORs and centromeres [15]. The pericentromeric and to a very limited extent also  
526 centromeric regions are represented in the assemblies by small contigs of 50-100 kbp.  
527 However, absence of telomeric sequences from some pseudochromosome ends was  
528 observed before even for a very high quality assembly [14, 15]. The detected telomeric  
529 repeats at the centromere positions support previously reported hypothesis about the  
530 evolution of centromeres out of telomere sequences [61], and telomeric or centromeric  
531 sequences, respectively, at the end of at least some pseudochromosomes indicated the  
532 completeness of the Ath-Nd1\_v2c assembly at these points. However, almost 20 years after  
533 the release of the first chromosome sequences of *A. thaliana*, we are still not able to  
534 assemble complete centromere sequences continuously.

535 Sequence differences observed on the short arm of chromosome 2 between Col-0 and Nd-1  
536 could be due to the integration of mtDNA into the Chr2 of Col-0 [2]. This region was reported  
537 to be collapsed in the Col-0 reference genome sequence, harboring in real Col-0 DNA about  
538 600 kbp of mtDNA instead of the 270 kbp represented in the reference genome sequence  
539 [62]. However, Ath-Nd1\_v2c did not reveal such a duplication of the mtDNA located in  
540 NdChr2. Filtering of plastome and chondrome sequences during the polishing process could  
541 be responsible for this if copies would have been fragmented into multiple contigs. In this  
542 region, we detected in the Nd-1 assembly sequences unrelated to the corresponding Col-0  
543 reference sequence. Since Nd-1 genes of this region show similarity to gene clusters on  
544 other chromosomes, they could be relics of a whole genome duplication as reported before  
545 for several regions of the Col-0 gold standard sequence [63]. This difference on Chr2 at  
546 about 3.2 to 3.5 Mbp is only one example for a large variant region between Col-0 and Nd-1.  
547 Similar though shorter such differences detected around centromeres could be explained by  
548 TEs and pseudogenes which were previously reported as causes for intra-species variants in  
549 these regions [62, 64].

550

551 **Plastome and chondrome**

552 Size and structure of the Nd-1 plastome is very similar to Col-0 [2] or *Ler* [27]. In accordance  
553 with the overall genome similarities, the observed number of small differences between the  
554 plastome sequences of Col-0 and Nd-1 is slightly higher than the value reported before for  
555 the Col-0 comparison to *Ler* [27].

556 The size of the Nd-1 chondrome matches previously reported values for the large chondrome  
557 configuration of other *A. thaliana* accessions [65]. Large structural differences between the  
558 Col-0 chondrome [65] and the Nd-1 chondrome could be due to the previously described  
559 high diversity of this subgenome including the generation of substoichiometric DNA  
560 molecules [66, 67]. In addition, the mtDNA level was reported to differ between cell types or  
561 cells of different ages within the same plant [68, 69]. The almost equal read coverage of the  
562 assembled Nd-1 chondrome could be explained by the young age of the plants at the point of  
563 DNA isolation, as the amount of all chondrome parts should be the same in young leaves [69].

564

#### 565 **Nd-1 gene space**

566 Many diploid plant genomes contain on average around or even slightly below 30,000 protein  
567 encoding genes [35, 70] with the *A. thaliana* genome harboring 27,445 nuclear protein-  
568 coding genes according to the most recent Araport11 annotation [40]. Since there are only  
569 few other chromosome-level assembly sequences of *A. thaliana* available at the moment, we  
570 do not know the precise variation range of gene numbers between different accessions. The  
571 number of 27,247 predicted non-TE genes in Nd-1 is further supported by the identification of  
572 24,453 RBHs with the Araport11 [40] annotation of the Col-0 gold standard sequence. This  
573 number exceeds the matches between Col-0 and *Ler*-0 [14]. Our chromosome-level  
574 assembly further enhances the gene prediction quality since at least 89.1% of all Col-0  
575 genes were recovered. This reinforces previous studies that also reported annotation  
576 improvements through improved assembly quality [71]. The number of 35,636 TEs annotated  
577 for Ath-Nd-1\_v2c exceeds the number of 33,892 such elements identified in the previous  
578 assembly version Ath-Nd-1\_v1 [13] by 1,744. Such an increase in the number of resolved

579 TEs as well as an improved assembly of TE-rich regions in assemblies based on long reads  
580 was reported before [72].

581 Due to the high proportion of genes within the *A. thaliana* genome assigned to paralogous  
582 groups with high sequence similarity [73, 74], we speculated that the identification of  
583 orthologous pairs via RBH analysis might be almost saturated. Gene prediction with the  
584 same parameters on the Col-0 gold standard genome sequence prior to RBH analysis  
585 supported this hypothesis. However, the incorporation of accession-specific RNA-Seq  
586 derived hints could further increase the accuracy of the Nd-1 gene prediction. Since there are  
587 even some RBHs at non-syntenic positions between the control Col-0 annotation and the  
588 Araport11 annotation, our Nd-1 annotation is already of very high accuracy. The precise  
589 annotation of non-canonical splice sites via hints as described before [39] contributed to the  
590 new GeneSet\_Nd-1\_v2.0. Gene duplication and deletion numbers, or accession specific  
591 presence/absence variations of genes, in Nd-1 and Col-0 are in the same range as  
592 previously reported values of up to a few hundred [14, 75]. Since we were searching  
593 genome-wide for copies of genes without requiring an annotated feature in each genome  
594 sequence, both numbers might include some pseudogenes due to the frequent occurrence of  
595 these elements within plant genomes [76, 77]. Since all comparisons rely on the constructed  
596 sequences we cannot absolutely exclude that a small number of other genes were detected  
597 as amplified due to collapsed sequences similar to SEC10 (At5g12370) [5]. Removing TE  
598 genes based on sequence similarity to annotated features reduced the proportion of putative  
599 pseudogenes. However, it is impossible to unequivocally distinguish between real genes and  
600 pseudogenes in all cases, because even genes with a premature stop codon or a frameshift  
601 mutation could function as a truncated version or give rise to regulatory RNAs [74, 77-79]. In  
602 addition, the impact of copy number variations involving protein encoding genes in *A.*  
603 *thaliana* might be higher than previously assumed thus supporting the existence of multiple  
604 gene copies [80]. Gene expression analysis could support the discrimination of  
605 pseudogenes, because low gene expression in *A. thaliana* was reported to be associated  
606 with pseudogenization [81]. Despite the unclear status of the gene product, the mere

607 presence of these sequences revealed fascinating insights into genome evolution and  
608 contributed to the pan-genome [82, 83].

609

## 610 **Conclusions**

611 We report a high quality long read *de novo* assembly (AthNd-1\_v2c) of the *A. thaliana*  
612 accession Nd-1, which improved significantly on the previously released NGS assembly  
613 sequence AthNd-1\_v1.0 [13]. Comparison of the GeneSet\_Nd-1\_v2.0 with the Araport11  
614 nuclear protein coding genes revealed 24,453 RBHs supporting an overall synteny between  
615 both *A. thaliana* accessions except for an approximately 1 Mbp inversion at the north of  
616 chromosome 4. Moreover, large structural variants were identified in the pericentromeric  
617 regions. Comparisons with the Col-0 gold standard sequence also revealed a collapsed  
618 locus around At4g22214 in Col-0. Therefore, this work contributes to the increasing  
619 *A. thaliana* pan-genome with significantly extended details about genomic rearrangements.

620

## 621 **List of abbreviations**

622 NGS next generation sequencing

623 NOR nucleolus organizing region

624 RBH reciprocal best hit

625 SMRT single molecule real time

626

627

## 628 **Declarations**

### 629 **Ethics approval and consent to participate**

630 Not applicable

631

### 632 **Consent for publication**

633 Not applicable

634

635 **Availability of data and materials**

636 The data sets supporting the results of this article are included within the article and its  
637 additional files. The Ath-Nd-1\_v2 assemblies (Table 1) and derived files are available as  
638 external downloads from <http://doi.org/10.5447/IPK/2019/4>. Sequencing reads were  
639 submitted to the SRA (SRP066294). Python scripts developed and applied for this study are  
640 available on github: [https://github.com/bpucker/Nd1\\_PacBio](https://github.com/bpucker/Nd1_PacBio)  
641 (<http://doi.org/10.5281/zenodo.2590750>).

642

643 **Competing interest**

644 The authors declare that they have no competing interest.

645

646 **Funding**

647 We acknowledge the financial support of the German Research Foundation (DFG) and the  
648 Open Access Publication Fund of Bielefeld University for the article processing charge. The  
649 funding body did not influence the design of the study, the data collection, the analysis, the  
650 interpretation of data, or the writing of the manuscript.

651

652 **Author's contributions**

653 BP, DH and BW conceived and designed research. BP, KS, KF, BH and RR conducted  
654 experiments. BP, DH and BW interpreted the data. BP and BW wrote the manuscript. All  
655 authors read and approved the final manuscript.

656

657 **Acknowledgements**

658 We thank Willy Keller for isolating high molecular DNA, Katharina Kemmet for extensive  
659 genotyping of plants for the genetic map, Helene Schellenberg, Ann-Christin Polikeit and  
660 Prisca Viehöver for Sanger sequencing, and Melanie Kuhlmann as well as Andrea Voigt for  
661 taking excellent care of the plants. We are very grateful for support from the Bioinformatics  
662 Resource Facility and Stefan Albaum.



663

664

665

666 **Additional Files**

667 **AdditionalFile1. Sequencing Statistics.**

668 Statistical information about the generated SMRT sequencing data for the *A. thaliana* Nd-1  
669 genome assembly are listed in this table. The expected genome size is based on several  
670 analyses reporting values around 150 Mbp [84, 85].

671

672 **AdditionalFile2. Canu assembly parameters.**

673 Listing of all parameters that were adjusted for the Canu assembly of the Nd-1 nucleome.  
674 While most default parameters were kept, some were specifically adjusted for this plant  
675 genome assembly.

676

677 **AdditionalFile3. FALCON assembly parameters.**

678 All parameters that were adjusted for the FALCON assembly of the Nd-1 nucleome are listed  
679 in this table. While most default parameters were kept, some were specifically adjusted for  
680 this plant genome assembly.

681

682 **AdditionalFile4. Molecular markers for genetic linkage analysis.**

683 All markers require the amplification of a genomic region using the listed oligonucleotides  
684 under the specified conditions (annealing temperature, elongation time). Depending on the  
685 fragment size differences, the resulting PCR products can allow the separation of both alleles  
686 by agarose gel electrophoresis (length polymorphism) or might require Sanger sequencing to  
687 investigate single SNPs.

688

689 **AdditionalFile5. Distribution of genetic markers over physical map.**

690 The positions of all genetic markers on the pseudochromosome sequences are illustrated.  
691 Assembled sequences were positioned based on the genetic linkage information. Some  
692 genetic marker combinations allowed the investigation of recombination frequencies within  
693 continuous sequences.

694

695 **AdditionalFile6. Oligonucleotide sequences for genetic linkage analysis.**

696 Sequences, names and recommended annealing temperatures of all oligonucleotides used  
697 in this work are listed in this table. Usage remarks for the oligonucleotides are provided as  
698 well.

699

700 **AdditionalFile7. Alignment of Ath-Nd1\_v2c and Ath-Nd1\_v2f.**

701 Assemblies generated by Canu and FALCON, respectively, were compared via BLASTn  
702 search of 10 kb sequence chunks. Color and position of dots in the figure indicate the  
703 position of the best hit on the respective sequence.

704

705 **AdditionalFile8. TE positions in the Nd-1 genome sequence.**

706 TE genes, TEs, and TE fragments in the Nd-1 genome sequence were identified based on  
707 sequence similarity to annotated TEs from the Col-0 gold standard sequence (Araport11)  
708 [40].

709

710 **AdditionalFile9. BUSCO analysis of the Col-0 and Nd-1 genome sequences.**

711 BUSCO v2.0 was run on the genomic sequences of Col-0 and Nd-1 using AUGUSTUS 3.3  
712 with default parameters for the gene prediction process.

713

714 **AdditionalFile10. Nd-1 plastome map.**

715 The GC content (black) and GC skew (green for positive GC skew, purple for negative GC  
716 skew) of the plastome sequence were analyzed by CGView [31]. The sequence and its  
717 properties are very similar to the Col-0 plastome sequence.

718

719 **AdditionalFile11. Nd-1 chondrome map.**

720 The GC content (black) and GC skew (green for positive GC skew, purple for negative GC  
721 skew) of the chondrome sequence were analyzed by CGView [31]. The sequence and its  
722 properties are very similar to the Col-0 chondrome sequence.

723

724 **AdditionalFile12. Experimental validation of 1 Mbp inversion on chromosome 4.**

725 The identified inversion between Nd-1 and Col-0 on chromosome 4 is different from the  
726 inversion described before between Col-0 and Ler [14]. However, the left breakpoint is the  
727 same for both alleles enabling the use of previously published oligonucleotide sequences  
728 [14]. The right breakpoint was identified by manual investigation of sequence alignments.  
729 Both breakpoints were validated via PCR using the oligonucleotides (for sequences see  
730 AdditionalFile6) as illustrated in (a). The results support the expected inversion borders (b).

731

732 **AdditionalFile13. Genome-wide distribution of genes inserted on chromosome 2 in Nd-  
733 1.**

734 AthNd-1\_v2c and the Col-0 gold standard sequence display a highly diverged region at the  
735 north of chromosome 2, which is about 300 kbp long. BLASTn of the complete Nd-1 gene  
736 sequences from this region revealed several regions on other Nd-1 chromosomes with  
737 copies of these genes.

738

739 **AdditionalFile14. Genome-wide distribution of large structural variants.**

740 The distribution of structural variants (SVs) >10 kbp (red dots) between Col-0 and Nd-1 over  
741 all five pseudochromosome sequences (black lines) is illustrated. Additionally, the assumed  
742 centromere (CEN) positions are indicated (blue dots). Most SVs are clustered in the (peri-  
743 )centromeric region.

744

745 **AdditionalFile15. Clustering of SVs around centromeres.**

746 The correlation between the number of SVs in a given part of the genome sequence (1 Mbp)  
747 and the distance of this region to the centromere position is illustrated. SVs are clustered  
748 around the centromeres (Spearman correlation coefficient = -0.66, p-value =  $1.7 \times 10^{-16}$ ).

749

#### 750 **AdditionalFile16. Functional annotation of GeneSet\_Nd-1\_v2.0.**

751 Functional annotations were transferred from Araport11 to corresponding RBHs in  
752 GeneSet\_Nd-1\_v2.0. In addition, genes were annotated based on the best BLAST hit  
753 if annotation via RBH was not possible.

754

#### 755 **AdditionalFile17. TE overlap with GeneSet\_Nd-1\_v2.0.**

756 The overlap between annotated TEs (AdditionalFile8) and predicted protein coding genes  
757 was analyzed to identify TE genes. This figure illustrates the fraction of a gene that is  
758 covered by a TE. Since TEs might occur within the intron of a gene, only genes with at least  
759 80% TE coverage were flagged as TE genes (AdditionalFile18).

760

#### 761 **AdditionalFile18. TE genes in GeneSet\_Nd-1\_v2.0.**

762 These genes were predicted by AUGUSTUS as protein coding genes. Due to their positional  
763 overlap with TEs (AdditionalFile8), they were flagged as TE genes and excluded from further  
764 gene set analysis.

765

#### 766 **AdditionalFile19. EST mapping.**

767 Percentage of nucleotides in ESTs matching predicted transcripts are displayed.

768

#### 769 **AdditionalFile20. Gene set overlap between Araport11, GeneSet\_Nd-1\_v1.1, and** 770 **GeneSet\_Nd-1\_v2.0.**

771 RBHs were identified pairwise between gene sets. The overlap was identified by mapping all  
772 genes onto Araport11 identifiers. Venn diagram construction was performed at  
773 <http://bioinformatics.psb.ugent.be/webtools/Venn/>.

774

775 **AdditionalFile21. Reciprocal best hits (RBH) pairs between Col-0 and Nd-1.**

776 Reciprocal best hits between predicted peptide sequences of Nd-1 and the representative  
777 peptide sequences of Col-0 (Araport11).

778

779 **AdditionalFile22. Reciprocal best hits (RHB) indicates inversion between Nd-1 and**  
780 **Col-0.**

781 Genes in RBH pairs were sorted based on their position on the five pseudochromosomes of  
782 the two genome sequences to form the x (Col-0) and y (Nd-1) axes of this diagram. Plotting  
783 the positions of each RBH pair leads to a bisecting line of black dots representing genes at  
784 perfectly syntenic positions. Red and green dots indicate RBH gene pair positions deviating  
785 from the syntenic position. Red dots symbolize a unique match to another gene, while green  
786 dots indicate multiple very similar matches. Positions of the centromere (CEN4) on the  
787 chromosomes of both accessions are indicated by purple lines. An inversion involving 131  
788 genes in RBH pairs just north of CEN4 distinguishes Nd-1 and Col-0.

789

790 **AdditionalFile23. RBH outliers in GeneSet\_Nd-1\_v2.0.**

791 RBHs between Araport11 and GeneSet\_Nd-1\_v2.0 were identified based on encoded  
792 representative peptide sequences. All 242 RBHs at positions deviating from the syntenic  
793 diagonal line were collected. The functional annotation of these genes was derived from  
794 Araport11.

795

796 **AdditionalFile24. Duplicated genes in Nd-1.**

797 The listed 385 Col-0 genes (Araport11 [40]) have at least two copies in Nd-1. Exons of these  
798 genes showed an increased copy number in Ath-Nd-1\_v2c compared to the Col-0 gold  
799 standard sequence. The annotation was derived from Araport11.

800

801 **AdditionalFile25. Duplicated genes in Col-0.**

802 The listed 394 Nd-1 genes have at least two copies in Col-0. Exons of these genes showed  
803 an increased copy number in the Col-0 gold standard sequence compared to Ath-Nd-1\_v2c.

804

805 **AdditionalFile26. Classification of genes as core, dispensable, or TE.**

806 GeneSet\_Nd-1\_v2.0 genes were classified as core, dispensable, or TE genes based on  
807 coverage in a read mapping involving 1,137 Illumina read data sets.

808

809 **AdditionalFile27. Genome-wide distribution of core, dispensable, and TE genes.**

810 Visualisation of the position of core genes, dispensable genes and TEs along the  
811 chromosomes.

812

813 **AdditionalFile28. Critical regions in the Col-0 gold standard sequence.**

814 The high contiguity of the Ath-Nd-1\_v2c assembly enabled the investigation of 13 sequences  
815 corresponding to gaps in the Col-0 gold standard sequence. This figure illustrates the  
816 homotetranucleotide occurrence in these sequences (red dots) in comparison to some  
817 randomly selected reference sequences (green dots). While there is a clear enrichment of  
818 homotetranucleotides in the gap-homolog sequences, there was no clear correlation  
819 between the length of a gap and the composition of the corresponding sequence observed.

820

821 **References**

822

- 823 1. Koornneef M, Meinke D. The development of Arabidopsis as a model plant. *The Plant*  
824 *Journal*. 2010;61(6):909-21.
- 825 2. The Arabidopsis Genome Initiative. Analysis of the genome sequence of the flowering  
826 plant *Arabidopsis thaliana*. *Nature*. 2000;408(6814):796-815.
- 827 3. Kumezawa N, Hosouchi T, Tsuruoka H, Kotani H. The size and sequence  
828 organization of the centromeric region of Arabidopsis thaliana chromosome 5. *DNA*  
829 *Research*. 2000;7(6):315-21.
- 830 4. Long Q, Rabanal FA, Meng D, Huber CD, Farlow A, Platzer A, et al. Massive  
831 genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. *Nature*  
832 *Genetics*. 2013;45(8):884-90. Epub 2013 Jun 23.
- 833 5. Vukašinić N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, et al. Dissecting a  
834 hidden gene duplication: the Arabidopsis thaliana SEC10 locus. *PLoS ONE*.  
835 2014;9(4):e94077.
- 836 6. Kawakatsu T, Huang SS, Jupe F, Sasaki E, Schmitz RJ, Urich MA, et al. Epigenomic  
837 Diversity in a Global Collection of Arabidopsis thaliana Accessions. *Cell*. 2016;166(2):492-  
838 505.
- 839 7. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, et al. Reference-  
840 guided assembly of four diverse Arabidopsis thaliana genomes. *Proceedings of the National*  
841 *Academy of Sciences of the United States of America*. 2011;108(25):10249-54. Epub 2011  
842 Jun 6.
- 843 8. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, et al. De novo assembly of soybean  
844 wild relatives for pan-genome analysis of diversity and agronomic traits. *Nature*  
845 *Biotechnology*. 2014;32(10):1045-54. Epub 2014 Sep 14.
- 846 9. Consortium TG. 1,135 Genomes Reveal the Global Pattern of Polymorphism in  
847 Arabidopsis thaliana. *Cell*. 2016;166(2):481-91. Epub 2016 Jun 9.
- 848 10. Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, et al. Long-read,  
849 whole-genome shotgun sequence data for five model organisms. *Scientific Data*.  
850 2014;1:140045.
- 851 11. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM. Assembling large  
852 genomes with single-molecule sequencing and locality-sensitive hashing. *Nature*  
853 *Biotechnology*. 2015;33(6):623-30. Epub 2015 May 25.
- 854 12. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, et al.  
855 Phased diploid genome assembly with single-molecule real-time sequencing. *Nature*  
856 *Methods*. 2016;13(12):1050-4. Epub 2016 Oct 17.
- 857 13. Pucker B, Holtgräwe D, Rosleff Sørensen T, Stracke R, Viehöver P, Weisshaar B. A  
858 De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1  
859 Displays Presence/Absence Variation and Strong Synteny. *PLoS ONE*.  
860 2016;11(10):e0164321.
- 861 14. Zapata L, Ding J, Willing EM, Hartwig B, Bezdán D, Jiao WB, et al. Chromosome-  
862 level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion  
863 polymorphisms. *Proceedings of the National Academy of Sciences of the United States of*  
864 *America*. 2016.
- 865 15. Michael TP, Jupe F, Bemm F, Motley ST, Sandoval JP, Lanz C, et al. High contiguity  
866 Arabidopsis thaliana genome assembly with a single nanopore flow cell. *Nature*  
867 *Communications*. 2018;9(1):541.
- 868 16. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, et  
869 al. Discovery and genotyping of structural variation from long-read haploid genome sequence  
870 data. *Genome Research*. 2017;27(5):677-85. Epub 2016 Nov 28.
- 871 17. Simpson JT, Pop M. The Theory and Practice of Genome Sequence Assembly.  
872 *Annual Review of Genomics and Human Genetics*. 2015;16:153-72. Epub 2015 Apr 22.
- 873 18. Rhoads A, Au KF. PacBio Sequencing and Its Applications. *Genomics Proteomics*  
874 *Bioinformatics*. 2015;13(5):278-89. Epub 2015 Nov 2.

- 875 19. Lam KK, Khalak A, D. T. Near-optimal assembly for shotgun sequencing with noisy  
876 reads. *BMC Bioinformatics*. 2014;15:S4. Epub 2014 Sep 10. doi: 10.1186/1471-2105-15-S9-  
877 S4.
- 878 20. Koren S, Phillippy AM. One chromosome, one contig: complete microbial genomes  
879 from long-read sequencing and assembly. *Current Opinion in Microbiology*. 2015;23:110-20.  
880 Epub 2014 Dec 1.
- 881 21. Shoromony I, Courtade T, Tse D. Do Read Errors Matter for Genome Assembly?  
882 *IEEE International Symposium on Information Theory (ISIT)*; Hong Kong 2015. p. 919-23.
- 883 22. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, et al. Reducing  
884 assembly complexity of microbial genomes with single-molecule sequencing. *Genome*  
885 *Biology*. 2013;14(9):R101.
- 886 23. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, et al. Assembly  
887 and diploid architecture of an individual human genome via single-molecule technologies.  
888 *Nature Methods*. 2015;12(8):780-6. Epub 2015 Jun 29.
- 889 24. Payne A, Holmes N, Rakyan V, Loose M. BulkVis: a graphical viewer for Oxford  
890 nanopore bulk FAST5 files. *Bioinformatics*. 2018.
- 891 25. Istace B, Friedrich A, d'Agata L, Faye S, Payen E, Beluche O, et al. de novo  
892 assembly and population genomic survey of natural yeast isolates with the Oxford Nanopore  
893 MinION sequencer. *Gigascience*. 2017;6(2):1-13.
- 894 26. Healey A, Furtado A, Cooper T, Henry RJ. Protocol: a simple method for extracting  
895 next-generation sequencing quality genomic DNA from recalcitrant plant species. *Plant*  
896 *Methods*. 2014;10(21). doi: 10.1186/1746-4811-10-21.
- 897 27. Stadermann KB, Holtgräwe D, Weisshaar B. Chloroplast Genome Sequence of  
898 *Arabidopsis thaliana* Accession Landsberg erecta, Assembled from Single-Molecule, Real-  
899 Time Sequencing Data. *Genome Announcements*. 2016;4(5):e00975-16.
- 900 28. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable  
901 and accurate long-read assembly via adaptive k-mer weighting and repeat separation.  
902 *Genome Research*. 2017;27(5):722-36. Epub 2017 Mar 15.
- 903 29. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W. Scaffolding pre-assembled  
904 contigs using SSPACE. *Bioinformatics*. 2011;27(4):578-9. Epub 2010 Dec 12.
- 905 30. Kleinboelting N, Huet G, Appelhagen I, Viehoveer P, Li Y, Weisshaar B. The  
906 Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-  
907 Strand Break Repair-Based Insertion Mechanism. *Molecular Plant*. 2015;8(11):1651-64.  
908 Epub 2015 Sep 5.
- 909 31. Stothard P, Wishart DS. Circular genome visualization and exploration using CGView.  
910 *Bioinformatics*. 2005;21(4):537-9. Epub 2004 Oct 12.
- 911 32. Kolmogorov M, Yuan J, Lin YR, Pevzner PA. Assembly of Long Error-Prone Reads  
912 Using Repeat Graphs. 2018.
- 913 33. Li H. Minimap and miniasm: fast mapping and de novo assembly for noisy long  
914 sequences. *Bioinformatics*. 2016;32(14):2103-10. Epub 2016 Mar 19.
- 915 34. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-  
916 MEM. Oxford University Press. 2013:1-3.
- 917 35. Pucker B, Brockington SF. Genome-wide analyses supported by RNA-Seq reveal  
918 non-canonical splice sites in plant genomes. *BMC Genomics*. 2018;19(1):980.
- 919 36. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an  
920 integrated tool for comprehensive microbial variant detection and genome assembly  
921 improvement. *PLoS ONE*. 2014;9(11):e112963.
- 922 37. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA. Primer3Plus,  
923 an enhanced web interface to Primer3. *Nucleic acids research*. 2007;35:W71-4. Epub 2007  
924 May 7.
- 925 38. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B. An *Arabidopsis*  
926 *thaliana* T-DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse  
927 genetics. *Plant Molecular Biology*. 2003;53(1):247-59.
- 928 39. Pucker B, Holtgräwe D, Weisshaar B. Consideration of non-canonical splice sites  
929 improves gene prediction on the *Arabidopsis thaliana* Niederzenz-1 genome sequence. *BMC*  
930 *Research Notes*. 2017;10(1):667.



- 931 40. Cheng CY, Krishnakumar V, Chan A, Thibaud-Nissen F, Schobel S, Town CD.  
932 Araport11: a complete reannotation of the *Arabidopsis thaliana* reference genome. *The Plant*  
933 *Journal*. 2017;(89):789-804.
- 934 41. Arend D, Junker A, Scholz U, Schüler D, Wylie J, Lange M. PGP repository: a plant  
935 phenomics and genomics data publication infrastructure. Database. 2016.
- 936 42. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM. BUSCO:  
937 assessing genome assembly and annotation completeness with single-copy orthologs.  
938 *Bioinformatics*. 2015;31(19):3210-2. Epub 2015 Jun 9.
- 939 43. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile  
940 and open software for comparing large genomes. *Genome Biology*. 2004;5(2):R12. Epub  
941 2004 Jan 30.
- 942 44. Khelik K, Lagesen K, Sandve GK, Rognes T, Nederbragt AJ. NucDiff: in-depth  
943 characterization and annotation of differences between two sets of DNA sequences. *BMC*  
944 *Bioinformatics*. 2017;18(1):338.
- 945 45. Keller O, Kollmar M, Stanke M, Waack S. A novel hybrid gene prediction method  
946 employing protein multiple sequence alignments. *Bioinformatics*. 2011;27(6):757-63. Epub  
947 2011 Jan 6.
- 948 46. Schmid KJ, Sorensen TR, Stracke R, Torjek O, Altmann T, Mitchell-Olds T, et al.  
949 Large-scale identification and analysis of genome-wide single-nucleotide polymorphisms for  
950 mapping in *Arabidopsis thaliana*. *Genome Research*. 2003;13(6A):1250-7. Epub 2003/06/12.  
951 doi: 10.1101/gr.728603. PubMed PMID: 12799357.
- 952 47. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search  
953 tool. *Journal of molecular biology*. 1990;215(3):403-10.
- 954 48. Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7:  
955 improvements in performance and usability. *Molecular Biology and Evolution*.  
956 2013;30(4):772-80. Epub 2013 Jan 16.
- 957 49. Smit AFA, Hubley R, Green P. RepeatMasker Open-4.0 2013-2015. Available from:  
958 <http://www.repeatmasker.org>.
- 959 50. Paten B, Earl D, Nguyen N, Diekhans M, Zerbino D, Haussler D. Cactus: Algorithms  
960 for genome multiple sequence alignment. *Genome Research*. 2011;21(9):1512-28. Epub  
961 2011 Jun 10.
- 962 51. Fiddes IT, Armstrong J, Diekhans M, Nachtweide S, Kronenberg ZN, Underwood JG,  
963 et al. Comparative Annotation Toolkit (CAT)-simultaneous clade and personal genome  
964 annotation. *Genome Research*. 2018;28(7):1029-38. Epub 2018 Jun 8.
- 965 52. Nawrocki EP, Eddy SR. Infernal 1.1: 100-fold faster RNA homology searches.  
966 *Bioinformatics*. 2013;29(22):2933-5. Epub 2013 Sep 4.
- 967 53. Kalvari I, Argasinska J, Quinones-Olvera N, Nawrocki EP, Rivas E, Eddy SR, et al.  
968 Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. *Nucleic acids*  
969 *research*. 2018;46(D1):D335-D42.
- 970 54. Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA  
971 genes in genomic sequence. *Nucleic acids research*. 1997;25(5):955-64.
- 972 55. Silverstein KA, Graham MA, Paape TD, VandenBosch KA. Genome organization of  
973 more than 300 defensin-like genes in *Arabidopsis*. *Plant Physiology*. 2005;138(2):600-10.
- 974 56. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, et al. An ethnically relevant  
975 consensus Korean reference genome is a step towards personal reference genomes. *Nature*  
976 *Communications*. 2016;7:13637.
- 977 57. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, et al. De novo assembly and  
978 phasing of a Korean human genome. *Nature*. 2016;538(7624):243-7. Epub 2016 Oct 5.
- 979 58. Chandrasekhara C, Mohannath G, Blevins T, Pontvianne F, Pikaard CS.  
980 Chromosome-specific NOR inactivation explains selective rRNA gene silencing and dosage  
981 control in *Arabidopsis*. *Genes & Development*. 2016;30(2):177-90. Epub 2016 Jan 7.
- 982 59. Copenhaver GP, Pikaard CS. RFLP and physical mapping with an rDNA-specific  
983 endonuclease reveals that nucleolus organizer regions of *Arabidopsis thaliana* adjoin the  
984 telomeres on chromosomes 2 and 4. *The Plant Journal*. 1996;9(2):259-72.

- 985 60. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, et al. The *Arabidopsis*  
986 *lyrata* genome sequence and the basis of rapid genome size change. *Nature Genetics*.  
987 2011;43(5):476-81. Epub 2011 Apr 10.
- 988 61. Villasante A, Abad JP, Méndez-Lago M. Centromeres were derived from telomeres  
989 during the evolution of the eukaryotic chromosome. *Proceedings of the National Academy of*  
990 *Sciences of the United States of America*. 2007;104(25):10542-7. Epub 2007 Jun 8.
- 991 62. Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, et al. Complex mtDNA  
992 constitutes an approximate 620-kb insertion on *Arabidopsis thaliana* chromosome 2:  
993 implication of potential sequencing errors caused by large-unit repeats. *Proceedings of the*  
994 *National Academy of Sciences of the United States of America*. 2001;98(9):5099-103. Epub  
995 2001 Apr 17.
- 996 63. Kowalski SP, Lan TH, Feldmann KA, Paterson AH. Comparative mapping of  
997 *Arabidopsis thaliana* and *Brassica oleracea* chromosomes reveals islands of conserved  
998 organization. *Genetics*. 1994;138(2):499-510.
- 999 64. Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, et al.  
1000 Integrated cytogenetic map of chromosome arm 4S of *A. thaliana*: structural organization of  
1001 heterochromatic knob and centromere region. *Cell*. 2000;100(3):367-76.
- 1002 65. Unseld M, Marienfeld JR, Brandt P, Brennicke A. The mitochondrial genome of  
1003 *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. *Nature Genetics*.  
1004 1997;15(1):57-61.
- 1005 66. Martínez-Zapater JM, Gil P, Capel J, Somerville CR. Mutations at the *Arabidopsis*  
1006 CHM locus promote rearrangements of the mitochondrial genome. *The Plant Cell*.  
1007 1992;4(8):889-99.
- 1008 67. Christensen AC. Plant mitochondrial genome evolution can be explained by DNA  
1009 repair mechanisms. *Genome Biology and Evolution*. 2013;5(6):1079-86.
- 1010 68. Preuten T, Cincu E, Fuchs J, Zoschke R, Liere K, Börner T. Fewer genes than  
1011 organelles: extremely low and variable gene copy numbers in mitochondria of somatic plant  
1012 cells. *The Plant Journal*. 2010;64(6):948-59. Epub 2010 Nov 4.
- 1013 69. Woloszynska M, Gola EM, Piechota J. Changes in accumulation of heteroplasmic  
1014 mitochondrial DNA and frequency of recombination via short repeats during plant lifetime in  
1015 *Phaseolus vulgaris*. *Acta Biochimica Polonica*. 2012;59(4):703-9. Epub 2012 Dec 6.
- 1016 70. Wendel JF, Jackson SA, Meyers BC, Wing RA. Evolution of plant genome  
1017 architecture. *Genome Biology*. 2016;17(37):s13059-016-0908-1.
- 1018 71. Denton JF, Lugo-Martinez J, Tucker AE, Schridder DR, Warren WC, Hahn MW.  
1019 Extensive error in the number of genes inferred from draft genome assemblies. *PLoS*  
1020 *Computational Biology*. 2014;10(12):e1003998.
- 1021 72. Belser C, Istace B, Denis E, Dubarry M, Baurens FC, Falentin C, et al. Chromosome-  
1022 scale assemblies of plant genomes using nanopore long reads and optical maps. *Natural*  
1023 *Plants*. 2018;4(11):879-87. Epub 2018 Nov 2.
- 1024 73. Paquette SM, Bak S, Feyereisen R. Intron-exon organization and phylogeny in a  
1025 large superfamily, the paralogous cytochrome P450 genes of *Arabidopsis thaliana*. *DNA and*  
1026 *Cell Biology*. 2000;19(5):307-17.
- 1027 74. Panchy N, Lehti-Shiu M, Shiu SH. Evolution of Gene Duplication in Plants. *Plant*  
1028 *Physiology*. 2016;171(4):2294-316. Epub 2016 Jun 10.
- 1029 75. Tan S, Zhong Y, Hou H, Yang S, Tian D. Variation of presence/absence genes  
1030 among *Arabidopsis* populations. *BMC Evolutionary Biology*. 2012;12(86):1471-2148/12/86.
- 1031 76. Benovoy D, Drouin G. Processed pseudogenes, processed genes, and spontaneous  
1032 mutations in the *Arabidopsis* genome. *Journal of Molecular Evolution*. 2006;62(5):511-22.  
1033 Epub 2006 Apr 11.
- 1034 77. Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. Evolutionary  
1035 and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiology*.  
1036 2009;151(1):3-15. Epub 2009 Jul 29.
- 1037 78. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, et al. Empirical analysis of  
1038 transcriptional activity in the *Arabidopsis* genome. *Science*. 2003;302(5646):842-6.
- 1039 79. Siena LA, Ortiz JP, Calderini O, Paolocci F, Cáceres ME, Kaushal P, et al. An  
1040 apomixis-linked ORC3-like pseudogene is associated with silencing of its functional homolog

- 1041 in apomictic *Paspalum simplex*. *Journal of Experimental Botany*. 2016;67(6):1965-78. Epub  
1042 2016 Feb 2.
- 1043 80. Zmienko A, Samelak-Czajka A, Kozlowski P, Szymanska M, Figlerowicz M.  
1044 *Arabidopsis thaliana* population analysis reveals high plasticity of the genomic region  
1045 spanning MSH2, AT3G18530 and AT3G18535 genes and provides evidence for NAHR-  
1046 driven recurrent CNV events occurring in this location. *BMC Genetics*. 2016;17(1):893.
- 1047 81. Yang L, Takuno S, Waters ER, Gaut BS. Lowly expressed genes in *Arabidopsis*  
1048 *thaliana* bear the signature of possible pseudogenization by promoter degradation. *Molecular*  
1049 *Biology and Evolution*. 2011;28(3):1193-203. Epub 2010 Nov 8.
- 1050 82. Marroni F, Pinosio S, Morgante M. Structural variation and genome complexity: is  
1051 dispensable really dispensable? *Current Opinion in Plant Biology*. 2014;18:31-6. Epub 2014  
1052 Feb 16.
- 1053 83. Golicz AA, Batley J, Edwards D. Towards plant pangenomics. *Plant Biotechnology*  
1054 *Journal*. 2016;14(4):1099-105. Epub 2015 Nov 23.
- 1055 84. Arumuganathan K, Earle ED. Nuclear DNA Content of Some Important Plant Species.  
1056 *Plant Molecular Biology Reporter*. 1991;9(3):208-18.
- 1057 85. Höfte H, Desprez T, Amselm L, Chiapello H, Caboche M, Moisan A, et al. An  
1058 inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from  
1059 *Arabidopsis thaliana*. *The Plant Journal*. 1993;4(6):1051-61.
- 1060
- 1061

1062 **Figure Legends**

1063

1064 **Figure 1: Nd-1 genome structure.**

1065 Schematic pseudochromosomes are represented by black lines with positions of genomics  
1066 features highlighted with colored icons as indicated in the insert.

1067

1068

1069 **Figure 2: Inversion on chromosome 4.**

1070 The dotplot heatmaps show the similarity between small fragments of two sequences. Each  
1071 dot indicates a match of 1 kbp between both sequences, while the color is indicating the  
1072 similarity of the matching sequences. A red line highlights an inversion between Nd-1 and  
1073 Col-0 or *Ler* and Col-0, respectively. A red arrow points at the position where the inversion  
1074 alleles differ between Nd-1 and *Ler*. (a) Comparison of the Nd-1 genome sequence against  
1075 the Col-0 gold standard sequence reveals a 1 Mbp inversion. (b) The *Ler* genome sequence  
1076 displays another inversion allele [14].

1077

1078

1079 **Figure 3: Highly divergent region on chromosome 2.**

1080 There is a very low similarity region (light blue) between the sequences in region A and  
1081 almost no similarity between the sequences in region B (white). The complete region  
1082 between 3.29 Mbp and 3.48 Mbp on NdChr2 is missing in the *Ler* genome.

1083

1084

1085 **Figure 4: Hidden locus in the Col-0 reference sequence.**

1086 Differences between the Nd-1 and Col-0 genome sequences lead to the discovery of a  
1087 collapsed region in the Col-0 gold standard sequence. There are two copies of At4g22214  
1088 (blue) present in the Col-0 genome, while only one copy is represented in the Col-0 gold  
1089 standard sequence. This gene duplication was initially validated through PCR with outwards

1090 facing oligonucleotides N258 and N259 (purple) which lead to the formation of the expected  
1091 PCR product (black). Parts of this region were cloned into plasmids (grey) for sequencing.  
1092 Sanger and paired-end Illumina sequencing reads revealed one complete gene  
1093 (At4g22214b) and a degenerated copy (At4g22214a). Moreover, the region downstream of  
1094 the complete gene copy in Nd-1 indicates the presence of at least one additional  
1095 degenerated copy.  
1096  
1097

1098 **Table 1: Nd-1 *de novo* assembly statistics.**

1099 Metrics of assemblies of the Nd-1 nucleome sequence generated by Canu, FALCON,  
 1100 miniasm, and Flye. All described assemblies are the final version after polishing.

<b>parameter</b>	<b>Ath-Nd1_v2c</b>	<b>Ath-Nd-1_v2f</b>	<b>Ath-Nd-1_v2m</b>	<b>Ath-Nd-1_v2y</b>
Assembler	Canu	FALCON	Miniasm	Flye
number of contigs	69	26	72	44
total number of bases	123,513,866	119,540,544	120,159,079	116,964,092
average contig length	1,790,056 bp	4,597,713 bp	1,668,876 bp	2,658,274 bp
minimal contig length	50,345 bp	86,055 bp	50,142 bp	53,207 bp
maximal contig length	15,898,009 bp	15,877,978 bp	14,338,505 bp	14,857,908 bp
GC content	36.14 %	36.04%	36.07%	36.01%
N25	14,369,729 bp	14,534,675 bp	11,880,610 bp	12,510,540 bp
N50	13,422,481 bp	9,302,209 bp	8,595,164 bp	10,607,548 bp
N75	8,555,326 bp	6,666,836 bp	3,513,050 bp	6,001,858 bp
N90	2,928,047 bp	2,829,734 bp	1,430,525 bp	2,524,876 bp

1101

1102









