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

Boas Pucker[1], Daniela Holtgräwe[1], Kai Bernd Stadermann[1], Katharina Frey[1], Bruno Huettel[2], Richard Reinhardt[2] and Bernd Weisshaar[1*]

[1] Bielefeld University, Faculty of Biology & Center for Biotechnology, Bielefeld, Germany

[2] Max Planck Genome Centre Cologne, Max Planck Institute for Plant Breeding Research, Cologne, Germany

[*] Corresponding author (BW)

Email addresses:
BP: bpucker@cebitec.uni-bielefeld.de
DH: dholtgra@cebitec.uni-bielefeld.de
KS: kstaderm@cebitec.uni-bielefeld.de
KF: katharina.frey@uni-bielefeld.de.
BH: huettel@mpipz.mpg.de
RR: reinhardt@mpipz.mpg.de
BW: bernd.weisshaar@uni-bielefeld.de

ORCIDs
BP: https://orcid.org/0000-0002-3321-7471
DH: https://orcid.org/0000-0002-1062-4576
KS: https://orcid.org/0000-0002-8036-1492
BW: https://orcid.org/0000-0002-7635-3473

Abstract

Background

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

Results

Here we provide a chromosome-arm level assembly of the *A. thaliana* accession Niederzenz-1 (AthNd-1_v2) based on SMRT sequencing data. The assembly comprises 26 nucleome sequences and displays a contig length of up to 16 Mbp. Compared to an earlier Illumina short read-based NGS assembly (AthNd-1_v1), a 200 fold increase in continuity was observed for AthNd-1_v2. To assign contig locations independent from the Col-0 reference sequence, we used genetic anchoring to generate a truly *de novo* assembly. In addition, we assembled the chondrome and plastome sequences.

Conclusions

Detailed analyses of AthNd-1_v2 allowed reliable identification of large genomic rearrangements between *A. thaliana* accessions contributing to differences in the gene sets that distinguish the genotypes. One of the differences detected identified a gene that is lacking from the Col-0 reference sequence. This *de novo* assembly will extent the known proportion of the *A. thaliana* pan-genome.

Background

Introduction

Arabidopsis thaliana became the most important model for plant biology within decades due to properties valuable for basic research like short generation time, small footprint or a small genome [1]. Even before the availability of DNA sequencing technologies the *A. thaliana* genome was studied by biochemical methods like reassociation kinetics [2], quantitative gel blot hybridization [3], Feulgen photometry, flow cytometry [4, 5], chromatin staining, fluorescence *in situ* hybridization and southern blotting [6]. Molecular biology studies indicated a genome size between 145 Mbp [4] and 160 Mbp [5] as well as a GC content of 40.3% [5]. Construction of genomic clones in vectors like phage lambda derivatives and genome blotting without knowing the actual sequence revealed insights into genome sequence complexity. Examples are the detection of about 570 copies of the 45S transcription unit (rDNA) and 660 chloroplast genome copies per cell [7]. By *in situ* hybridization Chromosome 1 and 5 were classified as metacentric, chromosomes 2 and 4 as acrocentric with nucleolus organizing regions (NORs) located at the short arms, and chromosome 3 was shown to be submetacentric [8]. Moreover, rDNA position polymorphisms between *A. thaliana* accessions were detected [8]. Different genetic maps were constructed, initially mainly based on restriction fragment length polymorphism (RFLP) and cleaved amplified polymorphic sequences (CAPS) markers [9, 10]. High resolution genetic maps were developed based on recombinant inbred lines (RILs) derived from crosses of Col-0 and Landsberg *erecta* (L*er*) [11]. The impact and position of genomic features like the recombination reduction by NORs on the short chromosome arms of chromosome 2 and chromosome 4 and the centromere positions were investigated by tetrad analysis [12]. Genetic maps provided the scaffold for the positioning and orienting of continuous DNA sequences or contigs [5] leading to chromosome-level physical maps and centromere size estimations [13]. Gene and genome duplication events were studied based on BAC sequences prior to completion of the reference genome [14]. Generated by a BAC-by-BAC approach, the almost 120 Mbp long Col-0 reference sequence is currently the most accurate plant genome sequence [15]. However, even this excellent high-quality nuclear genome sequence contains remaining gaps in almost inaccessible regions like repeats in the centromeres [13], at the telomeres and throughout NORs. The most recent genome annotation in Araport11 [16], which served as reference annotation for this study, contains 27,445 protein encoding nuclear genes as well as 31,189 transposable element sequences. Information about genomic differences between *A. thaliana* accessions were mostly derived from short read data [17, 18]. The average proportion sequenced per line was around 100 Mbp covering 84% of the Col-0 reference sequence [19]. However, selected accessions were sequenced much deeper leading to an almost reference-size assembly [17, 20, 21]. The identification of structural variants had an upper limit of 40 bp for most of the investigated accessions [19]. Larger insertions and deletions, which will often result in presence/absence variations of entire genes, are often missed in short read data sets [22]. Arabidopsis assembly continuity was significantly increased from high quality reference-guided assemblies [17] over *de novo* assemblies [20, 21] to most recent assemblies reaching chromosome-level quality [23].

The assembly concept of whole genome shotgun sequencing which relies on contigs created from overlapping sequence reads shorter than many repeat sequences and subsequent scaffolding is now challenged by new technical developments. The strong increase in the length of sequencing reads that was technically realized during the last years is enabling new assembly approaches [24, 25]. Despite the high error rate of about 11 to 15% 'Single Molecule, Real Time' (SMRT) sequencing reads significantly improve the continuity of *de novo* assemblies due to an efficient correction of the almost unbiased errors [26-28], provided that sufficient read coverage is available. SMRT sequencing offered by PacBio results routinely in average read lengths above 10 kbp [20, 29, 30]. These long reads were incorporated into high quality hybrid assemblies involving Illumina short read data [23, 30], but increasing sequencing output supports the potential for so called 'PacBio only assemblies' [20, 27, 31, 32].

Since the routine construction of very high quality assemblies becomes more feasible, methods for genome sequence comparison, especially for the comparison of multiple sequences in one alignment, need to be developed [33, 34]. Reciprocal best BLAST hits (RBHs) are a suitable way to analyze the synteny of two genomes by identifying homologous sequences [35, 36]. Each RBH pair consists of two sequences, one from each of the two genome sequences to compare, which displays the highest scoring hit in the other data set in a reciprocal manner [37]. These RBH pairs can be used to guide an assembly [21].

Here we provide a SMRT sequencing-based *de novo* genome assembly of Nd-1 comprising contigs of chromosome-arm size anchored to chromosomes and orientated within pseudochromosome sequences based on genetic linkage information. The application of long sequencing reads abolished limitations of short read mapping and short read assemblies for genome sequence comparison. Based on this genome sequence assembly, we identified genomic rearrangements between Col-0 and Nd-1 ranging from a few kbp up to one Mbp. Gene duplications between both accessions as well as 'private' genes in Nd-1 and Col-0 were revealed by this high quality sequence. The current assembly version outperforms the Illumina-based version (AthNd-1_v1) about 200 fold with respect to assembly continuity [21] and is in the same range as the recently released L*er* genome sequence assembly [23].

Methods

Plant material

Niederzenz-1 (Nd-1) seeds were obtained from the European Arabidopsis Stock Centre (NASC; stock number N22619). The DNA source was the same as described earlier [21].

DNA extraction

The DNA isolation procedure was a modified version of previously published protocols (AdditionalFile1) [32, 38] and started with 5 g of frozen leafs.

Library preparation and sequencing

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

Assembly parameters

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

Read sequences derived from the plastome [GenBank: AP000423.1] or chondrome [GenBank: Y08501.2] were extracted from the raw data set by mapping to the respective sequence of Col-0 as previously described [39]. Canu v1.4 [40] was used for the assembly of the organell genome sequences. Scaffolding of initial contigs was performed with SSPACE-LongRead v1.1 [41]. The quality of both assemblies was checked by mapping of NGS reads from Nd-1 [21] and Col-0 [42]. Manual inspection and polishing with Quiver [32] let to the final sequences. The start of the Nd-1 plastome and chondrome sequences was set according to the corresponding Col-0 plastome and chondrome sequences to ease comparisons. Finally, small assembly errors were corrected via CLC basic variant detection based on mapped

Illumina paired-end reads (SRX1683594, [21]) and PacBio reads. Sequence properties like GC content and GC skew were determined and visualized by CGView [43].

A total of 166,600 seed reads consisting of 4,500,092,354 nt (N50 = 26,295 nt) covering the expected 150 Mbp genome sequence were used for the assembly thus leading to a coverage of 30 fold (see AdditionalFile2 for details). Release version 1.7.5 of the FALCON assembler https://github.com/PacificBiosciences/FALCON/ [32] was used for a *de novo* assembly (see AdditionalFile3 for parameters) of the nuclear genome sequence. Resulting contigs were checked for contaminations with bacterial sequences and organell genome sequences as previously described [21]. Small fragments with low coverage were removed prior to polishing and error correction with Quiver [32].

Construction of pseudochromosomes based on genetic information

All assembled contigs were sorted and orientated based on genetic linkage information derived from 63 genetic markers (AdditionalFile4, AdditionalFile5, AdditionalFile6), which were analyzed in about 1,000 F2 plants, progeny of reciprocal crossing of Nd-1xCol-0 and Col-0xNd-1. Genetic markers belong to three different types: (1) fragment length polymorphisms, which can be distinguished by agarose gel electrophoresis, (2) small nucleotide polymorphisms which can be distinguished by Sanger sequencing and (3) small nucleotide polymorphisms, which were identified by high resolution melt analysis. Design of oligonucleotides was performed manually and using Primer3Plus [44]. DNA for genotyping experiments was extracted from *A. thaliana* leaf tissue using a cetyltrimethylammonium bromide (CTAB) based method [45]. PCRs were carried out using GoTaq G2 DNA Polymerase (Promega) generally based on the suppliers' protocol. The total reaction volume was reduced to 15 µl and only 0.2u of the polymerase were used per reaction. Sizes of amplicons generated were checked on an agarose gels. If required, samples were purified for sequencing by ExoSAP-IT (78201.1.ML ThermoFisher Scientific) treatment as previously described [46]. Sanger sequencing on ABI3730XL was applied to identify allele-specific

SNPs for the genotyping. Manual inspection of gel pictures and electropherograms lead to genotype calling. High resolution melt analysis was performed on a CFX96 Touch Real-Time PCR Detection System (BioRad) using the Precision Melt Supermix according to suppliers instructions (BioRad).

All data were combined and processed by customized Python scripts to calculate recombination frequencies between genetic markers. Linkage of genetic markers provided information about relationships of assembled sequences. The north-south orientation of the chromosomes was transferred from the reference sequence based on RBH support. Afterwards, contigs were joined into pseudochromosome sequences (AthNd-1_v2). The produced research data, that is the basis for this article, is available upon request.

Genome structure investigation

Characteristic elements of the Nd-1 genome sequence were annotated by mapping of known sequences as previously described [21]. Fragments and one complete 45S rDNA unit were discovered based on gi|16131:848-4222 and gi|16506:88-1891. AF198222.1 was subjected to a BLASTn for the identification of 5S rDNA sequences. Telomeric repeats were used to validate the assembly completeness at the pseudochromosome end as well as centromere positions as previously described [21].

BUSCO analysis

BUSCO [48] was run on the Nd-1 pseudochromosomes and on the Col-0 reference sequence to produce a gold standard for Arabidopsis. AUGUSTUS 3.2.1 [49] was applied with previously described parameters [21]. The ‘embryophyta_odb9’ was used as reference gene set.

Genome sequence alignment

Nd-1 pseudochromosome sequences were aligned to the Col-0 reference sequence [15] via nucmer [50] using parameters described in [23]. The aligned blocks were extracted via showcoords function. The longest path of allelic blocks was identified by custom python scripts. Blocks were classified as allelic, transposition or inversion according to the Col-0 reference sequence [15]. Classified blocks were merged with adjacent blocks of the same type.

Gene prediction and RBH analysis

AUGUSTUS 3.2.1 [49] was applied to the Nd-1 assembly AthNd-1_v2 with previously optimized parameters [46]. Afterwards, the identification of RBHs at the protein sequence level between Nd-1 and Col-0 (Araport11, representative peptide sequences) was carried out with a custom python script as previously described [21].

Additionally, gene prediction was run on the nucleome TAIR10 reference sequence [15] as well as on the L*er* chromosome sequences [23]. Parameters were set as described before to generate two control data sets.

Transposable element annotation

All annotated transposable element (TE) regions of Araport11 (derived from TAIR) [16] were mapped via BLASTn to the Nd-1 assembly AthNd-1_v2 and against the Col-0 reference sequence. The top BLAST score for each element in the mapping against the Col-0 reference sequence was identified. All hits against Nd-1 with at least 90% of this top score were considered for further analysis. Overlapping hits were removed to annotate a final TE set. All predicted Nd-1 genes which overlapped TEs with more than 80% of their gene space were flagged as putative TE genes.

Identification of gene copy number variations

A BLASTn search of all Col-0 exon sequences against the Nd-1 genome assembly sequence AthNd-1_v2 and of all predicted Nd-1 exon sequences against the Col-0 reference sequence was used to determine copy number variations of genes. Only non-overlapping hits were considered for the following analysis. Genes were considered to be duplicated if at least half of their exons were found more than once. At5g12370 [51] served as an internal control, because the duplication of this *A. thaliana* gene is collapsed in the Col-0 reference sequence but resolved in the Nd-1 genome sequence assembly. Duplication candidates were functionally annotated based on the Araport11 [16] information. Afterwards, putative transposable element genes were removed based on the annotation or the overlap with annotated transposable element sequences (AdditionalFile7), respectively. Duplications were classified as ‘tandem’ if the distance between both copies was smaller than 1 Mbp. Distances between genes and the next TEs were measured from the center of each feature to determine the impact of TEs on gene duplications. Finally, g:profiler http://biit.cs.ut.ee/gprofiler/ [52] was applied to identify significantly overrepresented genes in Col-0 and Nd-1.

Beside genes with changed copy numbers, protein coding genes unique to each accession were identified. Annotated genes in AthNd-1_v2, which were absent from the TAIR10 reference genome sequence, were considered as unique to Nd-1. To avoid assembly-related issues in the identification of unique Col-0 genes, we searched the peptide sequences of all potential unique Col-0 genes against the complete set of Nd-1 subreads.

Validation of rearrangements and duplications

LongAmpTaq (NEB) was used for the generation of large genomic amplicons up to 18 kbp based on the suppliers’ protocol. Sanger sequencing was applied for additional confirmation of generated amplicons. The amplification of small fragments and the following procedures were carried out with standard polymerases as previously described [21].

Investigation of collapsed region

The region around At4g22214 was amplified in five overlapping parts using the Q5 High Fidelity polymerase (NEB) with genomic DNA from Col-0. Amplicons were checked on agarose gels and finally cloned into pCR2.1 (Invitrogen) or pMiniT 2.0 (NEB), respectively, based on the suppliers’ recommendations. Cloned amplicons were sequenced on an ABI3730XL by primer walking. Sequencing reads were assembled using CLC

GenomicsWorkbench (v. 9.5 CLC bio). In addition, 2x250 nt paired-end Illumina reads of Col-0 [42] were mapped to correct small variants in the assembled contigs and to close a small gap between cloned amplicons.

Identification of structural variants

The distances between all syntenic neighboring RBHs were taken into account to identify structural variants above 10 kbp in length. Differences in the distance between two neighboring genes in the Col-0 genome and the corresponding neighboring genes in the Nd-1 genome indicate a structural variation between them. Spearman correlation coefficient was calculated using the implementation in the Python module scipy to validate the indication of increased numbers of SV around the centromeres.

Analysis of gaps in the Col-0 reference sequence

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

Results

Nd-1 genome

The final *A. thaliana* Nd-1 assembly (AthNd-1_v2) comprised 119.5 Mbp (Table 1). AthNd-1_v2 exceeds the previously reported assembly version AthNd-1_v1 by 2.5 Mbp, while reducing the number of contigs by a factor of about 200.

The plastome and chondrome sequences comprise 154,443 bp and 368,216 bp, respectively (available upon request). A total of 148 small variants were identified from a global alignment between the Nd-1 and Col-0 plastome sequences. General sequence properties like GC content and GC skew (AdditionalFile8, AdditionalFile9) are almost identical to the plastome and chondrome of Col-0. Nevertheless, there are some rearrangements between the chondrome sequences of Nd-1 and Col-0.

The high assembly quality and completeness of AthNd-1_v2 is supported by the detection of 99.9% of all BUSCO genes detected in Col-0 (AdditionalFile10). Only two genes are missing in the Nd-1 assembly AthNd-1_v2, which are partly present in the Col-0 reference sequence. These genes are EOG09360D4T (At3g01060) and EOG09360DFK (At5g01010) located at the very north end of chromosome 3 and chromosome 5, respectively. Both regions are not represented in AthNd-1_v2, but can be detected in the subreads. Amplification via PCR and Sanger sequencing of the PCR products confirmed the presence of both genes in the Nd-1 genome. NGS read mappings did not indicate any complications at the end of both sequences.

Pseudochromosomes were constructed truly *de novo* from 3-7 contigs based on genetic linkage information. They reach similar lengths as the corresponding chromosome sequences in the Col-0 reference sequence. The Nd-1 genome sequence AthNd-1_v2 contains a complete 45S rDNA unit on pseudochromosome 2 as well as several fragments of additional 45S rDNA units on pseudochromosomes 2, 4, and 5 (Fig. 1). Centromeric and telomeric repeat sequences as well as 5S rDNA sequences were detected at centromere positions. Completeness of the assembled sequences representing the north of chromosome 1 and the south of chromosome 3 were confirmed by the occurrence of telomeric repeat sequences (Fig. 1).

Genome structure differences

Sequence comparison between AthNd-1_v2 and the Col-0 reference sequence revealed a large inversion on chromosome 4 involving about 1 Mbp (Fig. 2). The left break point is at 1,631,539 bp and the right break point at 2,702,549 bp on NdChr4. The inverted sequence is 120,543 bp shorter than the corresponding Col-0 sequence. PCR amplification of both inversion borders (AdditionalFile11) and Sanger sequencing of the generated amplicons was used to validate this rearrangement.

The recombination frequency in this region was analyzed using the marker pair M84/M74. Only a single recombination was observed between these markers while investigating 60 plants. Moreover, only 8 recombination events in 108 plants were observed between another pair of markers, spanning a larger region (AdditonalFile5). In contrast, the average recombination frequency per Mbp at the corresponding position on other chromosomes was between 12%, observed for M31/M32, and 18%, observed for M13/M14. Statistical analysis revealed a significant difference in the recombination frequencies between the corresponding positions on different chromosomes (p<0.001, prop.test() in R) supporting the hypothesis of a reduced recombination rate across the inversion on chromosome 4.

Comparison of a region on Chr2, which is probaly of mitochondrial origin (mtDNA), in the Col-0 reference sequence with the Nd-1 genome sequence revealed a 300 kbp highly divergent region (Fig. 3). Sequences between position 3.20 Mbp and 3.29 Mbp on NdChr2 display low similarity to the Col-0 sequence, while there is almost no similarity between 3.29 Mbp and 3.48 Mbp. However, the length of both regions is roughly the same. Comparison against the L*er* genome assembly revealed the absence of the entire region between 3.29 Mbp and 3.48 Mbp on chromosome 2. The Nd-1 sequence from this region lacks continuous similarity to another place in the Col-0 or Nd-1 genome sequence. The 28 genes encoded in this region in Nd-1 show weak similarity to other Arabidopsis genes. Comparison of gene space sequences from this region against the entire Nd-1 assembly revealed some similarity on chromosome 3, 4, and 5 (AdditionalFile12).

An inversion on chromosome 3 which was described between Col-0 and L*er* [23] is not present in Nd-1. The sequence similarity between Col-0 and Nd-1 is high in this region. In total, 175 structural variants larger than 10 kbp were identified between Col-0 and Nd-1. The genome-wide distribution of these variants indicated a clustering around the centromeres (AdditionalFile13). A Spearman correlation coefficient of -0.66 (p=1.7*10^{-16}) was calculated for the correlation of the number of SVs in a given interval and the distance of this interval to the centromere (AdditionalFile14). Therefore, these large structural variants are significantly more frequent in the centromeric and pericentromic regions.

Hint-based gene prediction

Hint-based gene prediction using AUGUSTUS with the *A. thaliana* species parameter set on the Nd-1 pseudochromosomes resulted in 30,132 nuclear protein coding genes (GeneSet_Nd-1_v2.0) with an average transcript length of 1,573 bp (median), an average CDS length of 1,098 bp (median) and an average exon number per transcript of four (median). The number of predicted genes exceeds the number of annotated nuclear protein coding Col-0 genes in Araport11 (27,445) by 2,687. At the same time, the number of predicted genes is reduced compared to the GeneSet_Nd-1_v1.1 [46] by 702 genes.

As controls we run the gene prediction with same parameters on Col-0 and L*er* chromosome sequences resulting in 30,352 genes and 29,302 genes, respectively. There were only minor differences concerning the average transcript and CDS length as well as the number of exons per gene.

Based on 31,748 annotated TEs in Nd-1 (AdditionalFile7) 2,738 predicted Nd-1 genes were flagged as putative TE genes (AdditionalFile15, AdditionalFile16). This number matches well with the difference between the predicted genes in Nd-1 and the annotated protein coding genes in Araport11, which is supposed to be free of TE genes.

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

A BLASTp-based comparison of all predicted Nd-1 peptide sequences and Col-0 Araport11 representative peptide sequences in both directions revealed 24,572 reciprocal best hits (RBHs). In total, 89.6% of all 27,445 nuclear Col-0 genes are represented in this RBH set. Analysis of the colinearity of the genomic location of all 24,572 RBHs (see AdditionalFile17 for a list) between Nd-1 and Col-0 showed overall synteny of both genomes as well as an inversion on chromosome 4 (AdditionalFile18). While most RBHs are properly flanked by their syntenic homologs and thus lead to a diagonal positioning of points in the scatter plot, there are 242 outliers (see AdditionalFile19 for a list). Outliers were distinguished into 214 "random" outliers (green), which have multiple BLASTp hits of similar quality for genes at different locations in the genome sequence, and 28 "real" outliers (red), which display a unique BLASTp hit. In general, outliers occur frequently in regions around the centromeres. Positional analysis revealed an involvement of most "real" outliers in the large inversion on chromosome 4. An NGS read mapping at the positions of randomly selected "real" outliers was manually inspected and indicated rearrangements between Nd-1 and Col-0. Structural variants, which affect at least three different genes in a RBH pairs, were identified from the RBH analysis. Examples beside the previously mentioned 1.2 Mbp inversion on chromosome 4 (*At4g03820-At4g05497*) are a translocation on chromosome 3 (*At3g60975-At3g61035*) as well as an inversion on chromosome 3 around *At3g30845*.

As a control we identified 25,454 (92.7%) RBHs between our gene prediction on Col-0 and the manually curated reference annotation Araport11. In addition, 24,302 (88.5 %) RBHs were identified between our gene prediction on the L*er* assembly and the Col-0 reference sequence annotated in Araport11.

In total, 385 protein encoding genes (AdditionalFile20) were detected to be copied at least once in Nd-1 compared to the Col-0 reference sequence. This includes *SEC10* (*At5g12370*) [51] which was previously described as an example for a tandem gene duplication collapsed in the Col-0 reference sequence. However, this region was already properly represented in AthNd-1_v1 [21]. Gene duplications of *At2g06555* (unknown protein), *At3g05530* (*RPT5A*) and *At4g11510* (RALFL28) in Nd-1 were confirmed by PCR amplification and Sanger sequencing of the sequences enclosed by both copies as well as through amplification of the entire event locus. On the other side, there are 394 predicted genes in Nd-1 (AdditionalFile21) which appeared at least duplicated in Col-0. A functional annotation is missing for about half of the duplicated genes. ENSEMBL-based enrichment analysis revealed significantly overrepresented functionalities due to different copy number of genes in Col-0 and Nd-1 (AdditionalFile22).

In addition to gene duplications, there were 43 genes unique to Nd-1 (AdditionalFile23) and 42 genes unique to Col-0 (AdditionalFile24). Most of the gene functions were unknown and the functionally annotated genes were randomly distributed over different gene families and pathways. The length of the encoded peptides is shorter than the genome-wide average and some peptide sequences display long amino acid repeats. It has not escaped our notice that some of these genes might be gene prediction artifacts.

Hidden locus in Col-0

At4G22214 was identified as a gene duplicated in Nd-1 in our analysis. During experimental validation, we did not detect the expected difference between Col-0 and Nd-1 concerning the locus around *At4G22214*. However, the PCR results matched the expectation based on the Nd-1 genome sequence thus suggesting a collapsed gene sequence in the Col-0 reference sequence. This hypothesis was supported by PCR results with outwards facing primers (Fig. 4). Cloning of the At4g22214 region of Col-0 in five overlapping fragments was done to enable Sanger sequencing. The combination of Sanger and paired-end Illumina sequencing reads revealed a tandem duplication with modification of the original gene (Fig. 4). The copies were designated At4g22214a and At4g22214b based on their position in the genome (GenBank: MG720229). While At4g22214b almost perfectly matches the Araport11 annotation of At4g22214, a significant part of the CDS of At4g22214a is missing. Therefore, the gene product of this copy is probably functionless.

Gaps in the Col-0 reference sequence

Despite its very high quality, the Col-0 reference sequence contains 92 gaps of varies sizes representing regions of unknown sequence like the NOR clusters or centromeres. Ath-Nd1_v2 enabled the investigation of some of these sequences based on homology assumptions. A total of 22 Col-0 gaps were spanned with high confidence by Ath-Nd1_v2 and therefore selected for homopolymer frequency analysis. The corresponding regions in Nd-1 are significantly enriched with homopolymers in comparison to randomly picked control sequences (p=0.000022, Mann-Whitney U test) (AdditionalFile 25).

Discussion

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

In order to further investigate large variations in the range of several kbp up to several Mbp between *A. thaliana* accessions, we performed a *de novo* genome assembly for the Nd-1 accession using long sequencing reads and cutting-edge assembly software. Based on SMRT sequencing reads the assembly continuity was improved by over 200 fold considering the number of contigs in the previously released NGS-based assembly [21]. Assembly statistics are comparable to other projects using similar data [23, 31, 53, 54]. Despite the very high continuity, regions like NORs still pose a major challenge. These sequences are not just randomly clustered repeats, but highly regulated [55]. Therefore, the identification of accession-specific differences could explain phenotypic differences. One NOR repeat unit sequence in the Nd-1 assembly is located at 2.5 Mbp on chromosome 2. If this repeat unit indicates a NOR position, this would be a structural difference to Col-0 where the NOR2 is located at the very north end [56]. In addition to NORs, the assembly of chromosome ends remains still challenging, since the absence of some telomeric sequences in a high quality assembly was observed before [23]. Despite the absence of challenging repeats, regions close to the telomeres including the genes EOG09360D4T (At3g01060) and EOG09360DFK (At5g01010) were not assembled by FALCON although sequence reads covering these regions were present in the input data.

Genome sequence differences

The increased continuity of this long read assembly was necessary to discover an 1 Mbp inversion through sequence comparison as well as RBH analysis. An earlier Illumina short read based assembly [21] lacked sufficient continuity in the region of interest to reveal both breakpoints of this variant between Col-0 and Nd-1 in one contig. The large inversion at the north of chromosome 4 is a modification of the allele originally detected in L*er* [23, 57]. The Nd-1 allele is different from the L*er* allele. This could explain previous observations in several hundred *A. thaliana* accessions, which share the left inversion border with L*er*, but show a different right inversion border [23].

Despite the long read length, there are only very small parts of pericentromeric sequences represented in the assembly. Assuming an almost complete absence of centromere and NOR sequences from the assembly, the true genome size is matching earlier predictions of around 145-160 Mbp, which were calculated based on flow cytometry [4, 5] and adjusted towards the lower end of this range in more recent estimations [21]. Since genome size differences between accessions have been reported, the investigation of different accessions might explain some of the observed discrepancies [58]. Detection of telomeric or centromeric sequences, respectively, at the end of pseudochromosomes indicated the completeness of the Ath-Nd1_v2 assembly at these points. Almost 20 years after the release of the first chromosome sequences of *A. thaliana*, we are still not able to assemble complete centromere sequences continuously. However, absence of telomeric sequences from some pseudochromsome ends was observed before even for a very high quality assembly [23]. Detected telomeric repeats at the centromere positions support previously reported hypothesis about the evolution of centromers out of telomere sequences [59].

Sequence differences observed on chromosome 2 between Col-0 and Nd-1 could be due to the integration of mtDNA into the chromosome 2 of Col-0 [15]. This region was reported to be collapsed in the Col-0 reference genome sequence, thus harboring about 600 kbp of DNA from the chondrome instead of the 270 kbp represented in the reference genome sequence

[60]. Since Nd-1 genes of this region show similarity to gene clusters on other chromosomes, they could be relicts of a whole genome duplication as reported before for several regions of the Col-0 reference sequence [61]. This difference on chromosome 2 is only one example for a large variant between Col-0 and Nd-1. Clusters of structural variants around centromeres could be explained by transposable elements and pseudogenes which were previously reported as causes for intra-species variants in these regions [6, 60].

Size and structure of the Nd-1 plastome is very similar to Col-0 [15] or L*er* [39]. In accordance with the overall genome similarities, the observed number of small differences between the plastome sequences of Col-0 and Nd-1 is slightly higher than the value reported before for the Col-0 comparison to L*er* [39].

The size of the Nd-1 chondrome matches previously reported values for the large chondrome configuration of other *A. thaliana* accessions [62]. Large structural differences between the Col-0 chondrome [62] and the Nd-1 chondrome could be due to the previously described high diversity of this subgenome including the generation of substoichiometric DNA molecules [63, 64]. In addition, the mtDNA level was reported to differ between cell types or cells of different ages within the same plant [65, 66]. The almost equal read coverage of the assembled Nd-1 chondrome could be explained by the young age of the plants at the point of DNA isolation, as the amount of all chondrome parts should be the same in young leafs [66].

Nd-1 gene space

Many diploid plant genomes contain close to 30,000 protein encoding genes [67] with the Arabidopsis genome harboring 27,655 genes according to the most recent annotation [16]. Since there are only two other chromosome-level assembly sequences of *A. thaliana* available at the moment, we do not know the precise variation range of gene numbers between different accessions. The number of 30,132 predicted genes in Nd-1 is further supported by the identification of 24,572 RBHs with the Araport11 [16] annotation of the Col-0 reference sequence. This number exceeds the values reported for Nd-1 before [21, 46] as well as the matches between Col-0 and Ler-0 [23]. Incorporation of hints improved the gene prediction on the NGS assembly sequence AthNd-1_v1.0 [46] and was therefore applied again. Our chromosome-level assembly further enhances the gene prediction quality as at least 89.6% of all Col-0 genes were recovered. Previous studies reported annotation improvements through an improved assembly sequence [68].

Due to the very high proportion of genes within the Arabidopsis genome assigned to paralogous groups with high sequence similarity [69, 70], we speculated that the identification of orthologous pairs via RBH analysis might be almost saturated. Gene prediction with the same parameters on the Col-0 reference sequences prior to a RBH analysis supported this hypothesis. Since there are even some RBHs at non-syntenic positions between our control Col-0 annotation and the Araport11 annotation, our Nd-1 annotation is already of very high accuracy. The precise annotation of non-canonical splice sites via hints as described before [46] contributed to the new GeneSet_Nd-1_v2.0. Slightly over 200 genes at non-syntenic positions designated as 'outliers' in our RBH analysis highlight structural differences in the local genome structure.

Gene duplication and deletion numbers in Nd-1 and Col-0 are in the same range as previously reported values of up to a few hundred accession specific presence/absence variations of genes [23, 71]. Since we were searching genome wide for copies of a gene space without requiring an annotated feature in both genome sequences, both numbers might include some pseudo genes due to the frequent occurrence of these elements within plant genomes [72, 73]. Since all comparisons rely on the constructed sequences we cannot absolutely exclude that a small number of other genes were detected as amplified due to a collapsed sequences like SEC10 (At5G12370) [51]. Removing transposable element genes based on sequence similarity to annotated features should reduce the proportion of putative pseudo genes. However, it is impossible to clearly distinguish between real genes and pseudo genes in all cases, because even genes with a premature stop codon or a frameshift mutation could function as a truncated versions or give rise to regulatory RNAs [70, 73-75]. In addition, the impact of copy number variations involving protein encoding genes in Arabidopsis might be higher than previously assumed thus supporting the existence of multiple gene copies [76]. Gene expression analysis could support the discrimination of pseudo genes, because low gene expression in Arabidopsis was reported to be associated with pseudogenization [77]. Despite the unclear status of the gene product, the pure presence of these sequences revealed fascinating insights into genome evolution and contributed to the pan-genome [78, 79].

To detect the most important gene differences between Col-0 and Nd-1 without a strong bias through the applied prediction mechanisms [14], we searched via tBLASTn for genes completely absent from the other genome sequence. The number of 43 unique genes in Nd-1 (AdditionalFile23) and 42 unique genes in Col-0 (AdditionalFile24) are in accordance with the number of 40 genes in Ler-0 and 63 genes in Col-0, respectively, reported before [23]. Since the fast evolution of plant genomes [70, 80] is mainly based on gene duplications, presence/absence variations should have a severe impact. Moreover, harboring over 60% of genes with paralogous copies in the same genome [70, 81] makes copy number alterations more likely [76] to occur than the loss of a single copy gene. Changing the function of redundant gene copies e.g. derived from whole genome duplications [67, 82, 83] or transposon-mediated duplications [84, 85] poses a much higher potential for the acquisition of new functions than the *de novo* emergence of so called orphan genes from intergenic regions [70, 86, 87]. Orphan genes are frequently defined as unique to a specific phylogenetic lineage [88, 89]. The identification of these genes originating from non-coding sequences is challenging e.g. due to unique structural properties [90] or fragmented assemblies [68]. Sufficient information about genome sequences of closely related species is needed to distinguish *de novo* developed orphan genes e.g. from gene duplications with a following deletion of the original gene copy [88]. Orphan genes were previously described as a potential source of species-specific differences [89, 91] posing one explanation for accession-specific phenotypic differences. Functional analysis of the orphan genes identified in the high quality genome assemblies of the first *A. thaliana* accessions with a high quality genome assembly is needed to check if this holds true for phenotypic differences between plant accessions. It will be interesting to see if the rise of novel genes is more important for speciation events than the accumulation of mutations in existing genes.

Conclusions

We report a high quality long read *de novo* assembly (AthNd-1_v2) of the *A. thaliana* accession Nd-1, which improved significantly on the previously released NGS assembly sequence AthNd-1_v1.0 [21]. Comparison of the GeneSet_Nd-1_v2.0 with the Col-0 reference sequence genes revealed 24,572 RBHs supporting an overall synteny between both *A. thaliana* accessions except for an 1 Mbp inversion at the north of chromosome 4. Moreover, large structural variants were identified in the pericentromeric regions. Comparisons with the reference sequence also lead to the identification of the collapsed locus around At4g22214 in the Col-0 reference sequence. Therefore, this work contributes to the increasing *A. thaliana* pan-genome with significantly extended details about genomic rearrangements.

List of abbreviations

NGS	next generation sequencing
NOR	nucleolus organizing region
RBH	reciprocal best hit
SMRT	single molecule real time

Declarations

Ethics approval and consent to participate

Not applicable

Consent for publication

Not applicable

Availability of data and materials

The data sets supporting the results of this article are included within the article and its additional files. The Ath-Nd-1_v2 assembly is available upon request. Sequencing reads were submitted to the SRA (SRP066294).

Competing interest

The authors declare that they have no competing interest.

Funding

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

Author’s contributions

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

Acknowledgements

We thank Willy Keller for isolating high molecular DNA, Katharina Kemmet for extensive genotyping of plants for the genetic map, Helene Schellenberg, Ann-Christin Polikeit and

Prisca Viehöver for Sanger sequencing, and Melanie Kuhlmann as well as Andrea Voigt for taking excellent care of the plants.

Additional Files

AdditionalFile1. Protocol for extraction of high molecular weight genomic DNA for SMRT sequencing.

This protocol was used to extract high molecular genomic DNA from leaves of *A. thaliana* Nd-1 plants suitable for SMRT sequencing.

AdditionalFile2. Sequencing Statistics.

Statistical information about the generated SMRT sequencing data for the *A. thaliana* Nd-1 genome assembly are listed in this table. The expected genome size is based on several analyses reporting values around 150 Mbp [4, 5].

AdditionalFile3. FALCON assembly parameters.

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

AdditionalFile4. Molecular markers for genetic linkage analysis.

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

AdditionalFile5. Distribution of genetic markers over physical map.

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

AdditionalFile6. Oligonucleotide sequences for genetic linkage analysis.

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

AdditionalFile7. Transposable element positions in the Nd-1 genome sequence.

TE genes, TEs and TE fragments in the Nd-1 genome sequence were identified based on sequence similarity to annotated TEs from the Col-0 reference sequence (Araport11) [16].

AdditionalFile8. Nd-1 plastome map.

The GC content (black) and GC skew (green for positive GC skew, purple for negative GC skew) of the plastome sequence were analyzed by CGView [43]. The sequence and its properties are very similar to the Col-0 plastome sequence.

AdditionalFile9. Nd-1 chondrome map.

The GC content (black) and GC skew (green for positive GC skew, purple for negative GC skew) of the chondrome sequence were analyzed by CGView [43]. The sequence and its properties are very similar to the Col-0 chondrome sequence.

AdditionalFile10. BUSCO analysis of the Col-0 and Nd-1 genome sequences.

BUSCO v2.0 was run on the genomic sequences of Col-0 and Nd-1 using AUGUSTUS 3.2.1 with default parameters for the gene prediction process. The main difference between both gene sets is the absence of At3g01060 and At5g01010 from the Nd-1 genome assembly sequence. However, this is only caused by an assembly error, since the presence of these genes in the genome was validated by PCR and Sanger sequencing.

AdditionalFile11. Experimental validation of 1 Mbp inversion on chromosome 4.

The identified inversion between Nd-1 and Col-0 on chromosome 4 is different from the inversion described before between Col-0 and L*er* [23]. However, the left breakpoint is the same for both alleles enabling the use of previously published oligonucleotide sequences [23]. The right breakpoint was identified by manual investigation of sequence alignments. Both breakpoints were validated via PCR using the oligonucleotides as illustrated in (a) (AdditionalFile6). The results support the expected inversion borders (b).

AdditionalFile12. Genome-wide distribution of genes inserted on chromosome 2 in Nd-1.

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

AdditionalFile13. Genome-wide distribution of large structural variants.

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

AdditionalFile14. Clustering of SVs around centromeres.

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

AdditionalFile15. Transposable element overlap with GeneSet_Nd-1_v2.0.

The overlap between annotated TEs (AdditionalFile7) and predicted protein coding genes was analyzed to identify TE genes. This figure illustrates the fraction of a gene that is covered by a TE. Since TEs might occur within the intron of a gene, only genes with at least 80% TE coverage were flagged as transposable element genes (AdditionalFile16).

AdditionalFile16. Transposable element genes in GeneSet_Nd-1_v2.0.

These genes were predicted by AUGUSTUS as protein coding genes. Due to their positional overlap with TEs (AdditionalFile7), they were flagged as TE genes and excluded from further gene set analysis.

AdditionalFile17. Reciprocal best hits (RBH) pairs between Col-0 and Nd-1.

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

AdditionalFile18. Reciprocal best hits (RHB) indicates inversion between Nd-1 and Col-0.

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

AdditionalFile19. RBH outliers in GeneSet_Nd-1_v2.0.

Reciprocal bidirectional best BLAST hits (RBHs) between the gene sets of Col-0 and Nd-1 were identified. All 242 RBHs at positions deviating from the syntenic diagonal line were collected. The functional annotation of these genes was derived from Araport11.

AdditionalFile20. Duplicated genes in Nd-1.

The listed 385 Col-0 genes (Araport11 [16]) have at least two copies in Nd-1. Exons of these genes showed an increased copy number in Ath-Nd-1_v2 compared to the Col-0 reference sequence. The annotation was derived from Araport11.

AdditionalFile21. Duplicated genes in Col-0.

The listed 394 Nd-1 genes have at least two copies in Col-0. Exons of these genes showed an increased copy number in the Col-0 reference sequence compared to Ath-Nd-1_v2.

AdditionalFile22. Duplicated genes with significantly enriched functions.

Copied genes leading to significantly overrepresented functions in Col-0 or Nd-1, respectively. The listed genes are located in the center of networks which are significantly enriched in one accession due differences in the gene copy numbers. g:profiler [52] predicted the enrichment of specific functions in the set based on the ENSEMBL 89 annotation.

AdditionalFile23. List of unique Nd-1 genes in GeneSet_Nd-1_v2.0.

tBLASTn of the encoded peptide sequenced did not reveal a significant hit against the Col-0 reference genome sequence.

AdditionalFile24. List of unique Col-0 genes in Araport11.

tBLASTn of the encoded peptide sequenced did not reveal a significant hit against the Nd-1 genome sequence or the Nd-1 subreads.

AdditionalFile25. Critical regions in the Col-0 reference sequence.

The high continuity of the Ath-Nd-1_v2 assembly enabled the investigation of 22 sequences corresponding to gaps in the TAIR10 reference sequence (Col-0). This figure illustrates the homotetranucleotide occurrence in these sequences (red dots) in comparison to some randomly selected reference sequences (green dots). While there is a clear enrichment of homotetranucleotides in the gap-homolog sequences, there was no clear correlation between the length of a gap and the composition of the corresponding sequence observed.

References

1. Koornneef M, Meinke D: **The development of Arabidopsis as a model plant.** *Plant J* 2010, **61**(6):909-921.
2. Leutwiler LS, Hough-Evans BR, Meyerowitz EM: **The DNA of Arabidopsis thaliana**. *Molecular Genome and Genetics* 1984, **194**:15-23.
3. Francis DM, Hulbert SH, Michelmore RW: **Genome Size and Complexity of the Obligate Fungal Pathogen, Bremia lactucae**. *Experimental Mycology* 1990, **14**:299-309.
4. Arumuganathan K, Earle ED: **Nuclear DNA Content of Some Important Plant Species**. *Plant Mol Biol Reptr* 1991, **9**(3):208-218.
5. Höfte H, Desprez T, Amselm L, Chiapello H, Caboche M, Moisan A, Jourjon M-F, Charpentau J-L, Berthomieu P, Guerrier D *et al*: **An inventory of 1152 expressed sequence tags obtained by partial sequencing of cDNAs from *Arabidopsis thaliana***. *Plant J* 1993, **4**(6):1051-1061.
6. Fransz PF, Armstrong S, de Jong JH, Parnell LD, van Drunen C, Dean C, Zabel P, Bisseling T, Jones GH: **Integrated cytogenetic map of chromosome arm 4S of A. thaliana: structural organization of heterochromatic knob and centromere region.** *Cell* 2000, **100**(3):367-376.
7. Pruitt RE, Meyerowitz EM: **Characterization of the genome of *Arabidopsis thaliana***. *J Mol Biol* 1986, **187**:169-183.
8. Fransz P, Armstrong S, Alonso-Blanco C, Fischer TC, Torres-Ruiz RA, Jones G: **Cytogenetics for the model system Arabidopsis thaliana.** *Plant J* 1998, **13**(6):867-876.
9. Chang C, Bowman JL, DeJohn AW, Lander ES, Meyerowitz EM: **Restriction fragment length polymorphism linkage map for Arabidopsis thaliana**. *Proc Natl Acad Sci USA* 1988, **85**:6856-6860.
10. Bell CJ, Ecker JR: **Assignment of 30 microsatellite loci to the linkage map of *Arabidopsis***. *Genomics* 1994, **19**(1):137-144.
11. Lister C, Dean C: **Recombinant inbred lines for mapping RFLP and phenotypic markers in *Arabidopsis thaliana***. *Plant J* 1993, **4**:745-750.
12. Copenhaver GP, Browne WE, Preuss D: **Assaying genome-wide recombination and centromere functions with Arabidopsis tetrads.** *Proceedings of the National Academy of Sciences of the United Stated of America* 1998, **95**(1):247-252.
13. Kumekawa N, Hosouchi T, Tsuruoka H, Kotani H: **The size and sequence organization of the centromeric region of arabidopsis thaliana chromosome 5.** *DNA Res* 2000, **7**(6):315-321.
14. Blanc G, Barakat A, Guyot R, Cooke R, Delseny M: **Extensive duplication and reshuffling in the Arabidopsis genome**. *Plant Cell* 2000, **12**(7):1093-1101.
15. The Arabidopsis Genome Initiative: **Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana***. *Nature* 2000, **408**(6814):796-815.
16. Cheng CY, Krishnakumar V, Chan A, Thibaud-Nissen F, Schobel S, Town CD: **Araport11: a complete reannotation of the Arabidopsis thaliana reference genome.** *Plant J* 2017(89):789-804.
17. Schneeberger K, Ossowski S, Ott F, Klein JD, Wang X, Lanz C, Smith LM, Cao J, Fitz J, Warthmann N *et al*: **Reference-guided assembly of four diverse Arabidopsis thaliana genomes.** *Proceedings of the National Academie of Sciences of the United States of America* 2011, **108**(25):10249-10254.
18. Li YH, Zhou G, Ma J, Jiang W, Jin LG, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L *et al*: **De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits.** *Nat Biotechnol* 2014, **32**(10):1045-1054.
19. Consortium TG: **1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana.** *Cell* 2016, **166**(2):481-491.
20. Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin CS, Rapicavoli NA, Rank DR, Li J *et al*: **Long-read, whole-genome shotgun sequence data for five model organisms.** *Scientific Data* 2014, **1**:140045.

21. Pucker B, Holtgräwe D, Rosleff Sörensen T, Stracke R, Viehöver P, Weisshaar B: **A De Novo Genome Sequence Assembly of the Arabidopsis thaliana Accession Niederzenz-1 Displays Presence/Absence Variation and Strong Synteny.** *PLoS ONE* 2016, **11**(10):e0164321.
22. Huddleston J, Chaisson MJP, Steinberg KM, Warren W, Hoekzema K, Gordon D, Graves-Lindsay TA, Munson KM, Kronenberg ZN, Vives L *et al*: **Discovery and genotyping of structural variation from long-read haploid genome sequence data.** *Genome Res* 2017, **27**(5):677-685.
23. Zapata L, Ding J, Willing EM, Hartwig B, Bezdan D, Jiao WB, Patel V, Velikkakam James G, Koornneef M, Ossowski S *et al*: **Chromosome-level assembly of Arabidopsis thaliana Ler reveals the extent of translocation and inversion polymorphisms.** *Proc Natl Acad Sci USA* 2016.
24. Simpson JT, Pop M: **The Theory and Practice of Genome Sequence Assembly.** *Annual Review of Genomics and Human Genetics* 2015, **16**:153-172.
25. Rhoads A, Au KF: **PacBio Sequencing and Its Applications.** *Genomics Proteomics Bioinformatics* 2015, **13**(5):278-289.
26. Lam KK, Khalak A, D. T: **Near-optimal assembly for shotgun sequencing with noisy reads.** *BMC Bioinf* 2014, **15**:S4.
27. Koren S, Phillippy AM: **One chromosome, one contig: complete microbial genomes from long-read sequencing and assembly.** *Current Opinion in Microbiology* 2015, **23**:110-120.
28. Shoromony **I**, Courtade T, Tse D: **Do Read Errors Matter for Genome Assembly?** In: *IEEE International Symposium on Information Theory (ISIT).* Hong Kong; 2015: 919-923.
29. Koren S, Harhay GP, Smith TP, Bono JL, Harhay DM, Mcvey SD, Radune D, Bergman NH, Phillippy AM: **Reducing assembly complexity of microbial genomes with single-molecule sequencing.** *Genome Biol* 2013, **14**(9):R101.
30. Pendleton M, Sebra R, Pang AW, Ummat A, Franzen O, Rausch T, Stütz AM, Stedman W, Anantharaman T, Hastie A *et al*: **Assembly and diploid architecture of an individual human genome via single-molecule technologies.** *Nature Methods* 2015, **12**(8):780-786.
31. Berlin K, Koren S, Chin CS, Drake JP, Landolin JM, Phillippy AM: **Assembling large genomes with single-molecule sequencing and locality-sensitive hashing.** *Nat Biotechnol* 2015, **33**(6):623-630.
32. Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A *et al*: **Phased diploid genome assembly with single-molecule real-time sequencing.** *Nature Methods* 2016, **13**(12):1050-1054.
33. Muir P, Li S, Lou S, Wang D, Spakowicz DJ, Salichos L, Zhang J, Weinstock GM, Isaacs F, Rozowsky J *et al*: **The real cost of sequencing: scaling computation to keep pace with data generation.** *Genome Biol* 2016, **17**(53).
34. Consortium. TCP-G: **Computational pan-genomics: status, promises and challenges.** *Briefings in Bioinformatics* 2016:1-18.
35. Li L, Stoeckert CJJ, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13**(9):2178-2189.
36. Ward N, Moreno-Hagelsieb G: **Quickly finding orthologs as reciprocal best hits with BLAT, LAST, and UBLAST: how much do we miss?** *PLoS ONE* 2014, **9**(7):e101850.
37. Tatusov RL, Koonin EV, Lipman DJ: **A genomic perspective on protein families.** *Science* 1997, **278**(5338):631-637.
38. Healey A, Furtado A, Cooper T, Henry RJ: **Protocol: a simple method for extracting next-generation sequencing quality genomic DNA from recalcitrant plant species.** *Plant Methods* 2014, **10**(21).
39. Stadermann KB, Holtgräwe D, Weisshaar B: **Chloroplast Genome Sequence of Arabidopsis thaliana Accession Landsberg erecta, Assembled from Single-

Molecule, Real-Time Sequencing Data. *Genome Announcements* 2016, **4**(5):e00975-00916.

40. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM: **Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation.** *Genome Res* 2017, **27**(5):722-736.

41. Boetzer M, Henkel CV, Jansen HJ, Butler D, Pirovano W: **Scaffolding pre-assembled contigs using SSPACE.** *Bioinformatics* 2011, **27**(4):578-579.

42. Kleinboelting N, Huep G, Appelhagen I, Viehoever P, Li Y, Weisshaar B: **The Structural Features of Thousands of T-DNA Insertion Sites Are Consistent with a Double-Strand Break Repair-Based Insertion Mechanism.** *Molecular Plant* 2015, **8**(11):1651-1664.

43. Stothard P, Wishart DS: **Circular genome visualization and exploration using CGView.** *Bioinformatics* 2005, **21**(4):537-539.

44. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, Leunissen JA: **Primer3Plus, an enhanced web interface to Primer3.** *Nucleic Acids Res* 2007, **35**:W71-74.

45. Rosso MG, Li Y, Strizhov N, Reiss B, Dekker K, Weisshaar B: **An *Arabidopsis thaliana* T-DNA mutagenised population (GABI-Kat) for flanking sequence tag based reverse genetics.** *Plant Mol Biol* 2003, **53**(1):247-259.

46. Pucker B, Holtgräwe D, Weisshaar B: **Consideration of non-canonical splice sites improves gene prediction on the Arabidopsis thaliana Niederzenz-1 genome sequence.** *BMC Res Notes* 2017, **10**(1):667.

47. Arend D, Junker A, Scholz U, Schüler D, Wylie J, Lange M: **PGP repository: a plant phenomics and genomics data publication infrastructure.** *Database* 2016.

48. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva EV, Zdobnov EM: **BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs.** *Bioinformatics* 2015, **31**(19):3210-3212.

49. Keller O, Kollmar M, Stanke M, Waack S: **A novel hybrid gene prediction method employing protein multiple sequence alignments.** *Bioinformatics* 2011, **27**(6):757-763.

50. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**(2):R12.

51. Vukašinović N, Cvrčková F, Eliáš M, Cole R, Fowler JE, Žárský V, Synek L: **Dissecting a hidden gene duplication: the Arabidopsis thaliana SEC10 locus.** *PLoS ONE* 2014, **9**(4):e94077.

52. Reimand J, Arak T, Adler P, Kolberg L, Reisberg S, Peterson H, Vilo J: **g:Profiler-a web server for functional interpretation of gene lists (2016 update).** *Nucleic Acids Res* 2016, **44**((W1)):W83-89.

53. Cho YS, Kim H, Kim HM, Jho S, Jun J, Lee YJ, Chae KS, Kim CG, Kim S, Eriksson A *et al*: **An ethnically relevant consensus Korean reference genome is a step towards personal reference genomes.** *Nature Communications7* 2016:13637.

54. Seo JS, Rhie A, Kim J, Lee S, Sohn MH, Kim CU, Hastie A, Cao H, Yun JY, Kim J *et al*: **De novo assembly and phasing of a Korean human genome.** *Nature* 2016, **538**(7624):243-247.

55. Chandrasekhara C, Mohannath G, Blevins T, Pontvianne F, Pikaard CS: **Chromosome-specific NOR inactivation explains selective rRNA gene silencing and dosage control in Arabidopsis.** *Genes Dev* 2016, **30**(2):177-190.

56. Copenhaver GP, Pikaard CS: **RFLP and physical mapping with an rDNA-specific endonuclease reveals that nucleolus organizer regions of Arabidopsis thaliana adjoin the telomeres on chromosomes 2 and 4.** *Plant J* 1996, **9**(2):259-272.

57. Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H *et al*: **The Arabidopsis lyrata genome sequence and the basis of rapid genome size change.** *Nat Genet* 2011, **43**(5):476-481.

58. Schmuths H, Meister A, Horres R, Bachmann K: **Genome size variation among accessions of Arabidopsis thaliana.** *Ann Bot (Lond)* 2004, **93**(3):317-321.

59. Villasante A, Abad JP, Méndez-Lago M: **Centromeres were derived from telomeres during the evolution of the eukaryotic chromosome.** *Proceedings of the National Academy of Sciences of the United Stated of America* 2007, **104**(25):10542-10547.

60. Stupar RM, Lilly JW, Town CD, Cheng Z, Kaul S, Buell CR, Jiang J: **Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats.** *Proceedings of the National Academy of Sciences of the United Stated of America* 2001, **98**(9):5099-5103.

61. Kowalski SP, Lan TH, Feldmann KA, Paterson AH: **Comparative mapping of Arabidopsis thaliana and Brassica oleracea chromosomes reveals islands of conserved organization.** *Genetics* 1994, **138**(2):499-510.

62. Unseld M, Marienfeld JR, Brandt P, Brennicke A: **The mitochondrial genome of Arabidopsis thaliana contains 57 genes in 366,924 nucleotides.** *Nat Genet* 1997, **15**(1):57-61.

63. Martínez-Zapater JM, Gil P, Capel J, Somerville CR: **Mutations at the Arabidopsis CHM locus promote rearrangements of the mitochondrial genome.** *Plant Cell* 1992, **4**(8):889-899.

64. Christensen AC: **Plant mitochondrial genome evolution can be explained by DNA repair mechanisms.** *Genome Biology and Evolution* 2013, **5**(6):1079-1086.

65. Preuten T, Cincu E, Fuchs J, Zoschke R, Liere K, Börner T: **Fewer genes than organelles: extremely low and variable gene copy numbers in mitochondria of somatic plant cells.** *Plant J* 2010, **64**(6):948-959.

66. Woloszynska M, Gola EM, Piechota J: **Changes in accumulation of heteroplasmic mitochondrial DNA and frequency of recombination via short repeats during plant lifetime in Phaseolus vulgaris.** *Acta Biochim Pol* 2012, **59**(4):703-709.

67. Wendel JF, Jackson SA, Meyers BC, Wing RA: **Evolution of plant genome architecture.** *Genome Biol* 2016, **17**(37):s13059-13016-10908-13051.

68. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW: **Extensive error in the number of genes inferred from draft genome assemblies.** *PLoS Computational Biology* 2014, **10**(12):e1003998.

69. Paquette SM, Bak S, Feyereisen R: **Intron-exon organization and phylogeny in a large superfamily, the paralogous cytochrome P450 genes of Arabidopsis thaliana.** *DNA Cell Biol* 2000, **19**(5):307-317.

70. Panchy N, Lehti-Shiu M, Shiu SH: **Evolution of Gene Duplication in Plants.** *Plant Physiol* 2016, **171**(4):2294-2316.

71. Tan S, Zhong Y, Hou H, Yang S, Tian D: **Variation of presence/absence genes among Arabidopsis populations.** *BMC Evolutionary Biology* 2012, **12**(86):1471-2148/1412/1486.

72. Benovoy D, Drouin G: **Processed pseudogenes, processed genes, and spontaneous mutations in the Arabidopsis genome.** *J Mol Evol* 2006, **62**(5):511-522.

73. Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH: **Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice.** *Plant Physiol* 2009, **151**(1):3-15.

74. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M *et al*: **Empirical analysis of transcriptional activity in the Arabidopsis genome.** *Science* 2003, **302**(5646):842-846.

75. Siena LA, Ortiz JP, Calderini O, Paolocci F, Cáceres ME, Kaushal P, Grisan S, Pessino SC, Pupilli F: **An apomixis-linked ORC3-like pseudogene is associated with silencing of its functional homolog in apomictic Paspalum simplex.** *J Exp Bot* 2016, **67**(6):1965-1978.

76. Zmienko A, Samelak-Czajka A, Kozlowski P, Szymanska M, Figlerowicz M: **Arabidopsis thaliana population analysis reveals high plasticity of the genomic region spanning MSH2, AT3G18530 and AT3G18535 genes and provides

evidence for NAHR-driven recurrent CNV events occurring in this location. *BMC Genetics* 2016, **17**(1):893.

77. Yang L, Takuno S, Waters ER, Gaut BS: **Lowly expressed genes in Arabidopsis thaliana bear the signature of possible pseudogenization by promoter degradation.** *Mol Biol Evol* 2011, **28**(3):1193-1203.
78. Marroni F, Pinosio S, Morgante M: **Structural variation and genome complexity: is dispensable really dispensable?** *Curr Opin Plant Biol* 2014, **18**:31-36.
79. Golicz AA, Batley J, Edwards D: **Towards plant pangenomics.** *Plant Biotechnol J* 2016, **14**(4):1099-1105.
80. Murat F, Van de Peer Y, Salse J: **Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes.** *Genome Biology and Evolution* 2012, **4**(9):917-928.
81. Blanc G, Wolfe KH: **Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes.** *Plant Cell* 2004, **16**(7):1667-1678.
82. Vision TJ, Brown DG, Tanksley SD: **The origins of genomic duplications in Arabidopsis**. *Science* 2000, **290**(5499):2114-2117.
83. Renny-Byfield S, Gallagher JP, Grover CE, Szadkowski E, Page JT, Udall JA, Wang X, Paterson AH, Wendel JF: **Ancient gene duplicates in Gossypium (cotton) exhibit near-complete expression divergence.** *Genome Biology and Evolution* 2014, **6**(3):559-571.
84. Bennetzen JL: **Transposable elements, gene creation and genome rearrangement in flowering plants.** *Current Opinion in Genetics & Development* 2005, **15**(6):621-627.
85. Sun W, Zhao XW, Zhang Z: **Identification and evolution of the orphan genes in the domestic silkworm, Bombyx mori.** *FEBS Lett* 2015, **589**:2731-2738.
86. Tautz D, Domazet-Lošo T: **The evolutionary origin of orphan genes.** *Nat Rev Genet* 2011, **12**(10):692-702.
87. Schmitz JF, Bornberg-Bauer E: **Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA.** *F1000Research* 2017, **6**(57).
88. Fischer D, Eisenberg D: **Finding families for genomic ORFans.** *Bioinformatics* 1999, **15**(9):759-762.
89. Khalturin K, Hemmrich G, Fraune S, Augustin R, Bosch TC: **More than just orphans: are taxonomically-restricted genes important in evolution?** *Trends Genet* 2009, **25**(9):404-413.
90. Klasberg S, Bitard-Feildel T, Mallet L: **Computational Identification of Novel Genes: Current and Future Perspectives.** *Bioinformatics and Biology Insights* 2016, **10**:121-131.
91. Xu Y, Wu G, Hao B, Chen L, Deng X, Xu Q: **Identification, characterization and expression analysis of lineage-specific genes within sweet orange (Citrus sinensis).** *BMC Genomics* 2015, **16**(995):s12864-12015-12211-z.

Figure Legends

Figure 1: Nd-1 genome structure.

Schematic pseudochromosomes are shown in black with centromere repeat positions in green. Red dots indicate positions of 45S rDNA fragments and an orange star represents a complete 45S rDNA transcription unit. Blue triangles indicate the positions of 5S rDNAs. The position of telomeric repeats is shown by purple triangles.

Figure 2: Inversion on chromosome 4.

The dotplot heatmaps show the similarity between small fragments of two sequences. Each dot indicates a match of 1 kbp between both sequences, while the color is indicating the similarity of the matching sequences. (a) Comparison of the Nd-1 genome sequence against the Col-0 reference sequence reveals a 1 Mbp inversion. (b) The L*er* genome sequence displays another inversion allele [23].

Figure 3: Highly divergent region on chromosome 2.

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

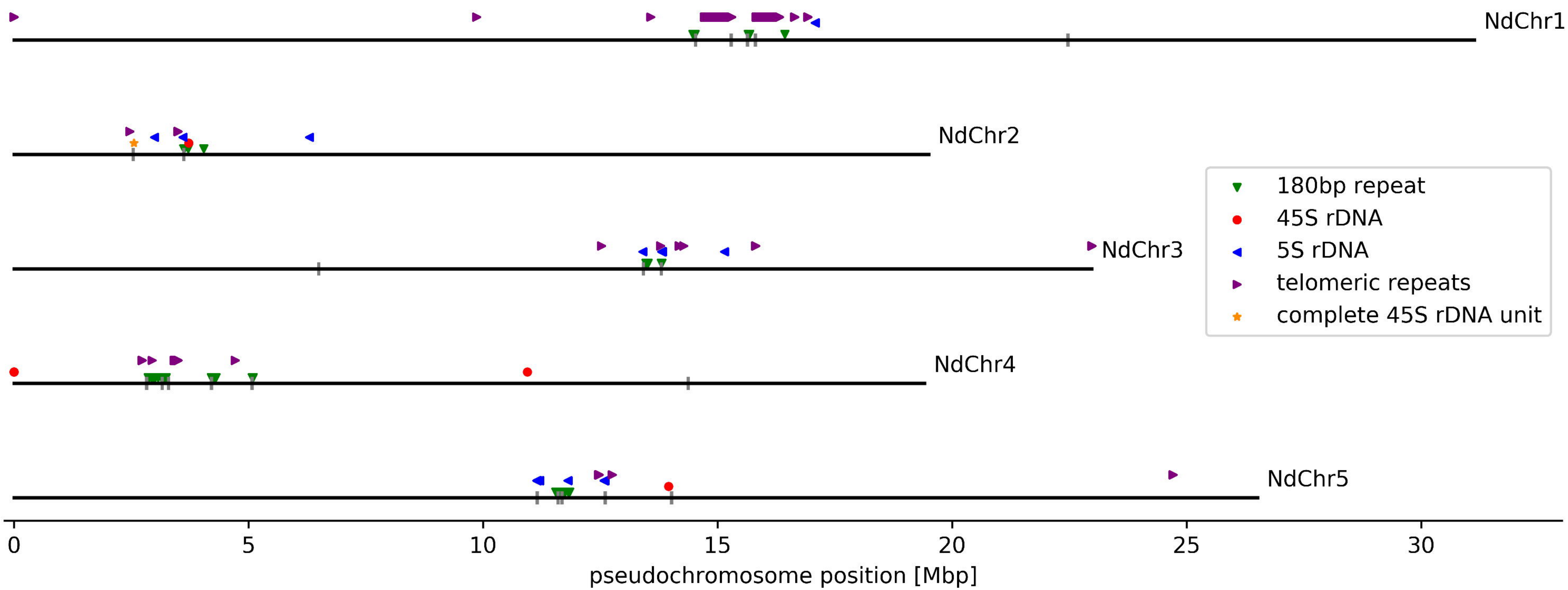
Figure 4: Hidden locus in the Col-0 reference sequence.

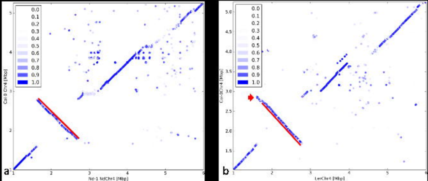
Differences between the Nd-1 and Col-0 genome sequences lead to the discovery of a collapsed region in the Col-0 reference sequence. There are two copies of At2g22214 (blue) present in the Col-0 genome, while only one copy is represented in the reference genome sequence. This gene duplication was initially validated through PCR with outwards facing oligonucleotides N258 and N259 (purple) which lead to the formation of the expected PCR product (black). Parts of this region were cloned into plasmids (grey) for sequencing. Sanger and paired-end Illumina sequencing reads revealed one complete gene (At4g22214b) and a degenerated copy (At4g22214a). Moreover, the region downstream of the complete gene copy in Nd-1 indicates the presence of at least one additional degenerated copy.

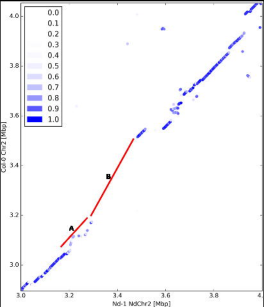
Table 1: Nd-1 *de novo* assembly statistics.

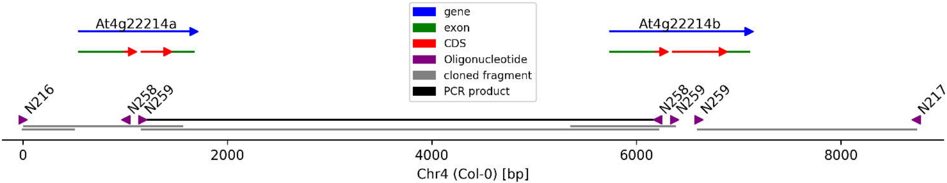
Metrics of the FALCON assembly of the Nd-1 nucleome sequence.

parameter	Nd-1 nucleome
number of contigs	26
total number of bases	119,540,544
average contig length	4,597,713 bp
minimal contig length	86,055 bp
maximal contig length	15,877,978 bp
GC content	36.04%
N25	14,534,675 bp
N50	9,302,209 bp
N75	6,666,836 bp
N90	2,829,734 bp


NdChr1
NdChr2
NdChr3
NdChr4
NdChr5
180bp repeat
45S rDNA
5S rDNA
telomeric repeats
complete 45S rDNA unit
0
5
10
15
20
25
30
pseudochromosome position [Mbp]


a
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Col-0 Chr4 [Mbp]
1
2
3
4
5
Nd-1 NdChr4 [Mbp]
1
2
3
4
5
6
b
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Col-0Chr4 [Mbp]
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
LerChr4 [Mbp]
1
2
3
4
5
6


0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1.0
A
B
Col-0 Chr2 [Mbp]
3.0
3.2
3.4
3.6
3.8
4.0
Nd-1 NdChr2 [Mbp]
3.0
3.2
3.4
3.6
3.8
4.


At4g22214a
At4g22214b
gene
exon
CDS
Oligonucleotide
cloned fragment
PCR product
N216
N258
N259
N258
N259
N259
N217
0
2000
4000
6000
8000
Chr4 (Col-0) [bp]