AMERICAN SOCIETY FOR MICROBIOLOGY | genomeAnnouncements

Chloroplast Genome Sequence of *Arabidopsis thaliana* Accession Landsberg *erecta*, Assembled from Single-Molecule, Real-Time Sequencing Data

Kai Bernd Stadermann,[a,b] Daniela Holtgräwe,[a] Bernd Weisshaar[a]

Chair of Genome Research, Faculty of Biology, Bielefeld University, Bielefeld, Germany[a]; Bioinformatics Resource Facility, Centre for Biotechnology, Bielefeld University, Bielefeld, Germany[b]

A publicly available data set from Pacific Biosciences was used to create an assembly of the chloroplast genome sequence of the *Arabidopsis thaliana* genotype Landsberg *erecta*. The assembly is solely based on single-molecule, real-time sequencing data and hence provides high resolution of the two inverted repeat regions typically contained in chloroplast genomes.

Received 19 July 2016 **Accepted** 8 August 2016 **Published** 22 September 2016

Citation Stadermann KB, Holtgräwe D, Weisshaar B. 2016. Chloroplast genome sequence of *Arabidopsis thaliana* accession Landsberg *erecta*, assembled from single-molecule, real-time sequencing data. Genome Announc 4(5):e00975-16. doi:10.1128/genomeA.00975-16.

Address correspondence to Kai Bernd Stadermann, kstaderm@cebitec.uni-bielefeld.de.

Arabidopsis thaliana is the most used model organism in plant genetics. During the last decades, various genotypes of *A. thaliana* have been sequenced using different sequencing technologies (1, 2). With third-generation sequencing technologies entering the market, novel techniques for DNA sequence generation have become available. The L*er*-0 genotype is currently the second most widely used accession of *A. thaliana* behind Columbia Col-0. The latest published L*er*-0 genome assembly used reads created by third-generation sequencing along with reads from other sources to create a high-quality nuclear genome assembly (3). One of the new technologies is single-molecule, real-time (SMRT) sequencing (4) developed by Pacific Biosciences, which is sometimes referred to as "PacBio sequencing". This sequencing technology allows long read lengths up to the 20- to 60-kb range, allowing single reads to span long repetitive elements or regions. The typical chloroplast genome contains two of these long (~26 kb) repeat regions, with the additional feature that the two regions are inverted repeats (IR) of each other. These IRs are hard to assemble using only short-read data from next-generation sequencing methods.

Recently, we described an approach to assemble a chloroplast genome based on an SMRT sequencing data set originally dedicated to sequencing a nuclear genome (5). We applied this approach to data from a single SMRT cell obtained from a publically available SMRT sequencing data set for *Arabidopsis thaliana* L*er*-0 provided by Pacific Biosciences (6). Sequence reads from sample number SAMN02724977 created by P5-C3 chemistry were used. We followed our protocol for identification of potential chloroplast reads, assembly, alignment, and polishing as described previously (5) with one exception: instead of using the spinach chloroplast genome sequence (7) as a template for read extraction, we used the chloroplast genome sequence of the *Arabidopsis thaliana* Col-0 genotype (8). We extracted 11,016 subreads summing up to 90,383,214 bp. The assembly resulted in a sequence with overlapping ends, and hence we assume that the chloroplast genome sequence is complete. We removed the additional overlap and aligned the sequence to the standard starting position of chloroplast genome sequences. The resulting assembly, designated Ler0_cp_smrt, has a total length of 154,515 bp and differs in 111 positions from the Col-0 chloroplast genome sequence (AP000423.1). We annotated the genome using CpAVAS (9) and identified 123 genes. Of these, 85 genes encode mRNA (i.e., proteins), 8 encode rRNA, and 30 encode tRNA.

Accession number(s). The complete sequence of the *Arabidopsis thaliana* L*er*-0 chloroplast genome was deposited in GenBank under accession number KX551970.

ACKNOWLEDGMENTS

We thank Pacific Biosciences for making such excellent data sets freely available to the public. We acknowledge support from the Deutsche Forschungsgemeinschaft and the Open Access Publication Fund of Bielefeld University for the article processing charge.

REFERENCES

1. **Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D.** 2008. Sequencing of natural strains of *Arabidopsis thaliana* with short reads. Genome Res **18**:2024–2033. http://dx.doi.org/10.1101/gr.080200.108.
2. **Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, Koenig D, Lanz C, Stegle O, Lippert C, Wang X, Ott F, Müller J, Alonso-Blanco C, Borgwardt K, Schmid KJ, Weigel D.** 2011. Whole-genome sequencing of multiple *Arabidopsis thaliana* populations. Nat Genet **43**:956–963. http://dx.doi.org/10.1038/ng.911.
3. **Zapata L, Ding J, Willing E-M, Hartwig B, Bezdan D, Jiao W-B, Patel V, Velikkakam JG, Koornneef M, Ossowski S, Schneeberger K.** 2016. Chromosome-level assembly of *Arabidopsis thaliana* Ler reveals the extent of translocation and inversion polymorphisms. Proc Natl Acad Sci U S A **113**:E4052-E4060. http://dx.doi.org/10.1073/pnas.1607532113.
4. **Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S.** 2009. Real-time DNA

sequencing from single polymerase molecules. Science **323:**133–138. http://dx.doi.org/10.1126/science.1162986.

5. **Stadermann KB, Weisshaar B, Holtgräwe D.** 2015. SMRT sequencing only *de novo* assembly of the sugar beet (*Beta vulgaris*) chloroplast genome. BMC Bioinformatics **16:**295. http://dx.doi.org/10.1186/s12859-015-0726-6.
6. **Kim KE, Peluso P, Babayan P, Yeadon PJ, Yu C, Fisher WW, Chin C-S, Rapicavoli NA, Rank DR, Li J, Catcheside DE, Celniker SE, Phillippy AM, Bergman CM, Landolin JM, Eid J, Clark TA, Flusberg BA, Travers KJ, Carneiro MO, Roberts RJ, Carneiro MO, Schatz MC, Koren S, Koren S, Chin CS, Li H, Durbin R, Chaisson MJ, Tesler G, English AC, English AC, Salerno WJ, Reid JG, Bankevich A, Mosher JJ, Thomas S, Underwood JG, Tseng E, Holloway AK, Tilgner H, Grubert F, Sharon D, Snyder MP, Voit RA, Hendel A, Pruett-Miller SM, Porteus MH, Bendall ML, Fang G, Kozdon JB, Song CX, Brown SD, Berlin K, Itsara A, Stankiewicz P, Lupski JR, Brizuela BJ, Celniker SE, Vogel HJ, Vogel HJ, Blattner FR, Engel SR, Galagan JE, Lamesch P, Yeadon PJ.** 2014. Long-read, whole-genome shotgun sequence data for five model organisms. Sci Data **1:**140045. http://dx.doi.org/10.1038/sdata.2014.45.
7. **Schmitz-Linneweber C, Maier RM, Alcaraz JP, Cottet A, Herrmann RG, Mache R.** 2001. The plastid chromosome of spinach (*Spinacia oleracea*): complete nucleotide sequence and gene organization. Plant Mol Biol **45:** 307–315.
8. **Sato S, Nakamura Y, Kaneko T, Asamizu E, Tabata S.** 1999. Complete structure of the chloroplast genome of *Arabidopsis thaliana*. DNA Res **6:**283–290. http://dx.doi.org/10.1093/dnares/6.5.283.
9. **Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X.** 2012. CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics **13:**715. http://dx.doi.org/10.1186/1471-2164-13-715.