The Plant Cell, Vol. 30: 525–527, March 2018, www.plantcell.org © 2018 ASPB.

LETTER TO THE EDITOR

Correction of Persistent Errors in Arabidopsis Reference Mitochondrial Genomes OPEN

Arabidopsis thaliana remains the foremost model system for plant genetics and genomics, and researchers rely on the accuracy of its genomic resources. The first completely sequenced angiosperm mitochondrial genome was obtained from Arabidopsis C24 (Unseld et al., 1997), and more recent efforts have produced additional Arabidopsis reference genomes, including one for Col-0, the most widely used ecotype in plant genetic research (Davila et al., 2011). These studies were based on older DNA sequencing methods, making them subject to errors associated with lower levels of sequencing coverage or the extremely short read lengths produced by early-generation Illumina technologies. Indeed, although the more recently published Arabidopsis mitochondrial reference genome sequences made substantial progress in improving upon earlier versions, they still have high error rates. By comparing publicly available Illumina sequence data to the Arabidopsis Col-0 reference genome, we found that it contains a sequence error every 2.4 kb on average, including 57 single-nucleotide polymorphisms (SNPs), 96 indels (up to 901 bp in size), and a large repeat-mediated rearrangement. Most of these errors appear to have been carried over from the original Arabidopsis mitochondrial genome sequence by reference-based assembly approaches, which has misled subsequent studies of plant mitochondrial mutation and molecular evolution by giving the false impression that the errors are naturally occurring variants present in multiple ecotypes. Building on the progress made by previous researchers, we provide a corrected reference sequence that we hope will serve as a useful community resource for future investigations in the field of plant mitochondrial genetics.

THE HISTORY OF ARABIDOPSIS MITOCHONDRIAL GENOME SEQUENCING

In 1997, a group led by Axel Brennicke reported the landmark achievement of sequencing a complete mitochondrial genome from Arabidopsis (Unseld et al., 1997), ushering flowering plants into the era of mitogenomics and providing numerous insights about the distinctive features of mitochondrial DNA (mtDNA) in plants (Mower et al., 2012). There has been some confusion over the source material used for this first sequencing effort. In the original 1997 publication and current GenBank accessions (Y08501.2 and NC_001284.2), the source ecotype is described as Columbia. However, the cosmid library used for sequencing was derived from C24 (Klein et al., 1994), which is genetically distinct from the widely used Col-0 (i.e., Columbia) ecotype (Lehle Seeds, 2004). The C24 source of the original published genome has been confirmed in subsequent studies (Davila et al., 2011). Notably, it is this original C24 sequence that is distributed with the current TAIR10 release of the Arabidopsis genome, even though the accompanying nuclear and plastid sequences are both derived from Columbia.

More recent efforts in the early phases of the "next-generation" sequencing revolution resequenced the mitochondrial genome of the C24 ecotype (GenBank accession no. JF729200) and produced reference sequences for the Col-0 (JF729201) and L*er* (JF729202) ecotypes (Davila et al., 2011). Resequencing of C24 yielded the same overall genome structure as the original sequence (Unseld et al., 1997) and earlier mapping efforts (Klein et al., 1994), but it also produced 416 sequence differences in the form of SNPs and small indels. At the time, there was no discussion or further investigation of these sequence differences, but they appear to represent corrections of sequencing errors from the original genome rather than true biological differences. Therefore, the work by Davila et al. (2011) has led to valuable increases and improvements in available mitogenomic resources for Arabidopsis. However, these efforts relied on some of the earliest implementations of Illumina sequencing technology. The extremely short read-lengths (35 bp) that were available at the onset of that study limited the researchers to reference-based assembly approaches, posing challenges for identification of variants in regions with multiple sequence differences. Therefore, the accuracy of the available Arabidopsis reference genomes has remained uncertain.

PERSISTENT SEQUENCING ERRORS IN PUBLISHED ARABIDOPSIS MITOCHONDRIAL GENOMES

While performing research to identify naturally occurring variants in Arabidopsis mtDNA (and being ignorant of some of the history described above), we were surprised to find that sequence data sets from Arabidopsis Col-0 exhibited numerous mitochondrial variants even when mapped against the Col-0 reference sequence. To investigate these discrepancies, we used a publicly available Illumina MiSeq data set (2 × 300-bp paired-end reads; NCBI SRA SRR5216995) to perform a de novo assembly of the Arabidopsis Col-0 mitochondrial genome (see Supplemental Methods). The resulting assembly differed by 57 SNPs and 96 indels relative to the published Arabidopsis Col-0 reference genome (Davila et al., 2011), amounting to a variant every 2.4 kb on average. Fifteen of these variants were found within introns of protein-coding genes, and two were in rRNA genes (Supplemental Data Set 1). The remainder were intergenic. To assess whether these variants represented

OPEN Articles can be viewed without a subscription.

www.plantcell.org/cgi/doi/10.1105/tpc.18.00024

sequencing artifacts or actual biological differences between the two Col-0 samples, we used custom scripts (https://github.com/dbsloan/arabidopsis_mitogenome) to extract diagnostic *k*-mers from the raw reads used in our analysis and those from the original Arabidopsis Col-0 sequencing effort (SRA SRR307226). We confirmed that all the variants identified in our assembly were strongly supported in both sets of sequencing reads (Supplemental Data Set 1), suggesting that the differences represent assembly errors in the published Col-0 reference sequence rather than real polymorphisms. We further validated these variants calls using the double-stranded consensus sequence from a data set (SRA SRR6420475) that was generated with a highly accurate technique known as duplex sequencing (Schmitt et al., 2012).

By comparing the same set of 57 SNPs and 96 indels to the raw reads in the resequenced C24 data set (SRA SRR307231), we identified 28 variants for which the original reference allele was supported in C24 (Supplemental Data Set 1). These cases, therefore, represent true polymorphisms that distinguish the C24 and Col-0 ecotypes but were not detected in the original reference-based assembly of the Col-0 mitochondrial genome, such that the published Col-0 sequence improperly retains the C24 allele. By contrast, we found that the raw C24 sequence reads did not support the original reference allele in the remaining 125 variants (82%) (Supplemental Data Set 1). These cases appear to result from errors in the original C24 genome sequence (Unseld et al., 1997) that were not detected in either the resequencing of C24 or the reference-based assembly of Col-0 and thus have been propagated across reported genome sequences from multiple ecotypes (Davila et al., 2011). Many of these errors are found in regions differing by multiple SNPs or by multinucleotide indels, so it is not surprising that they were difficult to detect with short-read sequencing data. However, there are also many individual SNPs and 1-bp indels in this set (Supplemental Data Set 1), so the source of the assembly artifacts is unclear in some cases.

Our newly assembled Arabidopsis Col-0 reference sequence also differs from the published Col-0 sequence in two major structural variants. First, it includes a 901-bp sequence that is absent from the published Col-0 genome. The full-length of this sequence is clearly detectable in the raw reads of the original Col-0 study (SRA SRR307226). It would be inserted after position 48,895 in the published Col-0 genome (JF729201) and would correspond precisely to the last 901 bp of the C24 reference genomes. The fact that this deletion occurs exactly at the point where the circular reference genome map had been arbitrarily “cut” for reporting as a linearized sequence suggests that it might have resulted from an inadvertent byproduct of sequence handling and reorientation. Second, our newly assembled Arabidopsis Col-0 reference sequence differs in a large rearrangement, apparently resulting from recombination between a pair of identical 453-bp inverted repeats at intergenic positions 36,362 to 36,818 and 143,953 to 144,409. We found evidence supporting the existence of both structural conformations, but the clear majority (30 of 33; 91%) of read pairs spanning these repeats support our reported conformation. We are not able to test for similar support in the raw Col-0 reads from Davila et al. (2011) because their insert sizes are too short to span the repeat copies, but we did verify that our reported configuration predominates in Illumina paired-end and PacBio sequencing reads from four other Col-0 data sets (NCBI SRA SRR1581142, SRR5012968, SRR5882797, and SRR3405242–SRR3405290). Furthermore, independent research by another group has previously used PCR-based screening and also found that this configuration predominates in Col-0 (José Gualberto, personal communication). Therefore, this is likely the most common configuration among different Col-0 seed stocks.

SUBSEQUENT RESEARCH IN ARABIDOPSIS MITOCHONDRIAL GENETICS

For good reason, Arabidopsis is the “go-to” model for studies of plant mitochondrial genome function, stability, mutation, and molecular evolution (Davila et al., 2011; Christensen, 2013; Cupp and Nielsen, 2014; Zampini et al., 2015; Gualberto and Newton, 2017). As such, there is great incentive to make the Arabidopsis reference mitochondrial genomes the gold standard in the field. Indeed, the extensive characterization of structural variation in these genomes has gone a long way toward accomplishing this goal (Arrieta-Montiel et al., 2009). However, sequence errors still exist in the reported reference genomes with potentially detrimental and far-reaching effects on related research efforts. This is especially true because the actual rate of sequence evolution in plant mtDNA is usually very low (Wolfe et al., 1987), so even a modest amount of sequencing errors can result in a problematic signal-to-noise ratio. For example, a recent study was performed to infer the distribution and spectrum of mutations across the Arabidopsis mitochondrial genome and used the sequence variants that distinguish published C24 and Col-0 mitochondrial DNA sequences (Christensen, 2013). Such comparative analyses of published genomic data are commonplace and can make substantial contributions to the field, but it is now clear based on our reexamination of the Col-0 sequence that ~40% of the analyzed variants in that study were artifacts (Supplemental Data Set 2).

Another recent investigation was conducted to detect de novo mutations in Arabidopsis organelle genomes using deep sequencing (Zampini et al., 2015). The authors applied a natural and seemingly conservative approach by rejecting any identified mitochondrial variant that did not differ from “both” published Col-0 mitochondrial genomes, but this choice highlights two pressing concerns. First, it illustrates the continued confusion in the field about the fact that original Arabidopsis reference mitochondrial genome is derived from C24 and not Col-0. Second, it reflects a misunderstanding about the extent to which the multiple available reference genomes constitute independent data points.

The reference-guided approach used to assemble mtDNA sequences from C24, Col-0, and L*er* (Davila et al., 2011) appears to have incorporated many errors and allelic variants from the reference genome into the new assemblies. Nevertheless, those new assemblies are still reported as separate accessions on GenBank rather than as a set of variant calls, so there is a risk that the many errors shared between them will be falsely perceived as having been independently validated in two or more sequencing data sets. This concern is particularly relevant for the L*er* sequence available on GenBank because it was generated with the same short 35-bp reads but a much lower level of sequence coverage—only

19× compared with 230× and 371× for Col-0 and C24, respectively (Davila et al., 2011). We have not attempted to build a de novo reference for L*er* ecotype, but researchers should be especially cautious about using the L*er* assembly currently available on GenBank (JF729202).

For all of the reasons outlined above, it is important that researchers in the field of plant mitochondrial genetics be more broadly aware of the history and methodologies that produced the currently available reference mitochondrial genome sequence for Arabidopsis. We have deposited our de novo assembly of the Arabidopsis Col-0 genome on GenBank (accession BK010421) in hopes that it will serve the community as a useful reference such that Arabidopsis can further develop as an outstanding model for elucidating mitochondrial genetic mechanisms.

Supplemental Data

Supplemental Methods. Data sources, code availability, read trimming, genome assembly, contig filtering and merging, assembly validation, and assessment of a plant mitochondrial genome assembly based on total cellular DNA.

Supplemental Data Set 1. *k*-mer based support for corrected Arabidopsis Col-0 sequence.

Supplemental Data Set 2. Identification of sequencing artifacts in comparative data set used to infer mutational spectrum in Christensen (2013).

ACKNOWLEDGMENTS

We thank *The Plant Cell* Editorial Board and three anonymous reviewers for their helpful comments. This research was supported by the National Institutes of Health (NIGMS R01 GM118046).

AUTHOR CONTRIBUTIONS

D.B.S., Z.W., and J.S. performed data analysis. D.B.S. wrote the manuscript.

Daniel B. Sloan
Department of Biology
Colorado State University
Fort Collins, Colorado
dbsloan@rams.colostate.edu

Zhiqiang Wu
Department of Biology
Colorado State University
Fort Collins, Colorado

Joel Sharbrough
Department of Biology
Colorado State University
Fort Collins, Colorado

REFERENCES

Arrieta-Montiel, M.P., Shedge, V., Davila, J., Christensen, A.C., and Mackenzie, S.A. (2009). Diversity of the Arabidopsis mitochondrial genome occurs via nuclear-controlled recombination activity. Genetics **183:** 1261–1268.

Christensen, A.C. (2013). Plant mitochondrial genome evolution can be explained by DNA repair mechanisms. Genome Biol. Evol. **5:** 1079–1086.

Cupp, J.D., and Nielsen, B.L. (2014). Minireview: DNA replication in plant mitochondria. Mitochondrion **19:** 231–237.

Davila, J.I., Arrieta-Montiel, M.P., Wamboldt, Y., Cao, J., Hagmann, J., Shedge, V., Xu, Y.Z., Weigel, D., and Mackenzie, S.A. (2011). Double-strand break repair processes drive evolution of the mitochondrial genome in Arabidopsis. BMC Biol. **9:** 64.

Gualberto, J.M., and Newton, K.J. (2017). Plant mitochondrial genomes: dynamics and mechanisms of mutation. Annu. Rev. Plant Biol. **68:** 225–252.

Klein, M., Eckert-Ossenkopp, U., Schmiedeberg, I., Brandt, P., Unseld, M., Brennicke, A., and Schuster, W. (1994). Physical mapping of the mitochondrial genome of *Arabidopsis thaliana* by cosmid and YAC clones. Plant J. **6:** 447–455.

Lehle Seeds (2004). More comments on C24 (Co-1/C24) glabra. http://www.lehleseeds.com/cgi-bin/hazel.cgi?action=detail&item=336&template=note1.html.

Mower, J.P., Sloan, D.B., and Alverson, A.J. (2012). Plant mitochondrial diversity: the genomics revolution. In Plant Genome Diversity, J.F. Wendel, ed (Vienna, Austria: Springer), pp. 123–144.

Schmitt, M.W., Kennedy, S.R., Salk, J.J., Fox, E.J., Hiatt, J.B., and Loeb, L.A. (2012). Detection of ultra-rare mutations by next-generation sequencing. Proc. Natl. Acad. Sci. USA **109:** 14508–14513.

Unseld, M., Marienfeld, J.R., Brandt, P., and Brennicke, A. (1997). The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366,924 nucleotides. Nat. Genet. **15:** 57–61.

Wolfe, K.H., Li, W.H., and Sharp, P.M. (1987). Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc. Natl. Acad. Sci. USA **84:** 9054–9058.

Zampini, É., Lepage, É., Tremblay-Belzile, S., Truche, S., and Brisson, N. (2015). Organelle DNA rearrangement mapping reveals U-turn-like inversions as a major source of genomic instability in *Arabidopsis* and humans. Genome Res. **25:** 645–654.