Release notes for *Dioscorea alata* reference genome assembly version 2

Overview

Yams (genus *Dioscorea*) are an important source of food and income for millions of smallholder farmers in the tropical and subtropical regions of Africa, Asia, the Pacific, and Latin America. Rich in carbohydrates, and containing protein and vitamin C, the year-round availability of yam tubers makes them preferable to seasonal crops. The importance of yams in West Africa is exemplified by their vital role in traditional culture, rituals, and religion. Yam production is declining, however, and under threat of pests and diseases. Thus, in the context of surging global population growth, improved yam breeding techniques are urgently needed.

Out of 600 different yam species, water yam, also called greater yam (*Dioscorea alata* L.) is the most widely distributed cultivated species in the world (Lebot 2009). It is superior to most cultivated yam species due to its potential to yield under low to average soil fertility, ease of propagation, early vigor for weed suppression, low post-harvest losses and high nutritional quality. Threats, however, include anthracnose disease, which can cause losses of over 80% of production (Nwankiti *et al.* 1984). Conventional breeding for desired traits in water yam is arduous due to its autopolyploid and heterozygous nature, long growth cycle, and erratic flowering. Thus, genomic tools are needed to facilitate accelerated improvement of this crop.

The water yam genome project is funded by an NSF-Gates Foundation BREAD grant “BREAD ABRDC: Development of Genomic Resources in Water Yam (*Dioscorea alata* L.) for Accelerated Breeding and Improvement” to Daniel Rokhsar (University of California Berkeley, California, USA), Ranjana Bhattacharjee (International Institute of Tropical Agriculture [IITA], Ibadan, Oyo, Nigeria), and Jude Obidiegwu (National Root Crops Research Institute [NRCRI], Umudike, Abia, Nigeria) (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1543967).

Our project is intended to provide a high-quality chromosome-scale *D. alata* genome assembly and consensus genetic map to the yam community, which will allow breeders to use modern genetic methods to breed the crop more efficiently. We will also characterize the natural genetic variability present in collections across different geographies, yielding insight into how they may be used to improve the crop, and contributing to an understanding of the relationship between water yam DNA sequence and desirable traits. Bringing water yam into the modern genomics era will facilitate the accelerated release to farmers of improved varieties.

An early draft version (v1) of the *D. alata* genome assembly, based on Illumina whole-genome shotgun and mate- pair sequence, was released in May 2018 in YamBase (ftp://yambase.org/genomes/Dioscorea_alata/), with accompanying annotation and release notes. It is also available as a bulk download in Phytozome (https://phytozome.jgi.doe.gov) and will be part of the next Phytozome release.

These notes describe the chromosome-scale reference *D. alata* genome assembly v2, a reference sequence generated from long-read single-molecule sequencing reads, Hi-C long-range linking information, and genetic linkage maps from four crosses, in addition to the data used for v1.

Sequencing, Assembly, and Annotation

Sequencing

The *D. alata* reference genome accession is TDa 95/00328, a breeding line from the IITA yam breeding collection. This accession was chosen because it was confirmed to be diploid by marker segregation analysis and is moderately resistant to anthracnose, a fungal disease caused by *Colletotrichum gloeosporioides* (Mignouna *et al*. 2002; Mignouna *et al.* 2001).

High molecular weight genomic DNA for Pacific Biosciences Single-Molecule Real-Time (SMRT) long-read sequencing was isolated by Dr. Prasad Hendre's group (World Agroforestry [ICRAF] and African Orphan Crops Consortium [AOCC]) in Nairobi, Kenya. The UC Davis DNA Technologies Core conducted library construction and sequencing. In total, we generated 21 Sequel SMRT cells (or 177× sequence depth). Half of the 106.4 Gb of generated bases were sequenced in reads 15.1 kb or longer. In addition, we sequenced 250 bp paired-end Illumina reads (625 bp insert size) to 108× depth and combined them with the 88× depth of 100 bp paired-end Illumina reads (255 bp insert size) generated by Saski *et al.* (2015).

Mate-pair libraries with 2 kb-, 4 kb-, and 7 kb-insert sizes were constructed by the UC Davis DNA Technologies Core from high molecular weight genomic DNA extracted by Dr. Ranjana Bhattacharjee's team at IITA Ibadan. Mate pairs were processed using the nxtrim (v0.4.2) software provided by Illumina and duplicates removed using Picard (v2.16.0). After filtering, 273.5, 217.9, and 72.2 million pairs (106×, 186×, and 108× depth) remained for each of the three libraries, respectively.

Hi-C long-range linking information was prepared by Dovetail Genomics LLC from fixed nuclei isolated in the laboratory of Dr. Jaroslav Doležel (IEB, Olomouc, Czech Republic), from TDa 95/00328 plants growing on site. The Hi-C libraries were sequenced by the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley, generating 297 million Hi-C read pairs.

Genetic Linkage Mapping

Four *D. alata* mapcross populations, ranging from 113 to 281 progenies, grown at IITA Ibadan were densely genotyped by DArT Canberra or Integrated Genotyping Service and Support (IGSS) at BecA-ILRI hub using the DArTseq reduced-representation method. SNP marker data were validated and filtered using an analytical pipeline for F_1 crosses adapted from ICGMC (2015, *G3*) and imputed using a custom script. Parent-specific linkage maps were constructed using the OneMap (commit 32def9f) R package.

Genome assembly

The PacBio long-read data were assembled into contigs with Canu (1.7-221-gb5bffcf) using the longest 90× depth of raw reads (*i.e*., reads 20 kb or longer). These contigs represented an incomplete subset of both TDa 95/00328 haplotypes and were subsequently filtered down to a single haploid complement of the genome via manual curation in JuiceBox (v1.8.9). The 532 remaining contigs were ordered and oriented into scaffolds using mate-pair data via SSPACE (v3). Hi-C reads were then aligned and duplicates removed using Juicer (v1.5.6), which identified 91 million genomic DNA-DNA contacts. These were used to create chromosome-scale scaffolds with the 3D-DNA (commit 2796c3b) genome scaffolding pipeline. Genetic linkage information was leveraged to identify large-scale misassemblies, which were later corrected via manual curation in JuiceBox. Finally, the assembly was polished twice via the Arrow (v2.2.2 from the smrtlink v6.0.0.47841 package) signal polishing tool using all 177× depth of coverage in long-reads, with any remaining detectable errors corrected by two subsequent rounds of Illumina-based polishing. This polishing was performed by calling variants from the combined 196× of TDa 95/00328

Illumina short-insert paired reads with FreeBayes (v1.1.0-54-g49413aa), then patching the assembly with a custom script. The final assembly is organized into 20 chromosomes and captures 480 Mb of total genomic sequence.

Annotation

From TDa 95/00328 plants growing on site, the Hendre group at AOCC Nairobi collected, extracted, and pooled RNA samples from 12 tissues. Aliquots of this RNA pool were sent to Jonathan Featherston at the AOCC partner Agricultural Research Council Biotechnology Platform (ARC-BTP) in Pretoria, South Africa; and to Gordon Simpson's group at the University of Dundee (UoD) and the James Hutton Institute (JHI), Dundee, UK. ARC performed the Illumina RNAseq library prep and sequencing, for a total of 41 million read pairs, sequenced 125 bp paired-end. The UoD/JHI group generated 626,000 Oxford Nanopore Technologies (ONT) direct RNA sequencing reads. These ONT reads were error-corrected with Proovread (v2.14.1), using merged reads created with FLASH (v1.2.11) from Illumina RNAseq read pairs. Corrected reads were aligned to the genome assembly with Minimap2 (v2.8) and collapsed into transcript models with Pinfish (https://github.com/nanoporetech/pinfish).

Protein-coding genes were annotated by Shengqiang Shu using the DOE Joint Genome Institute (JGI) plant genome annotation pipeline (Shu *et al.*, in preparation), which integrates transcript- and peptide-based gene prediction methods. Transcript assemblies were constructed with PERTRAN (Shengqiang Shu, unpublished) from 170 million pairs of Illumina RNAseq reads sourced from the above-described pooled library, Wu *et al.* 2015 (SRA: SRR1518381 and SRR1518382), and Sarah *et al.* 2017 (SRA: SRR3938623). The assemblies were then combined with 44k ESTs generated by Narina *et al.* (2011; SRA: SAMN00169815, SAMN00169801, SAMN00169798) and 18 full-length cDNAs collected from NCBI. Protein predictions from *D. rotundata* (Tamiru *et al.* 2017; BioProject PRJDB3383), arabidopsis, soybean, sorghum, rice, foxtail millet, amborella, eelgrass, banana, pineapple, grape, and Swiss-Prot proteomes were used as sources of homology evidence. Using BUSCO benchmarking (v3; embryophyta_odb9, N=1,440), the *D. alata* gene set is estimated to be 91.0% single copy, 3.6% duplicate, 1.9% fragmented, and missing 3.2% of genes.

Assembly and Annotation Statistics

Scaffold sequence total / count	480.0 Mb	25
Scaffold L50 / N50	24.0 Mb	9
Scaffold L90 / N90	19.5 Mb	18
Contig sequence total / count	479.5 Mb	532
Contig L50 / N50	4.5 Mb	31
Contig L90 / N90	565.0 kb	126

Annotation version	2.1
Taxonomy ID	5557
Primary transcripts (loci)	25,189
Alternate transcripts	13,414
Total transcripts	38,603

Primary transcripts:

Average number of exons	5.5
Median exon length	156
Median intron length	151
Number of complete genes	24,614
Number of incomplete genes with start codon	218
Number of incomplete genes with stop codon	281

Gene model support:

Number of genes with Pfam annotation	19,599
Number of genes with Panther annotation	23,183
Number of genes with KOG annotation	10,939
Number of genes with KEGG Orthology annotation	6,849
Number of genes with E.C. number annotation	7,654

Data Release

We are committed to early data release and, as a service to the yam research and breeding community, we are therefore making the chromosome-scale *D. alata* genome assembly version 2 available prior to publication. Please note that gene names are provisional and subject to change over time.

We (the data producers) encourage the pre-publication use of these data in accordance with the Toronto data sharing principles described at https://www.nature.com/articles/461168a. We anticipate publishing a description of the genome and genetic linkage maps in 2020. Questions about data use should be directed to Dr. Jessica Lyons (see below).
By accessing these resources, data users agree that the following analyses are reserved for first description by the data producers:

- Genomic landscape and higher-order chromatin structure
- Chromosome structure comparison with *D. rotundata*
- Gene and repeat content analyses
- Analyses of genome-wide intra-species variation
- Genome-wide inter-species comparisons

Contributors

UC Berkeley
Daniel S. Rokhsar (UC Berkeley, DOE JGI)
Jessen V. Bredeson
Jessica B. Lyons

IITA Nigeria
Ranjana Bhattacharjee
Ibukun Ogunleye
Antonio Lopez-Montes
Michael Abberton
Robert Asiedu

NRCRI
Jude Obidiegwu
Chiedozie Egesi (NRCRI, IITA Nigeria, Cornell Univ.)

AOCC
Allen Van Deynze (UC Davis, AOCC)
Prasad Hendre (ICRAF, AOCC)
Robert Kariba (ICRAF, AOCC)
Samuel Muthemba (ICRAF, AOCC)
Ramni Jamnadass (ICRAF, AOCC)
Alice Muchugi (ICRAF, AOCC)

DOE JGI
Shengqiang Shu
Joseph Carlson

IEB
Jaroslav Doležel
Eva Hřibová

Dundee Univ./JHI
Gordon Simpson (Dundee Univ., JHI)
Geoff Barton (Dundee Univ.)
Matthew Parker (Dundee Univ.)
Katarzyna Knop (Dundee Univ.)
Nick Schurch (Dundee Univ.)

ARC
Jonathan Featherston

Acknowledgements

UC Davis DNA Technologies Core; Dovetail Genomics; Vincent J. Coates Genomics Sequencing Center at UC Berkeley; Diversity Arrays Technology Pty Ltd; Integrated Genotyping Support and Service (IGSS). RNAseq was funded by the Illumina Greater Good Initiative. Nanopore sequencing was funded by the University of Dundee GCRF Challenge Fund.

Contacts

Jessica Lyons (UC Berkeley) (email: jblyons -AT- berkeley -DOT- edu)
Ranjana Bhattacharjee (IITA) (email: r.bhattacharjee -AT- cgiar -DOT- org)
Jude Obidiegwu (NRCRI) (email: ejikeobi -AT- yahoo -DOT- com)

References

International Cassava Genetic Map Consortium (ICGMC) (2015). High-resolution linkage map and chromosome-scale genome assembly for cassava (*Manihot esculenta* Crantz) from 10 populations. *G3: Genes, Genomes, Genetics* 5(1): 133.

Lebot, V (2009). Tropical root and tuber crops: cassava, sweet potato, yams and aroids. Vol. 17. Cabi.

Mignouna HD, Abang MM, Green KR, Asiedu R (2001). Inheritance of resistance in water yam (*Dioscorea alata*) to anthracnose (*Colletotrichum gloeosporioides*). *Theoretical and Applied Genetics* 103(1): 52.

Mignouna H, Mank R, Ellis T, Van den Bosch N, Asiedu R, Abang M, Peleman J (2002). A genetic linkage map of water yam (*Dioscorea alata* L.) based on AFLP markers and QTL analysis for anthracnose resistance. *Theoretical and Applied Genetics* 105(5): 726.

Narina SS, Buyyarapu R, Kottapalli KR, Sartie AM, Ali MI, Robert A, et al. (2011). Generation and analysis of expressed sequence tags (ESTs) for marker development in yam (*Dioscorea alata* L.). *BMC Genomics* 12(1): 100.

Nwankiti AO, Okpala EU, Odurukwe SO (1984). Effect of planting dates on the incidence and severity of anthracnose/blotch disease complex of *Dioscorea alata* L., caused by Colletotrichum gloeosporioides Penz., and subsequent effects on the yield. *Beit Trop Landwirtschaft und Veterinarmed* 22(3): 285.

Sarah G, Homa F, Pointet S, Contreras S, Sabot F, Nabholz B, et al. (2017). A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives. *Molecular Ecology Resources* 17(3): 565.

Saski CA, Bhattacharjee R, Scheffler BE, Asiedu R (2015). Genomic Resources for Water Yam (*Dioscorea alata* L.): Analyses of EST-Sequences, *de novo* Sequencing and GBS Libraries. *PLoS One* 10(7): e0134031.

Tamiru M, Natsume S, Takagi H, White B, Yaegashi H, Shimizu M, et al. (2017). Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination. *BMC Biology* 15(1): 86.

Wu ZG, Jiang W, Mantri N, Bao XQ, Chen SL, Tao ZM (2015). Transcriptome analysis reveals flavonoid biosynthesis regulation and simple sequence repeats in yam (*Dioscorea alata* L.) tubers. *BMC Genomics* 16(1): 346.