Release notes for *Dioscorea alata* genome assembly Version 1, May 2018

Overview

Yams (genus *Dioscorea*) are an important source of food and income for millions of smallholder farmers in the tropical and sub-tropical regions of Africa, Asia, the Pacific, and Latin America. Rich in carbohydrates, and containing protein and vitamin C, the year-round availability of yams 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, due to threats from 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. 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 up to 90% of production. Breeding for desired traits in water yam is arduous due to its autopolyploid and heterozygous nature, long growth cycle, and erratic flowering.

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” (https://www.nsf.gov/awardsearch/showAward?AWD_ID=1543967) to Daniel Rokhsar (University of California, Berkeley), Ranjana Bhattacharjee (International Institute of Tropical Agriculture [IITA], Ibadan, Nigeria), and Jude Obidiegwu (National Root Crops Research Institute [NRCRI], Umudike, Nigeria).

Our project is intended to provide a high-quality chromosome scale water yam genome assembly and consensus genetic map to the yam community, which will in turn 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 of improved varieties to the farmers that need them.

A major goal of our project is a high-quality chromosome-scale reference sequence for *D. alata.* Toward this end, we have generated Illumina shotgun and mate-pair data for the reference accession TDa 95/00328. Long-read PacBio sequencing of the reference accession is currently ongoing, along with the development of a dense genetic map from eight map-crosses. A high-quality chromosome scale genome sequence is anticipated by the end of 2018.

Data Release

We are committed to early data release. As a service to the research and breeding community, we are therefore making an early draft assembly of the *D. alata* genome available prior to the completion of the high-quality chromosome-scale genome sequence assembly. This sequence accounts for roughly half of the total genome, but an estimated 80–90% of protein-coding loci (that is, much of the missing sequence is repetitive, non-protein-coding portions of the genome).

We encourage the pre-publication use of this data according to the Fort Lauderdale data sharing principles described at http://www.sanger.ac.uk/legal/assets/fortlauderdalereport.pdf. We anticipate completion of the high-quality genome assembly (incorporating ongoing PacBio sequencing and genetic mapping) in the fall of 2018, and that we will submit a description of this genome and comparative analysis with related species in a manuscript to be submitted by the end of 2018. Questions about data use should be directed to Dr. Jessica Lyons <jblyons@berkeley.edu>.

Statistics

Genome

The early draft genome assembly of the reference accession TDa 95/00328 is not yet organized into chromosomes and is available only as scaffolds (reconstructed genome segments that may contain gaps whose size is known).

- Total scaffold sequence: 287.3 Mb (9.9% gap)
- Scaffold N/L50: 450 scaffolds longer than 145.7 kb account for half of the assembly
- Contig N/L50: 7,052 contiguous sequences longer than 9.0 kb account for half of the assembled contigs.

Protein-coding Loci

- 21,728 total loci containing protein-coding transcripts
- 29,190 total transcripts (including alternatively spliced isoforms)

Sequencing, Assembly, and Annotation Methods

Sequencing

The reference genome accession for *D. alata* is TDa 95/00328, from the IITA collection. We generated 108× sequence depth in 2 × 250 bp paired-end Illumina sequences (625 bp insert size), and 490×, 847×, and 1,305× clone coverage in mate pair libraries (with 2, 4, and 7 kb insert sizes, respectively). Genomic DNA was prepared at IITA Ibadan; library construction and sequencing were performed at UC Berkeley (shotgun) and UC Davis (mate pair). We also integrated 88× depth of 2 × 100 bp paired-end Illumina sequences with 255 bp insert size, from Saski *et al.,* 2015.

Assembly

The current early draft v1 assembly attempts to represent a single haplotype at non-repetitive regions of the genome. It captures about half of the genome but an estimated 80% of the protein-coding genes, due to a combination of the high repetitiveness and heterozygosity of the genome.

Briefly, all reads were trimmed to remove adapter sequences using fastq-mcf from the ea-utils packages (Aronesty 2011). Platanus v1.2.1 (Kajitani *et al*., 2014) was used for all stages of assembly. Contigs were generated using $k = 41$; all other parameters set to default. Due to high heterozygosity, all contigs with approximately half the expected genome-wide *k*-mer depth were removed. Filtered contigs were scaffolded requiring at least 5 mate pair links between contigs and a minimum contig overlap length of 40 bp for contig merging. Scaffolds were gap-filled with minimum overlap length between contig and assembled gap sequence in De Bruijn gap closing set to 40 bp, and minimum overlap length between reads in OLC gap closing stage also set to 40 bp.

Annotation

Protein-coding genes were annotated using the DOE Joint Genome Institute (JGI) plant genome annotation pipeline (Shu et al., unpublished), using genes predictions from *D. rotundata* (Tamiru *et al.* 2017, BioProject PRJDB3383) and other plant proteomes. We also integrated *D. alata* transcriptome data from Wu *et al.* 2015 (SRA Accessions SRR1518381 and SRR1518382), and Sarah *et al.* 2017 (SRA accession SRR3938623). 20,526 of the predicted peptides have homology to known peptides over more than 50% of their length, and 16,173 loci have RNAseq support across more than 50% of an associated transcript. Using the BUSCO methodology (v3; embryophyta_odb9, N=1,440), the *D. alata* gene set is estimated to be 79.5% complete (75.8% single copy, and 3.7% complete duplicate), 10% fragmented, and 10.5% missing.

Authors/Contributors

Jessen V. Bredeson (UC Berkeley), Jessica B. Lyons (UC Berkeley), Shengqiang Shu (DOE JGI), Ibukun Ogunleye (IITA Nigeria), Ranjana Bhattacharjee (IITA Nigeria), Jude Obidiegwu (NRCRI Nigeria), Daniel S. Rokhsar (UC Berkeley, DOE JGI)

Acknowledgements

UC Davis DNA Technologies Core; Vincent J. Coates Genomics Sequencing Center at UC Berkeley.

Contacts

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

References

Aronesty E (2011). “Ea-utils: Command-line tools for processing biological sequencing data.” https://github.com/ExpressionAnalysis/ea-utils

Kajitani R, Toshimoto K, Noguchi H, et al. (2014). “Efficient de novo assembly of highly heterozygous genomes from whole-genome shotgun short reads.” *Genome Res.* 24(8):1384-95. PMID: 24755901.

Sarah G, Homa F, Pointet S, et al. (2017). “A large set of 26 new reference transcriptomes dedicated to comparative population genomics in crops and wild relatives.” *Mol Ecol Resour.* 17(3):565-580. PMID: 27487989.

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. PMID: 26222616.

Tamiru M, Natsume S, Takagi H, et al. (2017). “Genome sequencing of the staple food crop white Guinea yam enables the development of a molecular marker for sex determination.” *BMC Biol.* 15(1):86. PMID: 28927400.

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* 16:346. PMID: 25924983.