The American Dahlia Society

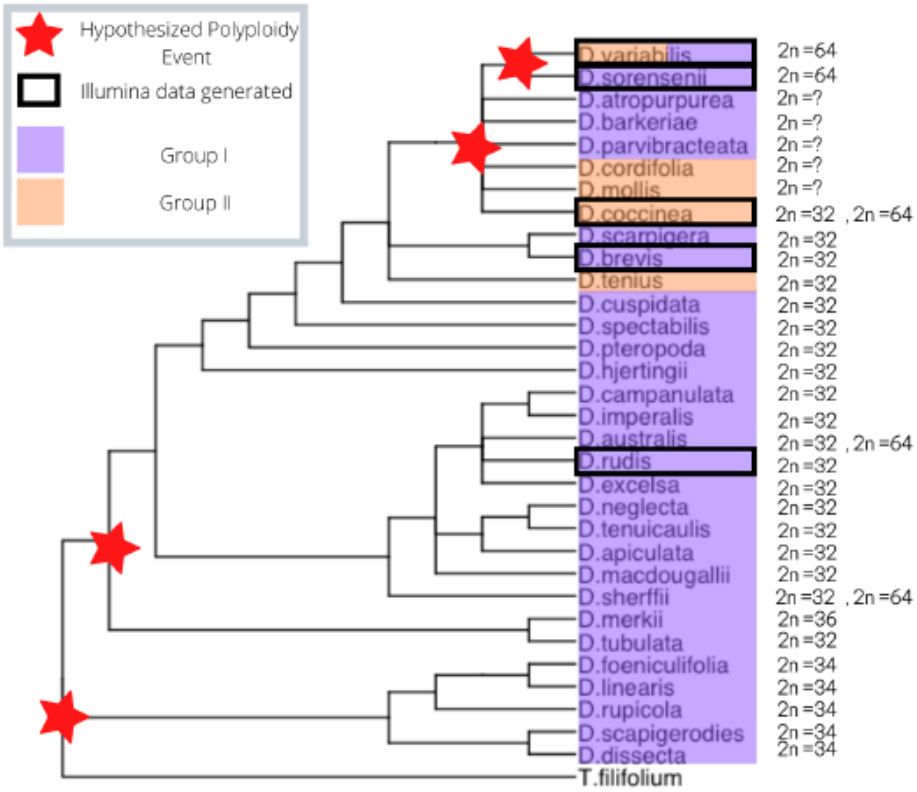
ADS Genome Project Update
March 2023

Project 1: Resolving the phylogeny of the *Dahlia* genus and identifying the putative parents of the cultivated octoploids

In collaboration with the Freeman Herbarium at Auburn University, Zach sourced herbarium material from 48 unique *Dahlia* species and subspecies, in replicate, totaling 105 sample sheets from herbaria across the United States and Mexico. DNAs will be enriched with the Compositae-1061 baits which targets 1061 highly conserved genes across all Asters. The libraries will be sequenced on a Illumina MiSeq. The sequence reads will be assembled using the Hybpiper pipeline. The fasta files created from Hybpiper will be aligned using MUSCLE and used as input for maximum likelihood tree estimation using RAxML. The resulting gene trees will be used in ASTRAL-Pro to construct multi-gene coalescent species trees using single- and multi-copy gene families. This tree will be compared with the phylogenies created by Gatt and Saar. With the addition of *Dahlia mixtecana, Dahlia tamaulipana, and Dahlia wixarika* to the genus and through the advancement in sequencing technology we expect this phylogeny to be a more accurate representation of the species relationships in the *Dahlia* genus.

Continued…

Using the updated *Dahlia* phylogeny, along with a k-mer based comparison pipeline, Zach will identify the most likely parents of *D. variabilis* and collect living samples from those species. Once grown, young leaf tissue will be collected and DNA will be extracted using a standard CTAB method before being Illumina whole genome shotgun sequenced. The data from the species will be added to the species he has already sequenced to >40X coverage: *D. brevis, D. coccinea, D. sorenesenii,* and *D. rudis* (Figure 1). K-mer lists for each species will be generated using jellyfish and mapped to each other using BWA to identify k-mers specific to each species. The species-specific k-mers will then be mapped to a cultivated dahlia (*D. variabilis*) to identify the putative ancestors to *D. variabilis*.



Continued…

By generating an updated phylogeny of the *Dahlia* genus Zach will be able to utilize gene trees to map the polyploidization events using PUG. Using BiSSE he will test for increased rate of diversification following the polyploidization events in the genus.

Zach has already generated Illumina sequencing data for 5 Dahlia species and is developing an analysis pipeline to test if *rudis, brevis, coccinea, and sorensenii* might be putative parents to the *D. variabilis* species complex. Zach is currently completing a pilot sequencing project for 24 individuals to test the efficacy of his bait capture sequencing approach, and we anticipate that data will be available in the next two months.

Zach recently wrote a grant proposal to the Botanical Society of America (2023 Bill Dahl Graduate Student Research Award) and the American Society of Plant Taxonomists (2023 ASPT Graduate Student) to fund some of this work, to complement the ADS funds we are grateful to receive.

Project 2: The Dahlia genome

Zach generated eight flowcells of data for the Edna C cultivar on a PacBio Sequel-II HiFi long-read platform, the greatest amount of Dahlia genomic data ever generated to our knowledge. He has currently assembled a first-pass genome assembly for Dahlia. Each of the "noodles" represented below is a contiguously assembled piece of chromosomal DNA. Like assembling a complex puzzle, not all of the pieces are in the correct order and orientation, but major blocks of the puzzle are correctly put together. We are generating an additional piece of data called Omni-C, which allows us to order and orient all of these large assembled pieces into full-length chromosomes. The ultimate challenge we will face is the true ability for us to fully assemble 4 complete genomes. Luckily, the current genome assembly has excellent statistics and metrics for quality. The pieces are large, meaning the input data was of high quality.