THE DAHLIA GENOME PROJECT

For 13 years, I served as the President of the Monterey Bay Dahlia Society. It's a group of about 60 dahlia growers on the Central Coast of California. As President, I was tasked with lining up speakers for our meetings. We are fortunate to have a well-known dahlia expert 40 miles away at Stanford University. Dr. Virginia Walbot is a biology professor specializing in corn genetics. She also maintains a dahlia garden at the University and uses dahlias, with their diverse forms and colors, to teach undergraduates about inheritance and plant genetics. Over the years, she and her students have done experiments with dahlias. I had heard that Dr. Walbot had a presentation on the science of flower color. In 2016, I invited her to one of our meetings to present her talk to our society.

Our group is typically pretty chatty. Members of the society have known each other for a long time, and our meetings were partially about socializing and catching up. The room grew quiet however once the lights went down and Dr. Walbot started speaking. With each progressive slide, our members became increasingly engrossed in the genetics of the pathways that produce color in dahlia blooms. As her presentation ended, she had a final slide, her dream slide. It expressed her desire that someday the dahlia genome would be sequenced. She explained that the only sequenced flower genomes were those that have significant commercial value, like roses and sunflowers. While dahlias have a lot of enthusiastic growers, there are no significant commercial interests that would foot the sizable bill for sequencing. She included the dream slide at the end of her presentation, hoping that perhaps someone in the audience would agree to fundraise for the sizable cost of genome sequencing.

That slide took me down a path that I am still on today. I had recently retired from a career as the Development Director of a Montessori school. Over eighteen years, my colleague and I raised $8 million in donations. I thought that Dr. Walbot's fundraising dream was achievable. So I decided to help make it happen. First, we discussed the project timeline and how much money would be needed. Not being a scientist, I was surprised to find out that the project would take many years. By the time the project is complete, we will have been working on it for about a decade.

Sequencing the dahlia genome is a huge task and requires hundreds of thousands of dollars. Fundraising to complete the job is a big part of the effort. On the scientific side, the dahlia genome is large compared to other plants and animals. Modern dahlias are octoploids, with four sets of genes from each parent. All of these extra chromosomes add complexity to the dahlia genome. Scientists compare the size of each genome using "base pairs." A base pair is

a tiny bit of DNA that forms a rung on the DNA ladder. Most plants have a genome of between 100 and 900 million base pairs. Strawberries (another complicated octopoid) have 813 million base pairs. The human genome consists of 3.2 billion base pairs. Each component of the dahlia genome is over 4 billion base pairs, thus as an octoploid with eight components, the dahlia genome is a 32 billion base pair species. The sheer size of the dahlia genome requires more time to document and more time to analyze and assemble into a usable data set.

I am often asked what will change for dahlias once the genome is fully sequenced and assembled. Not much will likely change in the first couple of years after completion. As the mechanics of sequencing DNA gets better, however, and the costs associated with genomics drop, dahlia breeders will have access to data that could significantly improve our breeding programs. It is helpful to use a hypothetical example to describe how the dahlia genome could change dahlia breeding.

Let's say I have a variety (we will call it variety A) that I like for its color; however, I want offspring with a different form. Variety A is an informal decorative form. I want to breed a variety with this same color but with a stellar petal form. I could cross variety A with all the stellar dahlias I can find. This may or may not get me closer to my goal. Because dahlias are packed with so many genes, the form each variety displays often differs from the offspring. My effort of crossing variety A with several potential parents is a game of chance with many rejects on the way and no guarantee of success.

Once the dahlia genome is sequenced and assembled, there will come a time when I can submit leaf samples to a lab for an inexpensive genetic scan. Estimates are the cost per variety will be less than ten dollars. Using our example above, I could send in 20 samples of potential mates for variety A and let the lab know I am looking for the genes resulting in the stellar petal form. This process will be similar to "23andMe" or "Ancestry," which we use to identify our human family tree. The lab will send back results showing which samples have the best chance of giving me the stellar form I desire. I could then make crosses with varieties more likely to give me the form I desire. Because breeders don't see the results of crosses for an entire year, this genetic “screening" could speed up the development of desired new varieties. This inexpensive screening could help breeders to find the genetic needles in the haystack they are looking for.

The ability to send in samples for screening could also provide the first significant steps in controlling the effects of viruses in dahlias. Dahlia viruses have been with us for centuries, and there is no cure for a plant with a virus. Several varieties appear to have a high tolerance

for viruses and grow and perform well while infected. In the future, geneticists may know what gene or genes confer virus tolerance in dahlias. Breeders could submit samples to determine which potential breeding parents carry the genes for virus tolerance. I can imagine an entire group of new varieties that are virus tolerant and asymptomatic.

For scientists, a reference genome will mean dahlias will be included in more scientific research. Scientists conducting genomic research on a particular plant species often include research on closely related species. The parallel track gives them greater insights. Often a specific gene and its function can't be determined in the target species, however, it may be successfully identified in the parallel species. If the two species are close relatives, the gene in question is likely in the same location on a particular chromosome. Locations on chromosomes are similar to street addresses in a neighborhood. In closely related species the order of genes is the same, just as the order of street addresses would be the same on parallel streets.

The closest relative to dahlias are sunflowers. Dahlias and sunflowers are both native to North America and share a wild habitat. The two plants originated only a few million years apart. Sunflower researchers, who support substantial research funding for this worldwide oil crop, have been waiting for a genome from a close cousin to compare to sunflowers. Once the dahlia reference genome is complete, dahlias are expected to be included in future sunflower research. For instance, if a scientist wants to search for the sunflower gene that may provide tolerance to higher temperatures, they would likely do parallel research on the dahlia genome simultaneously. In the process, they may discover what gene accounts for heat tolerance in dahlias.

Dr. Walbot and I worked quickly in those early days. There were just two of us, and we were excited to get started. We both committed to working as volunteers without compensation. This allowed 100% of the donations to go directly to the work. Although Dr. Walbot and I were excited and enthusiastic, we knew this project needed to be under the auspices of the American Dahlia Society (ADS). When looking for gifts, it is a benefit to donors to have tax deductibility through a non-profit organization.

I approached the ADS. Its mission includes spreading information and promoting the development of dahlias. The ADS Board agreed it was a project they should support. There was no professional fund-raiser on the ADS Board, so I was appointed to the Board's Finance Committee, responsible for Genome Project fundraising. We established a goal of $50,000 to achieve our early objectives. We reached our first fundraising goal in late 2016.

The following year Dr. Walbot received import permits from the USDA to import seeds collected from wild dahlias in Mexico.

In October 2017, Professor Walbot and her colleague Tim Culbertson collected dahlia leaf samples and seeds in Jalisco, Morelos, and Queretaro, Mexico. Professor Eduardo Ruiz Sanchez assisted them at the University of Guadalajara, along with botanists from the University of Morelos in Cuernavaca, Mexico.

In early 2018 I grew species dahlia seeds from the Mexico trip so scientists could have fresh leaf material. Leaf tissue was sent to a sequencing lab for RNA transcriptomes, an inexpensive way to start organizing the dahlia family tree. In the summer of 2018, Dr. Walbot, Tim Culbertson, and I traveled to Washington state to collect leaf samples from the gardens of Martin Kral, Wayne Lobaugh, and Brad and Rosemary Freeman. We were chauffeured around by Professor Alex Paradez from the University of Washington who also provided dry ice and liquid nitrogen needed for tissue preservation.

In the Fall of 2018, Dr. Walbot collected samples from several modern dahlia cultivars for RNA sequencing. Varieties sampled were 'Mexico', 'Jomanda', 'Rhonda', 'Thomas A. Edison', 'Emery Paul', and 'Pam Howden'. When Dr. Walbot shows up at my farm to take leaf samples, she arrives with stainless steel canisters of liquid nitrogen. The temperature inside the canister is -109° F (-78 C). Samples are taken carefully, documented, placed in labeled envelopes, and dropped into the cold canisters.

Later in 2018, samples from 15 species dahlias and 11 modern dahlias were sent out for RNA sequencing to Novogene, a sequencing lab in China. Again, the result of the RNA sequence was a surprise. The modern, fully double dahlias and the smallest species dahlias (open center with eight petals) are genetically indistinguishable. This suggests that modern and species dahlias spring from one shared gene pool. Similar to what we know about dogs: there is one shared gene pool with a wide range of characteristics.

As 2018 came to a close, a new scientist joined the effort: Dr. Alex Harkess, a plant biologist and National Science Foundation Plant Genome Initiative Postdoctoral Fellow. While a postdoctoral fellow he collaborated with Professor Walbot's lab, and she knew about his keen interest in floral evolution. He is a Faculty Investigator at the HudsonAlpha Institute for Biotechnology in Huntsville, Alabama. HudsonAlpha is the largest genome sequencing lab in the United States. Dr. Harkess is also an Assistant Professor in the Department of Crop, Soil, and Environmental Science at Auburn University in Auburn, Alabama. Dr. Walbot founded the Genome Project and did the initial ground-breaking work. She chose

Dr. Harkess to see the project through the subsequent phases, which will result in a fully assembled reference genome.

In 2019, Dr. Harkess and I started planning the second phase of the Dahlia Genome Project. This phase would require years of painstaking work by humans to recheck the automated initial assembly by computers. This human labor assembly is always understood to be the most time-consuming and expensive part of the project. Dr. Harkess compares the challenge of assembling the reference genome to a jigsaw puzzle. Assembly is like putting together a billion-piece jigsaw puzzle without a picture of the completed puzzle. To make things more complex, because modern dahlias are octoploids, imagine a box with eight complete sets of pieces all jumbled together. In addition, some puzzle pieces are double-sided, and 85% of the pieces look identical. Finally, many of the pieces don't even belong to dahlias. Sequencing plant material drags in DNA from fungi, bacteria, and viruses. These "foreign" puzzle pieces must be identified and thrown out of the mix. In the past, assembling the puzzle was done painstakingly by hand. Fortunately, today, computers are taught to recognize patterns in DNA fragments through computational biology to produce a draft assembly, making human assembly more efficient. Even with computers, figuring out what strings of DNA go where is the challenge that will take a passionate graduate student years to complete.

At the 2019 National Dahlia Show in Grand Rapids, Michigan, Dr. Harkess gave a presentation explaining what it means to sequence a genome and how the project should progress. Specifically, he expressed the need for a talented graduate student who could do the lab work for the full assembly while working toward a Ph.D. When his talk was over, a young man approached Dr. Harkess and told him that stimulated by his presentation, he was considering applying to be the graduate student Dr. Harkess sought. His name was Zach Meharg.

One result of moving the project to HudsonAlpha in Alabama is that my farm could no longer grow plant material for DNA sequencing. Stanford University is only 40 miles away. HudsonAlpha is 2,400 miles away. Using ADS membership lists, I searched for experienced dahlia growers within driving distance of the lab. Marcie Holt in Ringgold, Georgia, came to our aid. She is an experienced grower and agreed to grow the varieties Dr. Harkess would need for sampling. I sent her plants and tubers, both modern varieties and species dahlias. I had two other growers across the border in Alabama grow backup plants as a precaution.

After accepting Zach Meharg as a new graduate student, he and Dr. Harkess started growing dahlias at their facility for sampling. In addition, they collected samples from 800 varieties at

the ADS National Dahlia Show in Wooster, Ohio. Six hundred specimens have been sequenced to help put together the dahlia family tree. The variety used for the final reference genome is 'Edna C'.

One of my Genome Project responsibilities is to make monthly reports to the ADS Board on scientific progress. For me, this can be a challenge. The world of genomics has its own vocabulary. I am not a scientist (I was an art major in college), and it can be difficult to translate lab reports for the layperson. Here's an example of a recent report I was asked to explain.

"All tissues will be extracted using a standard CTAB method, checked for DNA quantity with the Qubit® 3.0 Fluorometer and purity using the Denovix DS-11 FX, and stored in a 4˚C refrigerator before library preparation starts. The extracted libraries will be enriched with the Angiosperms353 baits. The libraries will be sequenced on an 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."

Most of this goes over my head. There are frequent phone calls with Dr. Harkess to help me understand what I am reading and how best to communicate it to others.

The pace of work on dahlias at HudsonAlpha has picked up with graduate student Zach Meharg working full-time on the project. The work of understanding a species' genome is not a straight line. It involves discoveries from several angles.

One such project is exploring the genetic differences in sports from their parent plants. A sport is a dahlia bloom that is identical in form to the parent plant but displays a different color. For instance, a sport could be a yellow bloom on a plant that typically produces red blooms. Sports could provide a shortcut to understanding which genes are responsible for flower color. Searching the entire genome for color formation would be difficult. However, because a sport is genetically identical to the parent plant except for the genes controlling color, it might be possible to find those genes with a simple comparison.

In March of 2023, Zach and I hosted a Zoom meeting with several U.S hybridizers to gather up tubers or cuttings of sports and their parent cultivars. Zach will grow these plants in his greenhouse and take leaf tissue samples. He will then perform low-pass (inexpensive) sequencing to assemble a data set. Special software will then look for differences between

the parent and the sport. With luck, we may soon know which gene or genes control color variation in dahlias.

Studying sports is an example of looking for subtle differences between two data sets. An opposite strategy would determine what is similar in data sets. For example, Zach wants to know what genes control the floral forms in dahlias. What makes a cactus dahlia take that unique form? To answer this question, he intends to sequence multiple samples of a single form. Then, using software, he will compare several data sets of that form with those of another form, for example, cactus dahlias compared to informal decorative dahlias. By identifying what genes are common to the cactus form samples and different from the informal decorative samples, we may soon learn what genes control floral form.

Another area of scientific interest is when (in the growth of a dahlia bud) does it differentiate into its final floral form? If we dissected dahlia buds as they emerge at the tip of the stem, we would see undifferentiated meristem cells. As the bud develops, those cells differentiate and ultimately build the floral form we recognize as a particular dahlia form (like water lily or collarette). To look into a developing bud, Zach uses a laser-powered confocal microscope. This device allows scientists to see inside tissue with great clarity and depth of field. Using the microscope, Zach may be able to determine when the cells inside the bud start to differentiate. He can then collect RNA from the bud tissue. When switched on, genes produce RNA that encodes the proteins that carry out the genetic instructions. Identifying the RNA present when a bud starts differentiating may reveal the genes that determine floral form.

After many years of hard work to sequence the full dahlia genome, the Dahlia Genome Project is now exploring the above-mentioned projects and many more. By 2025 we should have a fully assembled genome. A few years later, breeders like myself may have the opportunity to get data similar to human genetic data derived from 23andMe today. That data could guide our future breeding goals and techniques. The American Dahlia Society maintains a web page on the Dahlia Genome Project at https://www.dahlia.org/docsinfo/genome-overview/