Growing Better Cotton for Future Generations

International collaboration led researchers to develop a cotton genotyping array, enabling breeders to produce more sustainable and valuable cotton.

Introduction

From the terry cloth in bath towels to the denim in blue jeans, cotton is used to make clothing and linens all over the world. The fiber is spun into yarn or thread to make textiles and the remaining cottonseed is used to produce oil, which is then refined and used for cooking. While most of the cotton supply is grown in China, India, and the United States, cotton is grown in over 75 countries. It is a hardy crop that can thrive in harsh climates and can be cultivated with little equipment, enabling small-farm holders in developing countries to grow and sell it easily.

Researchers and breeders are using genetic selection to develop new types of cotton that are more productive and resilient to challenging environments. Because cotton exists in both diploid and polyploid forms, it is difficult to crossbreed the 2 types of species, making genetic studies challenging. Until recently, few single nucleotide polymorphisms (SNPs) had been identified for cotton due to its highly complex genome. Researchers and breeders knew that certain cotton species had desirable characteristics, such as drought tolerance and insect resistance, but didn't know which genes conferred these traits. Breeders largely relied on visual inspection and quantitative assessments of the plants (eg, fiber yield, fiber quality, pest resistance) to select the best progeny for each breeding cycle, which is a time-consuming process.

Recent research in the lab of David Stelly, PhD at Texas A&M University is helping to change that. Amanda Hulse-Kemp, PhD, a postdoctoral research associate in Dr. Stelly's lab, focused her PhD research on identifying cotton SNPs and developing a genotyping tool to help breeders improve cotton crops. Through an international collaboration with researchers and Illumina, Dr. Hulse-Kemp and her colleagues developed a custom cotton genotyping array: the CottonSNP63K BeadChip.[1] International partners included the Commonwealth Scientific and Industrial Research Organization (CSIRO) in Australia, the Agricultural Research Centre for International Development (CIRAD) in France, the Council of Scientific and Industrial Research Organization's National Botanical Research Institute in India, and Illumina. The International Cotton SNP Chip Consortium designed the CottonSNP63K BeadChip to enable researchers and breeders to identify SNPs associated with valuable traits, leading to higher yield and hardier, disease-resistant crops.

iCommunity spoke with Dr. Hulse-Kemp to learn about how her work will transform the way researchers and breeders grow better cotton.

Q: Why did you get involved in cotton genomics?
Amanda Hulse-Kemp (AHK): Agriculture is the basis of society, and plants are a large part of that. By 2050, it's estimated that the global population will be approximately 9 billion. With the rapidly increasing population size, agriculture will play an even more important role. There's not only an immediate need to feed people, but also to provide clothing and linen supplies. For that reason, cotton will be an important crop for the future.

Amanda Hulse-Kemp, PhD, is a postdoctoral research associate in the Department of Soil and Crop Sciences at Texas A&M University. She received the Outstanding Achievement in Graduate Research Award in 2014 and the Distinguished Graduate Student Award in 2015 for her work on cotton genomics.

Q: How many different cotton species are you studying?
AHK: Typically, most breeders are only interested in the cultivated species. Our lab is interested in leveraging the diversity found in wild species. Wild species are unique because they've evolved naturally in multiple environments, so they have many specialized characteristics. Cultivated species tend to have similar traits, and there isn't much variety.

The 2 main cultivated species that we study are upland cotton (*Gossypium hirsutum*) and Pima cotton, sometimes called Egyptian cotton (*G. barbadense*). Upland cotton produces a large yield of moderate-quality cotton and makes up about 95% of the cotton grown worldwide. Pima cotton produces a higher-quality product with long fibers, which is used to make expensive clothing and linens. The problem with Pima cotton is that even though it's of higher quality, it produces a much lower yield than upland cotton, so the overall value to the farmers is lower. A goal for many cotton breeders is to develop an optimal cultivar that combines the high-yielding characteristics of upland cotton with the high-quality characteristics of Pima. However, both of these species still have limited diversity due to multiple genetic bottlenecks, which leaves them vulnerable to various pests and climate changes. That's where wild cotton species become incredibly valuable.

Q: What approaches have cotton breeders used in the past to select for traits?
AHK: Historically, cotton breeders have largely relied on visually inspecting plants to identify traits. Modern breeding involves qualitative and quantitative testing, including multi-location replicated testing at advanced generations to select the best progeny during each breeding cycle. Today, breeders cross many plants, looking for cotton lines that perform well or have valuable characteristics. This approach involves a lot of work, because it's basically a numbers game. You have to go through perhaps a million plants to capture that right combination of traits that makes a great cultivar. Even if you evaluate many lines, you still might not find exactly what you're looking for, because a single significant fault can render a line unusable for production.

Q: When did agricultural researchers start using genomics to study cotton?
AHK: We're at a turning point in using genomics to study cotton. By integrating array technology into our operations, our lab is working to demonstrate the ease, affordability, and value of genomic tools so that others can incorporate them into their research and breeding programs. Breeders are beginning to discover the value that array technology can bring to tracking cotton traits. They need a tracking tool that enables them to look at large numbers of plants at a low cost per sample. Arrays can help them do that. Also, genomics can help breeders evaluate and select for genes that affect lowly heritable or polygenic traits, which are difficult to manipulate using conventional methods.

For breeders, it's a numbers game. They need a tracking tool that enables them to look at large numbers of plants at a low cost per sample. Arrays can help them do that.

Q: How did your group start the International Cotton SNP Chip Consortium?
AHK: We started the International Cotton SNP Chip Consortium as a way to bring researchers together to create a low-cost SNP genotyping tool. When I started working on cotton genomics in 2010, everyone relied on simple sequence repeat (SSR) markers, or microsatellites. The problem with SSRs is that they're limited and not necessarily evenly distributed throughout the genome. The advantage of SNPs over SSRs is that they have the possibility of existing at any given base in the genome, so they provide a larger number of more evenly distributed markers. SNP markers can also be assayed in various ways, in a high-throughput fashion, and at a low cost per data point. They're cheap and reliable.

Dr. Stelly realized this, too. Working with Dr. Sukhwan (Sam) Yang (a postdoctoral researcher at the time), and later with Dr. Allen Van Deynze and others, he started identifying cotton SNPs. Given the complexities of the cotton genome, there were many hurdles to overcome. When I started my research program, only about 200 SNPs had been identified and mapped for cotton, so our lab set out to find more markers.

The Complex Cotton Genome

Although 2 of the 4 cultivated species of cotton are diploid (2 sets of chromosomes), the other 2 cultivated species are tetraploid (4 sets of chromosomes). The tetraploid genome is similar to a deck of 52 playing cards, with 2 red and 2 black suits (genomes), each with 13 cards (chromosomes). Complicating the cotton genome even further is the fact that both historical and recent data indicate that whole-genome duplication (polyploidization) preceded formation of the ancestral genome that gave rise to diploid cotton. Thus, upland cotton is at least paleo-octaploid, abounding with genetic redundancy, which makes genetic analysis, sequencing, and breeding a challenge.

By 2012, our group had developed tens of thousands of markers, and we connected with Dr. Iain Wilson at CSIRO because his group had also developed many SNPs. We now had enough to start genotyping on a large scale. The challenge was getting the price per sample low enough. So at the International Cotton Genome Initiative meeting in 2012, we distributed an announcement asking if anyone wanted to be a part of a collaboration to develop a cotton array. The cotton community was very supportive and interested in the overall goal. Afterwards, we worked with Illumina to develop an interest list. We then followed our announcement with a worldwide message asking if researchers from other countries would like to join the initiative, contribute to the SNP markers, and purchase the array.

By 2013, the Consortium decided on the size and types of content on the array. I identified the attributes of the SNP populations, removed duplicates, and trimmed the content down to about 70,000 SNPs. We sent it to Illumina for production in the fall of 2013, and we were delighted to have such high assay success in conversion to polymorphic markers for the array. The CottonSNP63K BeadChip was released in early 2014. It has been adopted quickly by researchers and several companies. Now we're trying to spread the word to other researchers and breeders, so that they too can benefit from the CottonSNP63K BeadChip.

Q: What were the goals of the International Cotton SNP Chip Consortium?
AHK: One of the main goals of the Consortium was the development of a cluster file for upland cotton and the other tetraploid species, so that everyone's CottonSNP63K BeadChip results would be standardized throughout the cotton community. Dr. Martin Ganal's group at TraitGenetics GmbH contributed their expertise to the cluster file development.

One of the challenges with working with inbred crops is that most of them are homozygous, so the heterozygous clusters aren't normally captured. Localizing the heterozygous cluster is the most difficult type of pattern to identify. Illumina helped us create the best cluster file possible by representing more varieties to make the CottonSNP63K BeadChip a more broadly useful tool. With a broad representation of the germplasm, we could identify a large portion of the clusters by including samples that belong to both the

homozygous and heterozygous clusters. Now anyone, including breeders, can send in their DNA sample, receive the genotyping data in 3 days, and immediately use the information to identify which lines to keep, select, or cross for their program. This tool enables us to associate the markers on the array with important traits and determine if they might be causal alleles or linked to causal regions for beneficial characteristics.

> The unique thing about the cluster file is that it's designed for tetraploid cotton. Anyone can take a tetraploid sample, put it through the file, and get a genotype.

Q: Which cotton species are represented on the CottonSNP63K BeadChip?
AHK: The CottonSNP63K BeadChip contains markers for *G. hirsutum, G. barbadense, G. tomentosum, G. mustelinum, G. longicalyx,* and *G. armourianum.* The BeadChip is primarily designed for *G. hirsutum* breeding, or integrating features of the other species into *G. hirsutum* lines, because *G. hirsutum* is the most widely cultivated species. *G. tomentosum* and *G. mustelinum* are tetraploids from Hawaii and Brazil, respectively. They have some characteristics for disease resistance and seem to be able to withstand high amounts of rain, which is especially useful with the heavy rainfall we've recently seen here in Texas. *G. mustelinum* has been shown to have the highest leaf concentration of terpenoid aldehydes, which affect insect resistance. *G. tomentosum* also has resistance to cotton leaf hopper and a few other insects. The 2 diploids represented on the BeadChip are *G. longicalyx* and *G. armourianum.* Because it originated in Africa, *G. longicalyx* theoretically has some drought resistance features, and has also been shown to have high resistance to the reniform nematode, which we're studying in our lab. *G. armourianum* has resistance to the white fly, which is a vector for the leaf curl virus, a new form of which has been devastating to cotton crops, particularly in Pakistan and India.

Q: How did you select the SNPs to include on the CottonSNP63K BeadChip?
AHK: Because research on cotton SNPs has been limited and we had only recently identified the SNPs, we did not have the ability to select SNPs that were associated with traits. Instead, we chose groups of SNPs that were known to be genic in origin and others that were not. By combining both types of SNPs, we assumed that the SNPs would be more or less evenly distributed across the genome.

The only reference genome for the genus *Gossypium* available at the time was from a wild, diploid, Peruvian small tree species. We didn't want to bias our SNP selection in the tetraploid species based on a diploid, so we used a random spread of SNPs based on design scores from each species. Now the community can use the CottonSNP63K BeadChip to make those associations between SNPs and traits.

Q: How did you validate the CottonSNP63K BeadChip?
AHK: We used 1158 samples to validate the CottonSNP63K BeadChip, which included many different types of samples so that we could populate each of the genotype clusters. Because cotton is a tetraploid species, it is likely that these genotyping assays evaluate 2 positions in the genome, because there would be 2 identical positions in the different subgenomes. As a result, it required manual validation of the cluster positions. Because they had previous experience analyzing polyploid data, our collaborators at TraitGenetics manually positioned or changed the position of the clusters in the cluster file where the automated clustering algorithm was insufficient due to the polyploidy. Through validation, we're making sure that we have samples that represent the 3 different cluster types to distinguish a successful polymorphic marker from a monomorphic or intergenomic marker. The segregating markers were further validated using 2 mapping populations.

Q: How will the genetic maps and the cluster file that you created benefit cotton researchers and breeders?
AHK: The unique thing about the cluster file is that it's designed for tetraploid cotton. Anyone can take a tetraploid sample, put it through the file, and get a genotype; it doesn't require any modifications to determine the true genotype. Also, this is the first time that cotton researchers have an intraspecific genetic map that links into linkage groups that represent chromosomes. Before, you would not have enough markers to associate them into the correct number of cotton chromosomes. Our group is also developing a cluster file for diploid cotton.

> Illumina has been really helpful throughout the whole process. If we had a question, they knew the answer.

Q: What was the value of having a consortium when you were developing the CottonSNP63K BeadChip?
AHK: The International Cotton SNP Chip Consortium was instrumental in developing the CottonSNP63K BeadChip. Everyone in the Consortium was fabulous and helpful. We had breeders, molecular geneticists, public agencies, representatives from companies, and a representative from the commodity organization Cotton Incorporated. They provided ideas and input on what their groups were interested in. Their input was instrumental in helping us make the CottonSNP63K BeadChip relevant to people in different areas of cotton research.

Because the CottonSNP63K BeadChip was so marketable, many samples were reserved in the initial buy-in, reducing per-sample costs for consortium members. This enabled us to process many samples cost-effectively. Otherwise, we could never have afforded to run all the SNPs on our own.

Q: What was your experience with Illumina in developing the Consortium?
AHK: Illumina has been really helpful throughout the whole process. I'm pretty sure that we would not have made it this far so quickly without input from Illumina. If we had a question, they knew the answer. They helped us with any technical difficulties and streamlined the process from formation to application. Illumina continues to support the Consortium by distributing BeadChips so we can run them in our lab, enabling us to get quotes to collaborators or send and receive samples from collaborators easily. The fact that you can use the array technology anywhere or send it to a service provider is useful because many of the breeders who will want to use the CottonSNP63K BeadChip don't have the technology or capability to run the array themselves.

> Almost everyone who works on cotton will use the CottonSNP63K BeadChip in the future.

Q: How did the International Cotton SNP Chip Consortium affect your doctoral research?
AHK: While not initially a part of my doctoral research proposal, developing the CottonSNP63K BeadChip became a large part of my PhD research. Dr. Stelly helped me gain valuable experience by having me work directly with the consortium members and participate in all the meetings and teleconferences. Not only have I used the CottonSNP63K BeadChip, but everyone in our lab is already using it, and almost everyone who works on cotton will use it in the future.

Developing a tool that molecular researchers and breeders can use and that incorporates lines produced in the public and private sectors is remarkable. I've met many people through the Consortium that I never would have otherwise. I think it's valuable to know that there is a tool out there that you can use to assay the same position at a point in the cotton genome, and that your research can easily be correlated with someone else's research.

Q: Do you have any advice for other graduate students studying cotton genomics?
AHK: Using array technology can significantly speed up graduate research. As Dr. Stelly likes to emphasize, students can benefit enormously because array technology expedites their progress, giving them a "quantum leap" forward and changing the timeline and scope of their research.

Most research at universities is conducted by graduate students and postdoctoral fellows with limited time. For many graduate students today, their whole doctoral project is making a high-density map or fine mapping a small region. They don't know where many of the traits they're dealing with are located. With an array, you can quickly run a mapping population and determine which regions you're interested in within a month rather than years. So, the final output of your PhD project can be a cultivar line instead of fine mapping that someone else will work on during their doctoral research. The impact on student projects is enormous. They can now do about 10 times the amount of work that they could do before with the same time and effort. It's much cheaper and faster to use arrays.

Q: What are the next steps in your research?
AHK: We're now working on the diversity analysis with the inbred lines that were included on the CottonSNP63K BeadChip for validation and development of the cluster file. Researchers can use those results to find out how close their lines are, how many markers they have between the parents, and things like that. They will then be able to associate the markers on the array with actual phenotypes, leading to streamlined development of superior cultivars.

> Using array technology, graduate students can now do about 10 times the amount of work that they could before with the same time and effort.

Q: Where do you see the future of cotton genomics research going from here?
AHK: The CottonSNP63K BeadChip will allow a revolution in cotton breeding as long as we can get it into the breeders' hands. This will allow for faster development of varieties so cotton farmers will be better equipped to deal with problems like climate change, drought, and disease. The disease-resistant, higher-quality lines will enable growers to obtain a higher yield per acre, which theoretically will generate additional income, something that will benefit all cotton producers in the future.

References

1. Hulse-Kemp AM, Lemm J, Plieske J, et al. Development of a 63K SNP array for cotton and high-density mapping of intraspecific and interspecific populations of *Gossypium* spp. *G3 (Bethesda).* 2015;5:1187-1209.

Learn More

Learn more about Illumina consortia at www.illumina.com/applications/genotyping/consortia.html.

Follow the Illumina Agrigenomics Community

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

illumina®