## TOWARDS A CLUSTER FILE FOR AUTOMATED GENOTYPING OF GERMPLASM INTROGRESSED FROM GOSSYPIUM DIPLOIDS

## A. N. Kyanam A. M. Hulse-Kemp F. Wang K. J. Kochan P. K Riggs D. M. Stelly Texas A&M University System College Station, TX

## <u>Abstract</u>

The "CottonSNP63K Array" is a cotton SNP "Chip" that was released in early 2014 for high-throughput genotyping. It is major new tool for global cotton research and breeding. Each array includes 70,000 potential 1-bead type Illumina Infinium II assays for both intra-specific (50,000) and interspecific (20,000) in silico SNPs relative to Upland cotton, Gossypium hirsutum (L.). Some of the 20,000 interspecific SNPs on the Chip were chosen to detect SNPs from diploid (2n=26) relatives of the secondary gene pool, specifically the Mexican species G. armourianum (A2-1 genome) and the African species G. longicalvx (F1 genome). Use of Infinium arrays for automated genotype calling requires a "cluster file" that a computer can use to interpret relative strengths of allele-specific fluorescent signals of each SNP locus. The computer uses the Chip data and the cluster file together to determine if a genotype can be assigned at a given level of probability to each locus of each sample, and if so, to "call" each genotype. Cluster files are thus critical to automated genotyping; each cluster file depends on the sample population from which the SNP signal distributions are collected. While a cluster file for cotton tetraploids was developed by TraitGenetics for the Cotton SNP Chip Consortium and is soon to be published, it is not designed for genotyping of diploids. DNA samples of diploids and tetraploids that include introgressed diploid germplasm behave distinctly different in terms of the occurrence and distribution of SNP signals, so the automated genotyping of such samples will require a different cluster file. Here, we report on progress toward development of a customized cluster file for Gossypium diploid genomes. Using the GenomeStudio software by Illumina, we have screened a subpopulation of the SNPs available on the CottonSNP63K Array, and manually adjusted for any anomalous graphical results. Distributions and ratios for specific families and populations can be used to guide further classification refinements. The file will also denote markers whose behavior does not conform to the norms specific to diploids. Once completed, we expect to make the diploid file publically available, so that it can be used to report putative genotypes and segregation ratios specific for diploids, or be used by others for further specialization.