Population Level Analysis to Move from Massive Sequence Data Sets to Application

David Francis[1], Sung-Chur Sim[1], Heather Merk[1], Allen Van Deynze[2], Kevin Stoffel[2], John Hamilton[3], C. Robin Buell[3], Dan Zarka[4], and David Douches[4]

[1]The Ohio State Univ., OARDC, Dept. of Horticulture and Crop Science, Wooster, OH 44691; [2]Univ. of California, Seed Biotechnology Center, Davis, CA 95616; [3]Michigan State Univ. Dept. of Plant Biology, East Lansing, MI, 48824; [4]Department of Crop and Soil Sciences, Michigan State University, East Lansing, MI 48824

Translating genome sequence data into applied outcomes in plant breeding requires that sequence data be available for germplasm that is relevant to crop improvement programs. To assess the distribution of genetic variation and inform future tomato breeding, "next generation" sequence data were generated for transcribed sequences from six tomato varieties and analyzed in a single nucleotide polymorphism (SNP) discovery pipeline. A public SNP array was developed from these analyses using allele frequency data and genome coverage as principle criteria. A panel of 410 tomato accessions ranging from land-race and vintage classes to elite parents was assembled and genotyped with 7,720 SNPs. Variation was visualized relative to both genetic maps (units in cM) and physical maps (base pair) as minor allele frequency, loading contribution to principal components, and F_{ST}-outlier as determined by deviation from the expected F_{ST}/He ratio. These analyses revealed regions of the genome that appear to be under selection due to crop improvement. Some of these regions are expected as they contain known introgressions; others provide unexpected insight into the biology of cultivated populations. Analysis of the SNPs suggested that recombination is limited in breeding populations and that selection and mapping within breeding programs could be accomplished with fewer SNPs. Optimized sets of 384 SNPs were developed for fresh-market and processing tomato germplasm pools, and these were used to genotype breeding populations. Processing tomato populations ranged from nested recom-

binant inbred populations to complex breeding populations developed to combine multiple sources of resistance to bacterial spot. The optimized SNPs performed well for both discovery (with QTL identified for yield, fruit size, and disease resistance) and selection. Next generation sequencing and high-throughput genotyping are providing resources to help achieve the long-standing promise of marker-assisted selection. In order to leverage these resources to increase gain under selection, breeding programs will benefit from structural changes to (1) increase recombination by leveraging larger populations; (2) systematic collection of objective data for key traits; (3) economic valuation of traits; (4) incorporation of pedigree information or kinship to strengthen estimates of breeding value using a random effects statistical framework; and (5) methods to shorten the breeding cycle and/or increase opportunities for selection during a given unit of time. Implementing these changes will improve gain under selection even without markers.