Utilizing low-coverage sequence data in tomato recombinant inbred lines (*S. lycopersicum x S. pimpinellifolium*)

Rianne van Binsbergen[ab], Marco C.A.M. Bink[c], Richard Finkers[d], Mario P.L. Calus[b], Roel F. Veerkamp[b] and Fred A. van Eeuwijk[a]

[a] *Biometris, Wageningen UR, The Netherlands*
[b] *Animal Breeding and Genomics Centre, Wageningen UR, The Netherlands*
[c] *Hendrix Genetics Research, Technology & Services B.V., The Netherlands*
[d] *Wageningen UR Plant Breeding, Wageningen UR, The Netherlands*

rianne.vanbinsbergen@wur.nl

The availability of the *Solanum Lycopersicum* 'Heinz 1706' reference genome (The Tomato Genome Consortium (2012)), together with rapid development of next generation sequencing techniques provides opportunities for modern tomato breeding. A cost effective approach would be to use 'old' SNP array genotyping data, together with genotype imputation to obtain sequence data information for large segregating populations. Building on these developments, the first objective of our study was to perform genotype imputation and asses the accuracy in a tomato RIL population. A second objective was to investigate the added value of (imputed) whole-genome sequence data relative to a SNP array when performing QTL linkage analysis.
A set of 51 RILs (F8) was available from a cross between *S. lycopersicum* (cv. Money maker) and *S. pimpinellifolium* G1.1554, which were sequenced at low-coverage (Viquez-Zamora et al. (2014)), with a mean sequencing depth of 7.3. For both parents sequence data were also available. All sequence genotypes of the RIL population were masked, except for the positions that were present on a custom made SNP array (Víquez-Zamora et al. (2013)). These masked SNPs were imputed using PlantImpute (Hickey et al. (2015)). Imputation accuracy was assessed as the proportion of SNP genotypes imputed correctly. QTL linkage analyses were performed per SNP using an independent two-sample t-test, assuming equal variances. The studied traits were average fruit weight and soluble solid content (brix).
Based on regression coefficients between SNP positions on the physical and genetic maps, each chromosome was divided into three regions: left- and right arm, and centromere. The sequence data consisted of 2,787,027 SNPs with 83.4% located in the centromere region. This high percentage is in contrast to the SNP array, which consisted of 1663 SNPs with 34.9% in the centromere region. Due to computational issues, we were only able to run PlantImpute for the chromosome arms. The number of SNPs per arm ranged between 7,002 and 46,630 SNPs on sequence data, and between 5 and 166 on SNP array. The largest arms were split in two runs of maximal 25,000 sequenced SNPs. Per run between 24.5 % and 97.4% of the genotypes (across all SNPs and individuals) were imputed, with an average proportion imputed correctly between 0.879 and 0.999.
Linkage analyses showed several significant QTL regions ($P < 0.01$) that were consistent across the three data sources (SNP array data, imputed sequence data, and true sequence data). In addition, QTL regions were detected when using (imputed) sequence data, which were not found with SNP array data. Although the number of observations is very low (~50), the imputation and QTL results suggest that (imputed) low coverage sequence data could have additional value to modern plant breeding.

Hickey, J. M., Gorjanc, G., Varshney, R. K. et al. (2015). *Crop Sci* 55.
The Tomato Genome Consortium. (2012). *Nature* 485: 635-641.
Viquez-Zamora, M., Caro, M., Finkers, R. et al. (2014). *BMC Genomics* 15: 1152.
Víquez-Zamora, M., Vosman, B., Geest, H. v. d. et al. (2013). *BMC Genomics* 14: 1-13.