דו"ח מסכם לתכנית מחקר מספר 20-10-0071

שנות המחקר 2016-2019

גישה מערכתית למיפוי תכונות כמותיות בטיפוח: שימוש בנתוני ריצוף מתקדם לצורך בניית סכימה לטיפוח אינטגרטיבי מבוסס סמנים לשיפור איכות הפרי ומרכיבי היבול במלון

A system-level perspective of QTL analysis: From raw-next generation sequence data to an integrated marker-assisted breeding scheme for improved fruit quality and yield related traits in melon.

מוגש לקרן המדען הראשי משרד החקלאות ע"י:

shiri@volcani.agri.gov.il	Shiri Freilich	שירי פרייליך
amitgur@volcani.agri.gov.il	Amit Gur	עמית גור
adif@volcani.agri.gov.il	Adi Faigenboim-Doron	עדי פייגנבוים-דורון
burgery@agri.gov.il	Joseph Burger	יוסי בורגר
tadmory@agri.gov.il	Yaakov Tadmor	יעקב תדמור
twefraim@volcani.agri.gov.il	Efraim Lewinshon	אפריים לווינסון
vcaris@volcani.agri.gov.il	Arthur Schaffer	ארתור שפר
portnoyv@volcani.agri.gov.il	Vitaly Portnoy	ויטלי פורטנוי
galilt@volcani.agri.gov.il	Galil Tzuri	גליל צורי
lahav008@gmail.com	Tamar Lahav	תמר להב
meira@volcani.agri.gov.il	Ayala Meir	איילה מאיר
fabianba@volcani.agri.gov.il	Fabian Baomkoler	פאביאן באומקלר
yarden@volcani.agri.gov.il	Uzi Sa'ar	עוזי סער

המכון למדעי הצמח, מנהל המחקר החקלאי The institute of Plant Science, Agricultural Research Organization

Introduction

Next generation sequencing (NGS) technologies have been revolutionizing crop genomics by promoting the efficient and detailed characterization of genetic variation. Genotyping by sequencing (GBS) and other NGS-based methods promoted the ability to carry out high resolution mapping studies in crop plants (Huang, et al., 2009; Liu, et al., 2014; Spindel, et al., 2013). Platforms allowing processing of raw sequence data, construction of high resolution genetic mapping and an integrative cross-experiment interpretation of results are a natural necessity for the comprehensive exploitation of the full potential derived from the new types of data. To date, though independent solutions are available for each step in the route, a complete bioinformatic pipeline is currently not available for geneticists and breeders for these purposes. The assembly of next generation breeding scheme is to a large–extent dependent on the establishment of multidisciplinary teams covering breeding, genetics, bioinformatics and systems biology aspects.

In the Israeli Agricultural Research Organization (ARO), high-throughput technologies are being applied by the cucurbit group for studying the genetic basis of diverse fruit-quality traits in melon using segregating mapping populations derived from diverse lines. A comprehensive work was performed on Recombinant Inbred Lines (RILs) population produced from a cross between PI414723 (*C. melo momordica* group) and an American muskmelon, 'Dulce' (*C. melo reticulatus* group). Multiple fruit-related traits were mapped in the past using this population (Diaz, et al., 2011; Harel-Beja, et al., 2010) including the recent identification of causative gene for fruit acidity QTL (Cohen, et al., 2014). This population was recently subjected to genome-wide genotyping-by-sequencing of coding sequences (i.e. RNA-Seq) that resulted in an extremely high-resolution mapping and re-mapping of QTLs for multiple fruit-quality traits. Despite the long standing experience of the cucurbit group in QTL analysis, conventional tools were insufficient for coping with volumes of next generation sequence data. Hence recent GBS-based projects were carried through out-sourcing to commercial companies (Galpaz, et al., 2018), significantly increasing the costs of each project and impeding the conductance of multiple-parameter iterations that are in many cases essential for the fine tuning of the analysis. In the current research we construct a pipeline that tackle current challenges in high-resolution mapping aiming at covering all aspects discussed above. The pipeline is applied for high-resolution mapping of the PI X Dulce population, used as a training set, as well as for additional three mapping populations. An

integrative network approach for delineating interactions between genes and traits was applied in attempt to understand the level of genetic dependency between key agronomic traits.

Objectives of the research

1. Establishment and implementation of a complete GBS-based pipeline for genetic mapping.
2. Integrated trait analyses based on (i) cross-experiment data mining, (ii) analyses of higher order genetic interactions, and (iii) functional contextualization of interaction networks.
3. High resolution QTL analysis of fruit quality and yield-related traits in melon populations.
4. Delineating cellular processes involved in shaping selected traits based on the integrated analyses.
5. QTL-based breeding for improved flavor, shape, color intensity and other complex fruit quality and yield-related traits in melon based on results from stage 3.
6. Mediating the use of the software and data resulted from the analyses for public use and breeding practice, through the construction of community tools.

An account of the research carried & discussion of scientific efforts and achievements

The research have led to completion of the construction of the pipeline for NGS-based genetic map building. The pipeline was applied for the production of genetic maps of five melon populations. In parallel, extensive phenotypic data was collected for two segregating populations (TADxDUL RILs and NAxDUL F3:F4). QTL detection based on the collected phenotypic data and map construction was carried and pipeline additions were integrated into the protocol. An integrative network approach was applied for dissecting the genetic landscape controlling phenotyped traits.

1. <u>Completion of a pipeline for construction of physical and linkage-based genetic maps based on NGS</u>

As described in previous reports, a pipeline aiming to tackle key challenges in construction of genetic maps based on NGS data was defined, constructed and applied for five mapping populations. Read processing, filtration, alignment to the reference genome and identification of variants at different sites was performed using standard tools including Bowtie, BWA, GATK and TASSEL-GBS V2 pipeline (Glaubitz, et al., 2014; Langmead, et al., 2009; Li and Durbin, 2009; McKenna, et al., 2010). The product is a standard variant call format or VCF (Danecek, et al., 2011). Imputation was done using the full-sib families LD algorithm (Swarts K, et al., 2014) under different parameters. The published melon reference genome (Garcia-Mas, et al., 2012) serves as the reference for marker positioning and grouping them into bins representing recombination blocks. Before binning, data processing included imputing missing data, filtering for segregation distortion and excess of heterozygosity across lines and sites. Binning was then carried using an in house script implementing the procedure outlined in (Chang, et al., 2017) and SNPbinner (Gao, et al., 2019). Linkage maps were constructed using R/qtl, ASMap and Joinmap (Broman and Sen, 2009; Stam, 1993; Taylor and Taylor, 2016). The linkage approach required stringent filtering of missing data and segregation distortion. A set of robust markers derived from the linkage analysis served as anchors and the physical dependence between them and the remaining (non-anchor) markers were tested using a Fisher exact test applying concepts described in (Huttley, et al., 1999; Yu, et al., 2001). Markers that fail the test were removed from the SNP panel. For this integrated approach we developed an in-house script that makes use of the anchor markers from the linkage map to test all available markers from the initial dataset.

The pipeline was initially applied (over the first year) for processing data derived from the PIxDUL RILs population (F_7, RNA-seq) that was previously analyzed by NRGene (Galpaz, et al., 2018) and served here as a training set for the calibration of the in-house pipeline. We carried preliminary calibration of several parameters such as, initial proportion of missing genotypic data, imputed versus non-imputed data, and different imputation parameters to characterize their effect on final number of SNPs, bins, and population size, considering that the final data set must comply with certain constraints such as allelic frequency, heterozygosity and final proportion of missing data. Allowed fraction of missing data per site was calibrated according to saturation point in terms of recombination event detection.

Mapping procedure - both the sequence-based and linkage approaches - were applied for five segregating populations (Table 1), which allows comparing the rate of recombination events within and across the different populations. Below we discuss in detail map construction for the TAMXDUL population.

2. Application of the pipeline to TADXDUL RIL populations

We examined the significance of complementary approaches (sequence-based and genetic) to construct recombination maps based on the genotypic data for the TADxDUL RIL mapping population. The sequence-based approach, as the name suggests, relies on the reference sequence to determine the marker order and map structure while the linkage-based approach is independent of the physical sequence in determining the marker order. As a result, the sequence-based method is highly susceptible to map inaccuracies due to structural variation or assembly errors that the linkage-based method allows to account for. On the other hand, linkage-based mapping results are specific to their mapping population making it harder to employ comparative mapping between populations. The mapping population is derived from a cross of a non-climacteric Honey Dew line (TAD) and a climacteric reticulatus line (DUL). The population is composed of 166 RILs (F7) that display wide uncharacterized variation in fruit traits and is therefore a key population in this project. The population was genotyped using GBS approach (Elshire, et al., 2011) and according to the pipeline that was implemented on the training set. Briefly, DNA from 164 F7 lines was extracted and shipped to Novogene, China for GBS. DNA libraries were constructed and sequenced on an Illumina HiSeq 2000/2500 platform (144 bp, paired-end reads) and mapped to the C.melo reference genome, DHL92 v3.5.1 (available at https://melonomics.net/fles/Genome/Melon_genome_v3.5.1/). Over 570 million reads were produced covering nearly 21% of the genome across more than 35 million tags at an average read depth of 9 reads per site. SNP calling was carried out using Broad Institute's genome analysis toolkit (GATK) following best practices pipeline (McKenna, et al., 2010) resulting in 1,205,528 SNPs. Sites with a depth of less than three reads per site or more than 50 percent missing data were filtered using TASSEL v5.2.43 (Bradbury, et al., 2007), then imputed using full-sib families LD algorithm (Swarts K, 2014) followed by the removal of individuals with excess heterozygosity and finally phased to ABH format consisting of 89,343 SNPs across 146

lines used to construct the sequence-based map. SNP filtration, imputation and phasing was done with TASSEL v.5.2.43 (Bradbury et al. 2007).

Sequence-based recombination map: SNPs were grouped to their respective recombination bins as detected in the sequence using SNPbinner (Gao, et al., 2019) with a minimum ratio between crosspoints set at 0.001 and minimum bin size of 1000 bp. Bin statistics and genetic distance were calculated using in-house script developed in python, based on the Kosambi mapping function (KOSAMBI, 1943). The final set included 2,853 bins across 146 lines, representing recombination events, and were used to build a 3,563 cM sequence-based genetic map.

Linkage map: a parallel linkage based genetic map was constructed using the ASMap R package (Taylor and Taylor 2016). Prior to linkage analysis, the initial raw SNPs were filtered for minimum depth of three reads per site, maximum of 20% missing data per site, a minimum of 20% minor allele frequency (MAF) and transformed to ABH format - nucleotides are masked according to the parental allele and only polymorphic sites between parents are kept) resulting in 65,564 markers across 161 lines. Since the FSFHap imputation algorithm is partially dependent on the physical marker order, we preferred to avoid any bias that this step might introduce and continued the linkage analysis using non-imputed SNP set with r/qtl - ASMap package (Broman et al. 2003; Taylor and Butler 2017). Further filtering for less than 10% missing data and exclusion of markers showing segregation distortion, resulted in 30,526 SNPs. This set was further narrowed to 10,357 SNPs after removing markers that showed co-segregation. These markers were then clustered, using a 30 cM threshold, into 12 major and 5 minor linkage groups. Next step was scanning for genotypes with inflated recombination rates, as number of crossover and double crossover events exceeding the expected theoretical median value of 14 per linkage group – these genotypes were masked as missing data within their respective linkage groups, resulting in a total of 20 lines removed from the analysis. Excess heterozygosity in lines was analyzed separately for each linkage group, removing on average 20 lines from each linkage group. After the removal of the aberrant lines, marker order was calculated again and co-segregating markers removed leaving 1,626 skeletal markers, representing recombination bins across 17 linkage groups and total genetic map of 2,187 cM. Minor linkage groups were then merged into the major LGs, based on their physical chromosomal affiliation,

adding 105 cM to the total map length. The final genetic map is 2,292 cM long, with 12 linkage groups and coveres 95% of the published reference genome used in this study, v3.5.1 and 90% of the recently published version, v3.6.1.

Comparison and evaluation: Alignment of the constructed genetic map vs. the reference genome allows to identify mismatches representing potential assembly error in the genome, or alternatively structural variations (Figure 1). A total of 29 bins were in conflict between chromosomes and linkage groups, most of them from chromosome 5. For example, two neighbouring recombination bins situated on the proximal position of chromosome 5, showed up on distal positions of linkage groups 1 and 8, most likely due to structural variations or assembly errors in the reference genome. Some of these events in chromosome 8 are highlighted in Figure 1. While the average genome-wide recombination rate was 10.3 cM/Mb - within the chromosomes the results varied greatly (Figure 1). While on chromosomes 9 and 11 they were as low as 15-20 cM/Mb, on chromosomes 2, 3, 5 and 6 they were double. Recombination events rate dropped to almost zero around the centromeres and increased toward the telomeres on all chromosomes. To evaluate the accuracy of the maps and the efficiency of the TAD x DUL RILs population for high-resolution trait mapping, we mapped to both the sequence and linkage-based maps two traits that segregate in this population and for which the causative genes are known. Flesh color in melon is determined by the *CmOr* gene (MELO3C005449), located on chromosome 9 (Tzuri et al. 2015). We phenotyped the RILs for mature fruit flesh color (orange or green that segregate at 1:1 ratio) and mapped this trait. A single QTL for flesh color (linkage-based LOD = 77.1, sequence-based LOD = 139.6, r^2 = 0.99) is mapped to an interval of 241Kb surrounding the *CmOr* gene (Figure 1).

3. Phenotyping the TADXDUL RIL population

A total of 3,963 fruits from the TADXDUL population were sampled over the course of three years, averaging 10 fruits per entry per experiment. During the summers of 2016 and 2017, experiments were conducted in open fields and in the summer of 2018, the population was grown trellised in a net-house. Plants were grown in a randomized block design and extensively phenotyped to

construct a solid dataset for QTL mapping. Our focus in this project is on the dissection of wide array of fruit traits as listed in Table 2 that also provides perspective on the scale of these experiments. A single fruit per plant was harvested at maturity based on abscission in climacteric fruits, or days after anthesis, rind color and sugar content (TSS) in non-climacteric fruits. Five mature fruits per plot were weighed, photographed externally and evaluated for external netting. Netting density (NDen), visually evaluated as the proportion of several 5x5 cm rind areas covered in netting, was scored on a categorical scale from zero to nine, zero meaning no netting and nine meaning nearly 100% netting. In cases where the netting density varied across the rind, the densest area was scored. The individual fruit NDen scores were then averaged and the trait was analyzed as a quantitative genotype mean. Fruits were then cut along the longitudinal section and one half scanned for internal imaging, using a standard document scanner (Canon, Lide120) as described previously (Gur et al. 2017). Scanned images were analyzed using the Tomato Analyzer software (Rodríguez et al. 2010) for the following morphological traits: fruit area (FAr) – total area of the longitudinal fruit section in cm^2, fruit length (FLn) - length measured at mid-width in cm, fruit width - width measured at mid-height in cm and fruit shape index (FSI) - ratio of the height at mid-width to the width at mid-height. Flesh TSS was measured from juice squeezed from each fruit by refractometer (Atago PAL-1, Atago, Japan). Ethylene emission which was measured according to the method described in (Galpaz, et al., 2018).The population display extensive variation for morphological and fruit ripening traits as shown through frequency distributions in Figure 2. Broad-sense heritabilities across the population ranged between 0.65 and 0.77, implying a strong genetic component for the fruit traits measured. Most traits displayed homogeneity of variances within a normal distribution with three exceptions: (i) Fruit weight (FW) displayed a natural logarithmic distribution. (ii) Netting coverage (NCOV) displayed a bi-modal distribution, with higher frequency across the netted fruits and was later analyzed as a binary trait in addition to non-parametric methods. (iii) Netting density (NDEN) was scored in the 2016 experiment as categorical trait with only X classes, essentially aimed at discriminating between netted and smooth with less focus on the . Traits repeatability among the open field trials was between 0.77 for netting density and 0.84 for fruit width (WD). Between the open field and net-house the correlations were between 0.62 for area (AR) and 0.79 for fruit shape (FSI). TSS repeatability was much lower among all trials (open field $r = 0.56$, open field - net-house $r = 0.30$) implying this trait was more

susceptible to environmental interactions. Genotype x year interactions among open field experiments ranged between non-significant (WD) to highly significant (WT, AR, TSS) while genotype x environment interactions as calculated between open-field and net house experiments were significant for all traits.

The comprehensive phenotypic analysis allow us to learn also on the patterns of trait distribution patterns**.** Principal component analysis based on trait means across the 3 experiments resulted in four main clusters is shown in Figure 3 (top left). Principal component 1 accounted for 63.6% of the variation and was positively loaded with fruit weight, area, diameter, and length and negatively with netting coverage and density. Within this group was a clear distinction between the growing environments as fruit size and weight were significantly smaller in the net-house (2018). Principal component 2 accounted for 17.7% of the variation and was loaded negatively with fruit shape, suggesting a non-linear relationship between shape and the other morphological traits, with length being the dominant factor for this trait, based on trait correlations.

Another population that was analyzed as part of this project is NAxDUL F3: F4 that was grown in a randomized block design and phenotyped on summer 2017. NA is a non-climacteric inodorous line and this population extend the variation for fruit and ripening behaviour traits.

4. Mapping QTLs in the phenotyped TADxDUL population

The genotypic and phenotypic data collected on the TADxDUL RILs (described above) was used to map QTLs for multiple traits. Genome-wide linkage analysis of the traits on the sequence based bin map was performed in TASSEL ver. 5.2.51 (Bradbury et al. 2007) using a generalized linear model (GLM) with 1000 permutations and p. value of 0.05 as threshold. Stepwise interval mapping on the linkage based map was performed with R/qtl. The method used was Haley–Knott regression, except for non-parametric traits, with 1000 permutations and *p*.value of 0.05 as detection threshold. The search was limited to ten QTLs and the confidence interval was calculated based on 1.5 LOD scores. QTL interactions were confirmed using a two-way ANOVA in JMP ver. 13.1. QTLs were mapped for each year separately. For fruit area, weight, length, width and fruit shape, overall 46 significant QTLs were detected on

eight of the twelve melon chromosomes. Robust QTLs – that is QTLs that appear in at least two years are listed in Table 3, explaining 8-33% of the variance of the respective traits. In order to better explain the phenotypic diversity by genetic terms we systematically examined the percentage of the variance that can be explained by multi-allelic models. To this end we used stepwise mapping models that screen for QTLs with secondary effects following masking the effect of the primary QTLs. In some of the cases, combined models, considering the effect of both primary and secondary QTLs can better explain the variance. For example, stepwise mapping identified four significant QTLs associated with fruit size (AR, single year, Figure 4). A joint model, based on the four QTLs better explains the phenotypic variance in comparison to a single gene model based solely on the strongest QTL (Figure 4). Finally, the four significant markers were systematically screened *vs.* all markers in the map in order to detect hidden markers - (markers that by themselves are not significantly associated with traits). An example for such hidden marker revealed by whole-marker screen is shown in Figure 5, pointing at a marker whose effect on the phenotype depends on the genetic background.

5. Network based integrative view for phenotypic and genotypic landscapes

Typically, results from association studies provide a list of discrete candidate loci per trait (Table 3). Complexity of phenotypes can be reduced by the grouping of traits according to co-appearance patterns in a monitored collection (for example in Figure 3), suggesting common genetic control and inheritance. As illustrated in Figure 3 (top left), comparative view of the phenotypic vs. genotypic landscape further allows to dissect the molecular mechanism that determine examine whether phenotypic dependence/independence is associated with genetic determinants. The network can be crossed with the phenotypic landscape in order to identify pleotropic effect of loci and/or traits that are controlled by independent genetic elements. Such view can serve the development of marker assisted selection programs as loci can be ranked not only according to the strength of effect on a single trait but according to the pleiotropy of their effect on related and unrelated phenotypes. In cases of trade-offs between traits, integrative view can lead to the selection of markers that best balance between selection programs. For example, in cases of two QTLs with a similar effect on a single phenotype, integrative view can disqualify one of the loci based on its pleiotropic effects (desired and

undesired). Similarly, integrative view can point at independent phenotypic groups and direct to selection programmers that are based on selection of multiple markers with non-redundant contribution to the phenotype. Here, we applied to integrative view for examining the genetic landscape controlling fruit size and harvesting time. The TAD parent is a slow riper with large fruits whereas DUL is a fast riper with small fruits. The phenotypic vs genotypic landscape controlling these traits is shown in Figure 3 (bottom left). The integrative view indicates that fruit size and ripening time traits display independence in their phenotypic behavior and the main QTLs associated with the traits are different, pointing at the possibility of identifying QTLs that dissect these traits. Notably, significant QTLs for both traits display parental patterns hence election only for size might preserve parental patterns and can be associated with long ripening time. In order to select for QTL/s leading to maximal fruit size and minimal ripening length we systematically screened for the best tradeoff between these traits by crossing values measures for these traits across all pairwise allelic combinations, leading to prediction of potential markers.

6. Application of gentotypic pipeline for the development of advanced breeding schemes

One of the scopes of this project is the implementation of tools and concepts developed as part of the project into next generation breeding schemes. High resolution mapping allowing such implementation was achieved only during the last year of the project. The genetic infrastructure achieved as part of the project had served so far few projects and continues serving the current breeding efforts of the cucurbit group in Newe Yaar. Key focuses of the applicative directions of the research are highlighted below,

Mapping *APRR2 -a* gene that is associated with fruit pigment accumulation in melon and watermelon: The high resolution linkage and physical maps constructed for the TADXDUL mapping population were used for the identification of a QTL associated woth a gene that controls pigment accumulation in the young melon fruit. The high resolution maps allowed identification of a relatively narrow QTL with 300kb containing 30 genes. The mapping further allowed identification and isolation of *APRR2* - the gene controlling the trait. Markers within this genes allows an exact phenotypic characterization of the allele ('dark' or 'bright') at a range of genetic

backgrounds. A paper describing this work was recently published (Oren, et al., 2019). As a continuation of this work, we develop new mapping population by crossing parents with different alleles, considering different genetic backgrounds.

Development of MAS breeding scheme for rind netting: Netting is a trait of a high commercial value, yet despite its importance to breeders and the high phenotypic diversity of the traits, controlling genes were not characterized and identified. The TAD x DUL mapping population was extensively characterized for the trait together with other fruit shape associated traits along the three years of the project (described in point 3 above). Traits were mapped to the high-density sequence- and linkage-based recombination maps (described in point 4 above). Four QTLs associated with netting were identified and an additive interaction between two QTL were characterized. These QTLs now serve the development of PCR markers that will support MAS breeding for netted vs non-netted fruits. MAS, using these markers is also tested in different genetic backgrounds. A paper describing this work is being written these days and is aimed to be soon submitted.

Development of mapping populations for ripening: TAMXDUL population has a non-climacteric and a climacteric parent, respectively. Hence, the population serves the mapping of ripening associated traits allowing exploring the genetic landscape of the trait. Markers are being developed based on the analysis. A manuscript describing this work is being written and is expected to be completed in a few months. Similar work is carried for earliness, another traits with high commercial relevance.

7. Construction of multi-allelic phenomic and genomic database in melon and design of a website for data integration

One of the scopes of this project is the implementation of tools and concepts to leverage multi-allelic data for trait mapping. In addition to the populations that were phenotyped as part of the project, as described above, we initiated a process of collection and standardization of data from additional populations that were developed at Newe-Yaar and phenotyped earlier, as part of other projects. For all these populations, genome-wide genotypic data is available and will be used for mapping. The multi-allelic database

structure and nomenclature that was developed and will provide a base for the further development of a solid data repository for future projects.

Summary

In the current project we have developed a pipeline for the construction of high resolution maps based on NGS data, process that were previously handed to external support. Map construction were carried in parallel by using sequence and genetic information. Maps were crossed and used for the identification of structural variations and assembly errors. The pipeline was applied for a careful construction of maps for the TAMXDUL that was extensively phenotypes for several key fruit quality, harvesting time and fruit size traits. The high-density genotyping enabled high-resolution whole-genome mapping of complex traits. Beyond the identification of discrete markers associated with trait, multiple locus-trait interaction pairs were integrated to describe to promote the understanding of the network of interactions that determine the phenotypic landscape. Data and approaches that were produced as part of this work aimed at improving the quality and resolution agricultural trait mapping and simplify the presentation of multi-faceted data. The framework developed as part of the project served and is serving breeding efforts in the cucurbit group in Newe Yaar.

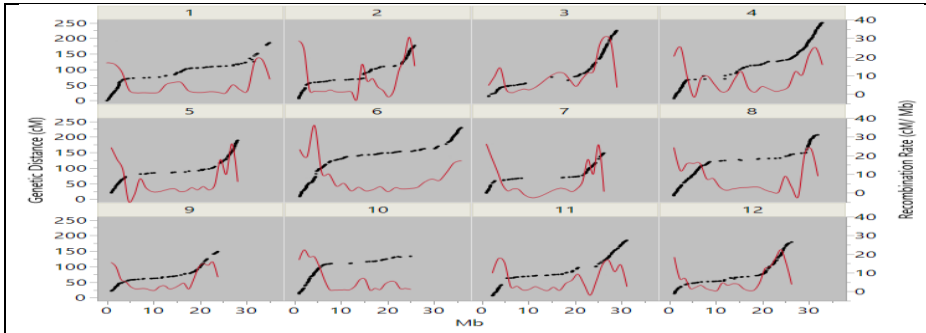
The work carried as part of this project is summarized in three papers, including a paper that was just accepted for publishing (after the submission of the original report). The publication - Oren *et al* 2019 (Oren, et al., 2019), describes an application of the pipeline for high resolution mapping and is enclosed as an additional file. Additional two papers based on this research are currently being written and will be submitted this coming year. Tentative titles are:

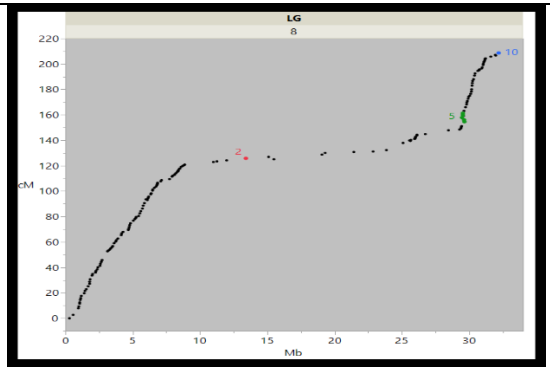
1. High-density map construction and genetic dissection of fruit size, shape and rind netting in *Cucumis melo* This publication will include a detailed description and discussion of the pipeline. All in house scripts that were developed and used for the project will become publically available as part of the publication.
2. Genetic characterization of fruit maturity and ripening behavior in melon using multi- and bi-allelic populations. This publication will discuss pre-breeding application of the computational platform.

References

Bradbury, P.J.*, et al.* (2007) TASSEL: software for association mapping of complex traits in diverse samples, *Bioinformatics*, **23**, 2633-2635.
Broman, K.W. and Sen, S.A. (2009) Guide to QTL Mapping with R/qtl.
Chang, C.W., Wang, Y.H. and Tung, C.W. (2017) Genome-Wide Single Nucleotide Polymorphism Discovery and the Construction of a High-Density Genetic Map for Melon (Cucumis melo L.) Using Genotyping-by-Sequencing, *Frontiers in plant science*, **8**, 125.
Cohen, S.*, et al.* (2014) The PH gene determines fruit acidity and contributes to the evolution of sweet melons, *Nat Commun*, **5**, 4026.
Danecek, P.*, et al.* (2011) The variant call format and VCFtools, *Bioinformatics*, **27**, 2156-2158.
Diaz, A.*, et al.* (2011) A consensus linkage map for molecular markers and quantitative trait loci associated with economically important traits in melon (*Cucumis melo L.*), *BMC plant biology*, **11**, 111.
Elshire, R.J.*, et al.* (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species, *PloS one*, **6**, e19379.
Galpaz, N.*, et al.* (2018) Deciphering genetic factors that determine melon fruit-quality traits using RNA-Seq-based high-resolution QTL and eQTL mapping, *The Plant journal : for cell and molecular biology*, **94**, 169-191.
Gao, L.*, et al.* (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor, *Nature genetics*.
Garcia-Mas, J.*, et al.* (2012) The genome of melon (*Cucumis melo L.*), *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 11872-11877.
Glaubitz, J.C.*, et al.* (2014) TASSEL-GBS: a high capacity genotyping by sequencing analysis pipeline, *PloS one*, **9**, e90346.
Harel-Beja, R.*, et al.* (2010) A genetic map of melon highly enriched with fruit quality QTLs and EST markers, including sugar and carotenoid metabolism genes, *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik*, **121**, 511-533.
Huang, X.*, et al.* (2009) High-throughput genotyping by whole-genome resequencing, *Genome Res*, **19**, 1068-1076.
Huttley, G.*, et al.* (1999) A scan for linkage disequilibrium across the human genome. , *Genetics*, **152**, 1711–1722.

Langmead, B.*, et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome, *Genome Biology*, **10**, R25.
Li, H. and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform, *Bioinformatics*, **25**, 1754-1760.
Liu, H.*, et al.* (2014) An evaluation of genotyping by sequencing (GBS) to map the Breviaristatum-e (ari-e) locus in cultivated barley, *BMC Genomics*, **15**, 104.
McKenna, A.*, et al.* (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data, *Genome research*, **20**, 1297-1303.
Oren, E.*, et al.* (2019) The multi-allelic APRR2 gene is associated with fruit pigment accumulation in melon and watermelon, *Journal of experimental botany*, **70**, 3781-3794.
Spindel, J.*, et al.* (2013) Bridging the genotyping gap: using genotyping by sequencing (GBS) to add high-density SNP markers and new value to traditional bi-parental mapping and breeding populations, *Theor Appl Genet*, **126**, 2699-2716.
Stam, P. (1993) Construction of integrated genetic linkage maps by means of a new computer package: Join Map. , *The Plant journal : for cell and molecular biology*, **3**, 739–744.
Swarts K*, et al.* (2014) Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. , *Plant Genome* **7**.
Swarts K, L.H., Romero Navarro JA, An D, Romay MC, Hearne S, Acharya C, Glaubitz JC, Mitchell S, Elshire RJ, et al (2014) (2014).
Taylor, J. and Taylor, M. (2016) Package "ASMap." **38**.
Yu, A.*, et al.* (2001) Comparison of human genetic and sequence-based physical maps, *Nature*, **409**, 951-953.





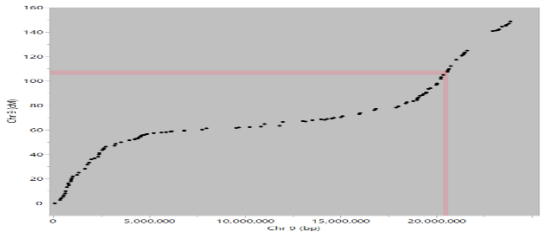


Figure 1. Top right panel - Marey map. In black a comparison between genetic and physical distance. In red chromosome-wide recombination rates. Top left panel - Re-localization of markers from reference genome sequence to chromosome 8, based on linkage analysis. The number next to the colored dots denominates their original chromosome. Bottom left panel - physical and genetic QTL intervals of CmOr QTL.

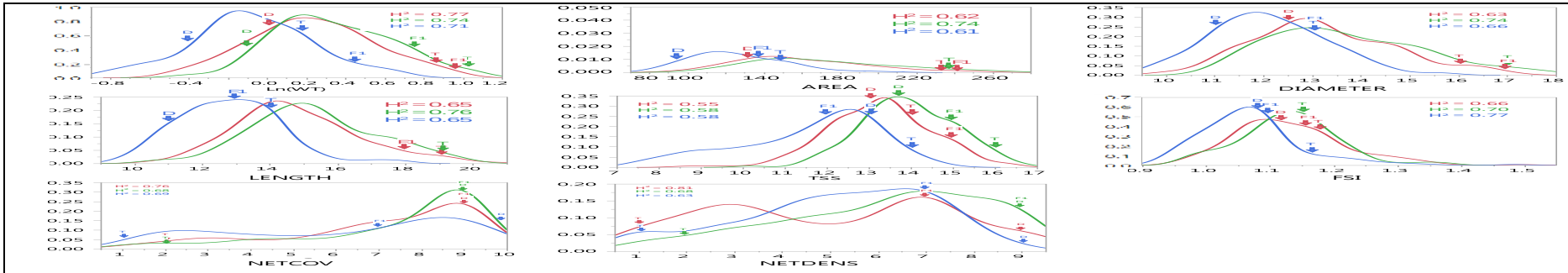


Figure 2. Distributions and heritability of morphological in TAD x DUL population across years. Arrows mark parental and hybrid scores.

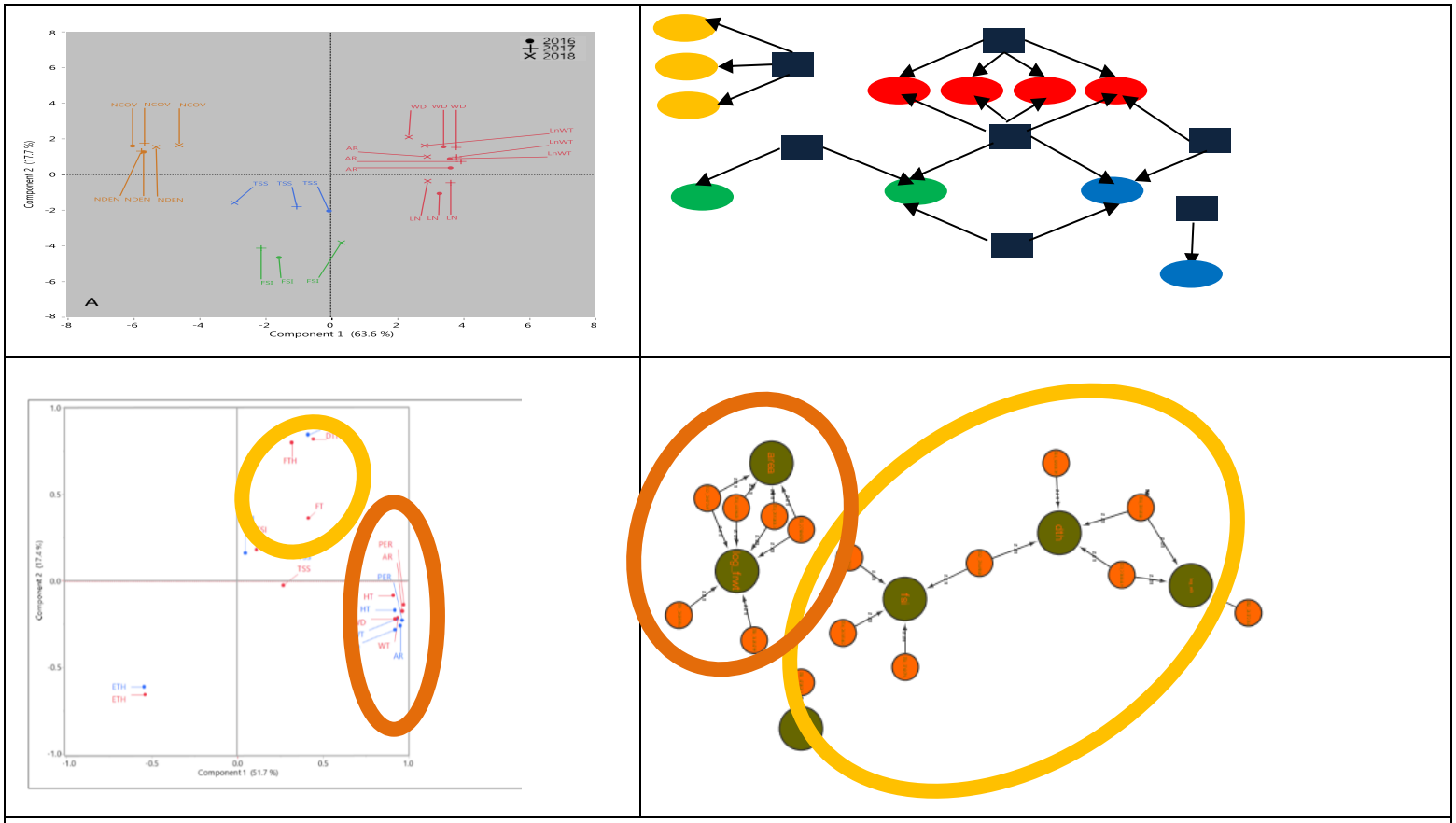


Figure 3. Top: Phenotypic vs genetic landscape. Top left: Phenotypic diversity of traits in Figure 2 projected on a principal component plot. Four groups of traits are indicated by text and arrow color. Right: an illustration of trait-locus network. Traits are indicated by colored circles according to group determined by distribution similarity (left), rectanges represent QTLs. Bottom: PCA of fruit size (width, height, area, perimeter, Length, diameter – circled in orange) and ripening time (days to harvest, days to flowering, flowering time to harvest, ethylene, circled in yellow) traits phenotyoed in the TADXDUL population on the left vs. QTL-trrait network om the right. Loci are shown in orange; traits in green. Mapping is based on preliminary results based on the previous draft of the map.

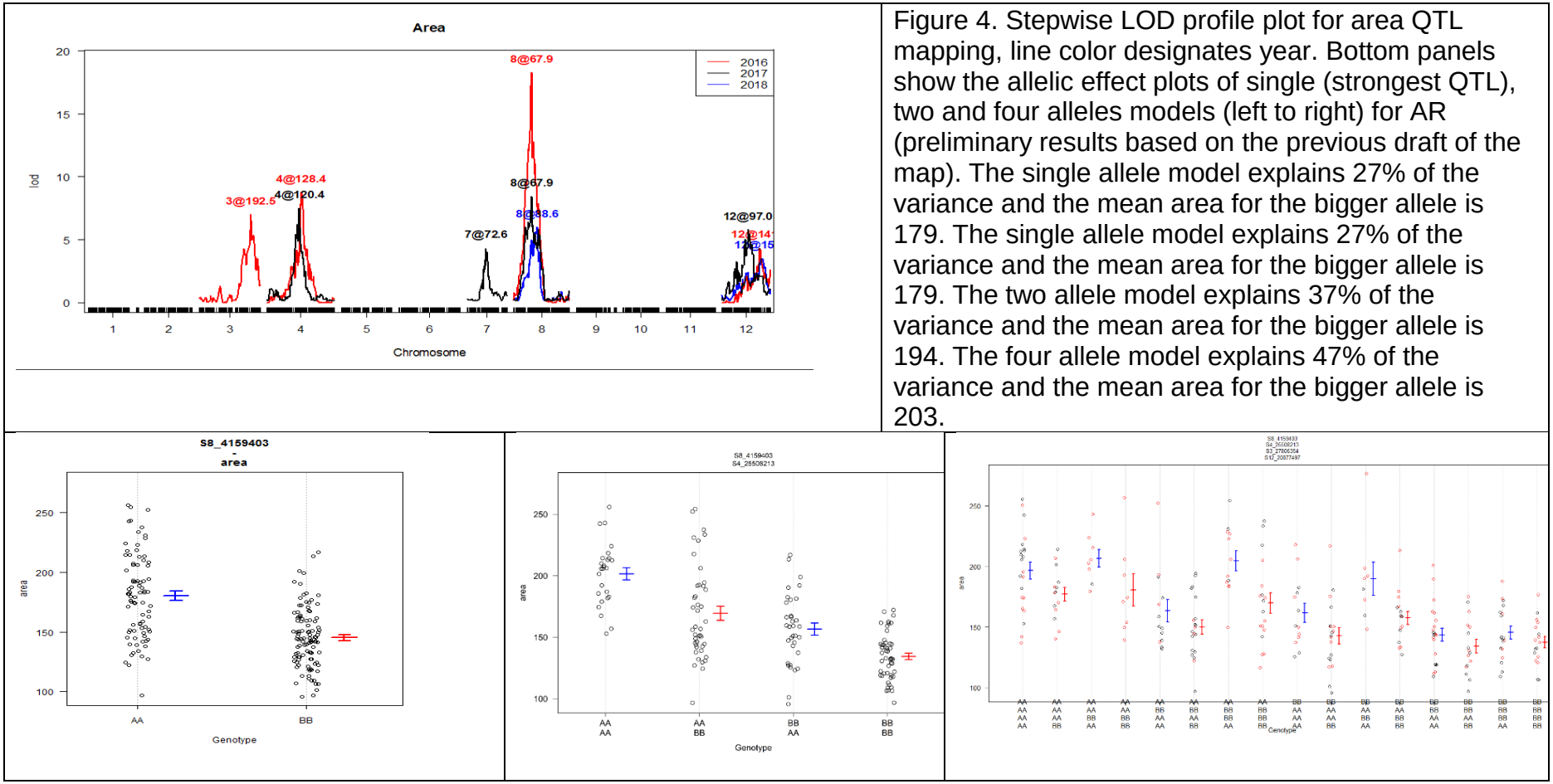
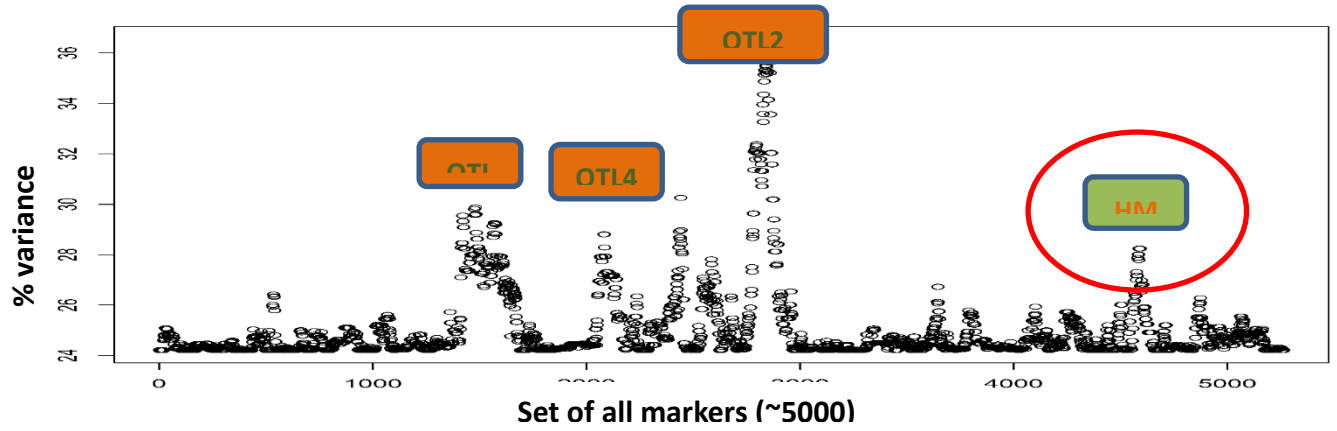
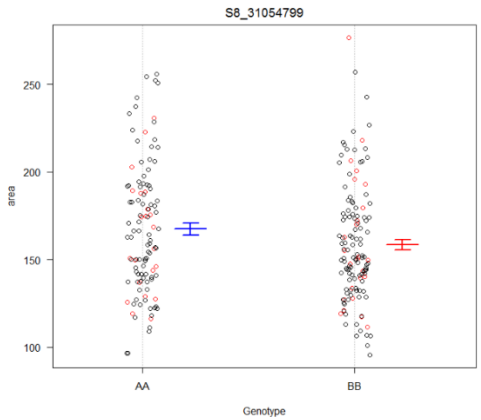


Figure 4. Stepwise LOD profile plot for area QTL mapping, line color designates year. Bottom panels show the allelic effect plots of single (strongest QTL), two and four alleles models (left to right) for AR (preliminary results based on the previous draft of the map). The single allele model explains 27% of the variance and the mean area for the bigger allele is 179. The single allele model explains 27% of the variance and the mean area for the bigger allele is 179. The two allele model explains 37% of the variance and the mean area for the bigger allele is 194. The four allele model explains 47% of the variance and the mean area for the bigger allele is 203.





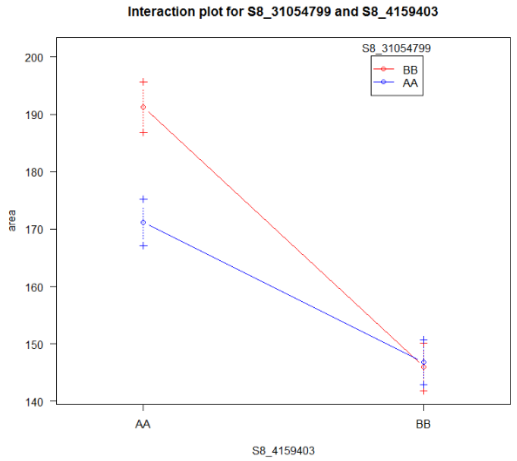


Figure 5. An example for a systematic screen for epistatic interactions QTL (strongest QTL for AR, Figure 4 and Table 3) and all markers in the genetic map (preliminary results based on the previous draft of the map). Y axis indicates the % of variance explained by epistatic interactions where basal line indicates the % explained by the main QTL itself. The three QTLs identified by the stepwise analysis (Figure 4) are show in orange. New markers are shown in green, where the effect of the marker circled (in red) is shown in the lower panels: individually (left) it has no effect but it has epistatic interaction with the main QTL (left).

Table 1. Summary of populations used in comparative genome wide recombination event analysis.

Population	Population Size	Genetic distance between parents	Initial Proportion of NA	Fraction of sites with less than 40% NA	Proportion of NA after imputation	Fraction of sites with less than 40% NA	Number of SNPs	Number of Bins
PI414xDUL F_{6-8}	98	0.40	0.52	41%	0.01	99.9%	16,524	1408
TADxDUL F_7	164	0.38	0.20	10%	0.03	98%	92,584	2853
DULxTOG F_3	121	0.50	0.44	33%	0.09	89%	31,039	2287
NAxDUL F_3	140	0.31	0.45	34%	0.13	85%	24,784	2623
SASxDOYA F_3	120	0.53	0.50	24%	0.09	88%	36,144	2234
SASxTOG F_3	119	0.10	0.30	62%	0.12	88%	499	-

Table 2. Summary of the traits phenotyped in the mapping population.

Trait name	Abbreviation	Units	Description
Ethylene emission	ETH	ln(ppm/kg*hr)	Ethylene emission of ripe fruits as described in methods
Fruit weight	WT	ln(kg)	Average fresh weight of fruits from plot
Total soluble sugars	TSS	° Brix	Total soluble sugars measured from ripe fruit flesh
Days to harvest	DTH	days	Time from sowing to harvest
Days to flowering	FT	days	Time from sowing to anthesis
Flowering time to harvest	FTH	days	Time from anthesis to harvest
Fruit area	AR	cm^2	Total area of the longitudinal fruit section
Fruit perimeter	PER	cm	Perimeter measured at mid-width
Fruit length	HT	cm	Height measured at mid-width

Fruit diameter	WD	cm	Width measured at mid-height
Fruit shape index	FSI	-\	Ratio of the height at mid-width to the width at mid-height

Table 3. Significant QTLs from at least two years, based on stepwise interval mapping

QTL[a] name	Trait	Chr	LOD[b]	GeneticQTL peak position (cM)	Genetic QTL confidence interval (cM)[c]	Physical QTL position (bp)	Physical QTL confidence interval (bp)[d]	% Var explained[e]	Additive effect[f]
AR4.2	AR	4	8.8	128.45	3.5	24,034,321	4,848,992	10.9	-13.98
AR8.1	AR	8	18.3	67.95	5.3	4,281,821	772,512	14	-21.37
AR12.1	AR	12	4.3	38.13	10.2	24,153,178	1,149,195	33.6	-9.05
FSI1.1	FSI	1	8.4	100.08	3.1	16,617,878	1,636,332	14.3	0.02
FSI6.1	FSI	6	8.3	153.87	3.1	25,300,403	6,624,001	6.4	-0.03
FSI6.2*	FSI	6	5.5	95.42	13.5	5,224,893	1,863,470	13	-0.03
FSI8.1*	FSI	8	4	58.96	6.3	3,615,451	1,611,434	9.2	-0.03
FSI9.1*	FSI	9	4.7	85.64	14.4	19,004,589	4,893,639	10.9	0.03
FSI12.1	FSI	12	11.3	82.66	17.3	21,520,282	1,366,214	14	-0.04
LN4.1	LN	4	5.6	122.03	15.3	20,753,062	5,190,287	20.1	-0.9
LN8.3	LN	8	13.5	67.95	5.3	4,281,821	1,541,494	28	-1.04
LN9.1	LN	9	4.3	67.23	11	13,024,318	14,472,904	7.8	0.52
LN12.1*	LN	12	5.2	27.97	20.7	24,529,333	1,348,987	10.8	-0.55
NDEN2.2	NDEN	2	13.7	62.43	2.4	7,105,437	5,765,990	13.1	0.92
NDEN2.3	NDEN	2	12.9	96.68	2.6	16,471,732	1,259,903	32.9	0.97
NDEN9.1	NDEN	9	5.4	6.78	6.8	535,980	1,905,449	12.1	0.33
NDEN9.2	NDEN	9	12.8	38.05	4.9	2,324,093	517,821	12.1	0.81
WD4.1	WD	4	6.4	119.23	1.4	20,135,408	876,056	19.6	-0.75
WD8.1	WD	8	13.7	67.95	5.3	4,281,821	772,512	23.2	-0.74
WD12.2	WD	12	5.3	81.4	14.8	21,544,132	1,810,973	16	-0.67

WT3.2	WT	3	6.8	192.55	13.5	27,490,349	705,613	10.9	0.12
WT4.1	WT	4	9.9	128.45	8	24,034,321	5,701,186	7.9	-0.13
WT8.2	WT	8	16.8	67.95	2.6	4,281,821	524,021	15.9	-0.18

[a] QTL names are composed of trait abbreviation, chromosome number and QTL number; [b] Maximum LOD score for consensus QTLs; [c] Interval based on at least 1.5 LOD score drop; [d] Interval bases on flanking markers physical position; [e] Maximum R square for each QTL; [f] Positive additive effect when Tamdew alleles contribute to trait score and negative for Dulce alleles;[*] QTL is unique for one net-house