Scientia Horticulturae 246 (2019) 1–11

Contents lists available at ScienceDirect

Scientia Horticulturae

journal homepage: www.elsevier.com/locate/scihorti

Application of genomic tools to avocado (*Persea americana*) breeding: SNP discovery for genotyping and germplasm characterization

D.N. Kuhn[a,*,1], D.S. Livingstone III[b,1], J.H. Richards[c], P. Manosalva[d], N. Van den Berg[e], A.H. Chambers[f]

[a] USDA-ARS Subtropical Horticulture Research Station, 13601 Old Cutler Rd, Miami, FL, 33158, USA
[b] Genetics and Breeding, Cocoa Research & Development, MARS Wrigley Confectionery, 13601 Old Cutler Rd, Miami, FL, 33158, USA USA
[c] Department of Biological Sciences, Florida International University, Miami, FL, 33199, USA
[d] Department of Microbiology and Plant Pathology, University of California, Riverside, Riverside CA, 92521, USA
[e] Department of Biochemistry, Genetics and Microbiology, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, 0002, South Africa
[f] Department of Horticultural Sciences, Tropical Research and Education Center, University of Florida, Homestead, FL, 33031, USA

ARTICLE INFO

Keywords:
Infinium genotyping array
Synonymous SNP
Non-synonymous SNP
GO slim annotation
Marker-assisted selection

ABSTRACT

Avocado (*Persea americana*) is an important tropical and subtropical fruit tree crop. Traditional tree breeding programs face the challenges of long generation time and significant expense in land and personnel resources. Avocado selection and breeding can be more efficient and less expensive through the development of molecular markers for the estimation of germplasm genetic diversity, marker-assisted selection (MAS), and creation of linkage maps. Two important breeding resources, the world's two largest avocado mapping populations and an extensive germplasm collection, are housed at the USDA-ARS Subtropical Horticulture Research Station (SHRS) in Miami, Florida. However, to use these resources to their greatest advantage, many thousands of genetic markers are necessary. Here, we describe the development of the first set of avocado genetic markers based on single-nucleotide polymorphism (SNP) variation in expressed genes. RNA sequencing was used both to build a reference transcriptome from 'Hass', the most widely grown avocado cultivar worldwide, and to identify SNPs by alignment of RNA sequences from the mapping population parents to the 'Hass' transcriptome. This study provides a new genomic tool for the avocado community that can be used to assess the genetic diversity of avocado germplasm worldwide and to optimize avocado breeding and selection programs by complementing traditional breeding methods with molecular approaches, thus increasing the efficiency of avocado genetic improvement.

1. Introduction

Avocado (*Persea americana*) is an economically important tropical and subtropical tree fruit crop. Avocado has three major horticultural subdivisions: *Persea americana* var. *americana* Mill. 'West Indian'; var. *guatemalensis* Williams, 'Guatemalan'; and var. *drymifolia* (Schlecht. & Cham.) Blake, 'Mexican'. These subdivisions are supported by microsatellite genotyping data (Schnell et al., 2003). Avocado, like many tree crops, is propagated clonally through grafting to preserve commercially desirable varieties. In 2014, worldwide production of avocado was 5.03 million metric tons (MMT), with over 1 MMT produced in Mexico (FAO, 2016; Russell et al., 2011). In the US, which ranks seventh in worldwide avocado production, California is the largest producer with Florida and Hawaii accounting for smaller percentages of the yearly crop. Almost all the cultivated avocado acreage in California is the cultivar 'Hass', which is currently described as a Mexican x Guatemalan hybrid (Ashworth and Clegg, 2003). Avocado production is the second largest agricultural industry in Florida after citrus and is based almost exclusively in Miami-Dade county, where growers use West Indian (WI) and Guatemalan x West Indian (G x WI) hybrid varieties that are suited to local climate and soil conditions. Currently, the most widely grown Florida avocados are the varieties 'Simmonds' (WI), 'Donnie' (WI), 'Monroe' (G x WI) and 'Lula' (G x WI) on either open pollinated 'Lula' or 'Waldin' (WI) rootstock (*personal communication*, Alan Flinn, Florida Avocado Administrative Committee). To support the domestic avocado industry, the USDA-ARS Subtropical Horticulture Research Station

* Corresponding author.
E-mail addresses: David.Kuhn@ars.usda.gov (D.N. Kuhn), Don.Livingstone@effem.com (D.S. Livingstone), Richards@fiu.edu (J.H. Richards), pmanosal@ucr.edu (P. Manosalva), noelani.vandenberg@up.ac.za (N. Van den Berg), ac@ufl.edu (A.H. Chambers).
[1] These authors contributed equally to the work.

https://doi.org/10.1016/j.scienta.2018.10.011
Received 10 April 2018; Received in revised form 12 September 2018; Accepted 5 October 2018
Available online 28 October 2018

(SHRS) in Miami, Florida curates a large avocado germplasm collection. In addition, the SHRS has generated two large mapping populations to identify quantitative trait locus (QTL) controlling desirable agronomical and horticultural traits using SNPs markers with the ultimate goal to implement marker-assisted selection (MAS) in avocado breeding programs.

Biotic and abiotic stresses are important targets for avocado breeding programs in the US. The most important losses worldwide are due to Phytophthora root rot caused by *Phytophthora cinnamomi* (Ramirez-Gil et al., 2017). In southern Florida, avocado production is currently threatened by laurel wilt, which causes the rapid death of mature trees (Ayala-Silva et al., 2012; Kendra et al., 2011; Ploetz et al., 2011). While acceptable rootstocks overcome the challenges of growing in calcareous soils, other concerns in southern Florida are cold tolerance and early or late flowering, which could extend the market season (Paritosh et al., 2014). In California, salt and drought tolerance together with dwarfism are important traits that breeders aim to incorporate in elite rootstocks. To improve the efficiency of tree breeding, which currently requires ~15 years before the release of a new cultivar, thousands of genetic markers distributed throughout the genome are needed. Single nucleotide polymorphisms (SNPs) are suitable for this purpose. Previous studies have developed molecular markers for avocado, including SNPs from targeted resequencing (Chen et al., 2008) and microsatellite markers (Alcaraz and Hormaza, 2007; Ashworth et al., 2004; Borrone et al., 2007; Lavi et al., 1994; Sharon et al., 1997). These markers have been combined with phenotypic data, including heritability estimates of nutritional components (Calderon-Vazquez et al., 2013), or to identify marker trait associations in avocado (Mhameed et al., 1995; Sharon et al., 1998). Genome-wide SNPs are required to construct a high-density linkage map, identify marker-trait associations to implement MAS, and identify suitable parents from germplasm collections for breeding purposes.

To develop MAS for avocado, we created two large mapping populations. The first is 'Simmonds' x 'Tonnage' and their reciprocal cross (Florida mapping population, FLAMP) (Borrone et al., 2008), and 'Hass' x 'Bacon' and their reciprocal cross (California mapping population, CAMP) (Schnell et al., 2009). These populations are planted at USDA-ARS SHRS, Miami, FL, and phenotypic evaluation of both populations is in progress. We initially made a genetic map of FLAMP using ~130 microsatellite markers (Borrone et al., 2009). This map had 12 linkage groups representing the 12 chromosomes of avocado, but two of these linkage groups had only two to four markers. The objective of this study was to mine transcriptomes of the mapping population parents for SNPs that could be used in the future to increase the resolution of the FLAMP map, create the CAMP map, and merge the two for a consensus map. This was accomplished through the use of thousands of SNP markers since they occur in greater numbers than microsatellites, can occur in the coding region of genes, and can be assayed unambiguously on any SNP assay platform. In addition, to reduce resource costs and increase data collection rates over the microsatellite approach, we produced an Infinium II 6000 SNP chip that will be used in the future to simultaneously genotype the mapping populations and several avocado germplasm collections.

Avocado SNPs were discovered by creating a reference transcriptome from `Hass` leaves and flowers and mapping RNA sequences from the FLAMP and CAMP parents onto the reference to identify variants. SNPs that were heterozygous in one or more of the parents were chosen for the Illumina SNP chip. Here we describe: i) the assembly and analysis of a reference 'Hass' transcriptome from leaf and flower RNA sequences, ii) the identification and analysis of ~600,000 SNPs, iii) filtering of the SNPs for the design of an Infinium II 6000 SNP chip, and iv) comparison of the congruence of SNP genotypes predicted from the RNA sequencing with those scored by SNP chip for the four mapping population parents assessed.

2. Materials and methods

2.1. Avocado cultivars

The parents of the mapping populations 'Simmonds', 'Tonnage', 'Bacon', and 'Hass' were obtained as clones from commercial nurseries (Pine Island, Florida, Limonera Grove, California). Parents were genotyped with informative microsatellite markers as previously described (Schnell et al., 2003) to validate their clonal identity.

2.2. Mapping populations

To create mapping populations, we bought fruit from commercial groves where only two cultivars were being grown, one as the producer and one as the pollinator. Seeds were germinated and genotyped using a small set of informative microsatellite markers to identify hybrids and selfed progeny. The female parent is listed first by convention. For FLAMP, the 'Simmonds' x 'Tonnage' cross had 249 individuals, and the 'Tonnage' x 'Simmonds' cross had 514 individuals, for a total of 763 individuals. For CAMP, the 'Bacon' x 'Hass' cross had 230 individuals, and the 'Hass' x 'Bacon' cross had 346 individuals, for a total of 576 individuals. The grand total was 1339 individuals, making these the largest avocado mapping populations in the world.

2.3. RNA isolation, sequencing, and transcriptome assembly

Leaves, unopened flowers, female-phase flowers and male-phase flowers were collected from each parent. Material was frozen in liquid nitrogen immediately after harvest and RNA was isolated as previously described (Kuhn et al., 2012). RNAs from these four tissues were quantified by fluorimetry (QuantIT riboGreen, Invitrogen), pooled in equimolar amounts, and used to generate each parent-specific Illumina sequencing library. Parent-specific RNA-Seq libraries were synthesized following the manufacturer (Illumina) protocols, followed by partial normalization with double-stranded nuclease (DSN) using the Trimmer Direct kit (Evrogen), according to published protocols (Matvienko et al., 2013). Sequencing was performed on a GAIIx instrument as paired 72 bp reads, and data were processed using Illumina Sequencing Control Software (SCS) v1.7.1. Read ends were trimmed to 60 bp to eliminate low quality and biased nucleotide representation. Raw sequence data were deposited in NCBI databases under BioProject PRJNA258225.

'Hass' reference RNA libraries were prepared for optimal sequencing on the 454 Titanium platform as previously reported (Kuhn et al., 2012). Manufacturer (Roche/454 Sequencing) protocols were followed for emulsion PCR and sequencing on two full two-region GS-FLX Titanium PicoTitre™ plates. Reads were cleaned of adaptor sequences (http://sourceforge.net/projects/estclean/) (Ballerini et al., 2013), and sequences ≤100 bp were removed. NEWBLER v2.3 (Roche/ 454 Sequencing) with default parameters for cDNA (40 bp overlap; 90% identity) was used for assembly.

2.4. Transcriptome analyses

A condensed dataset (CD) was produced by removing outliers from the 33,957 isotigs (OD, original dataset), as follows. For each parameter (e.g., isotig length), the distribution of the parameter for all 33,957 isogroups was determined. The top 2.5% and the bottom 2.5% of the parameter values were used to filter the data. The process was repeated for each parameter, starting always from the 33,957 isogroups. The 33,957 isogroups were filtered for outliers for isotig length, average mean reads, and average covered length. Since more than 2.5% of the isogroups had no SNPs, SNP number was not part of the outlier removal. An isotig is here defined as the longest contig in an isogroup. Using the parameters identified from each filtering step, a hierarchical filtering of the data was done to generate the dataset with outliers removed.

Further subsets were made from the CD. A BLASTn search of the avocado transcriptome against itself was performed to identify putative single copy transcript/gene (PSCI) isotigs. Isotigs that had a BLASTn e-score lower than 10^{-15} to any isotig but itself were considered to be potential paralogs and were removed from the CD to produce the PSCI subset. Avocado genes that are homologs of single copy genes conserved in Arabidopsis, Vitis, Populus and Oryza (APVO) were identified as follows. The gene identifiers for the *Arabidopsis* homologs of the single copy genes conserved APVO were retrieved from the Supplementary information of Duarte et al. (Duarte et al., 2010). The individual genes were retrieved from TAIR10 (Wells et al., 2013), made into a local BLAST searchable dataset, and a BLASTx search with the avocado single copy genes (isotigs) was used to identify the avocado isotigs that are APVO orthologs (Duarte et al., 2010).

GO Slim annotation was retrieved from the TAIR10 website (www.arabidopsis.org) for all the *Arabidopsis* genes. Using the BLASTx results of the avocado isotigs against the *Arabidopsis* genes, a dataset of GO Slim annotation for each isotig was generated. To make the annotation dataset uniform, a category for "no annotation available" for each of the three categories (Component C, Function F, and Process P) was created (Cgoslimterm, Fgoslimterm, Pgoslimterm), and a Perl script was written that produced a uniform dataset with only the most recent annotation entries for each gene. All Perl scripts mentioned here and in the other Methods sections are available upon request.

2.5. SNP calling

The longest isotig was selected from each isogroup of the assembly and these were concatenated to generate a non-redundant transcript sequence reference. To this reference, Illumina read sets representing each of the four cultivars were mapped using GenomeMapper, and SNPs in uniquely mapped reads were determined using SHORE (Ossowski et al., 2008). SHORE used a minimum reads support of three to call a heterozygous allele and > 80% total reads coverage to call a homozygous allele.

SHORE generated a variant report which was customized to include information on the isotig length. The variant report contained information on the isotig length, the average reads for each isotig for each parent, the average covered length for each isotig for each parent and the total number of unique SNPs for each isotig summed over all four parents. “Average mean reads” was calculated by averaging the mean number of reads for each isotig from each cultivar ('Bacon', 'Hass', 'Simmonds' and 'Tonnage'). “Average mean covered length” was calculated in a similar fashion. “Total SNPs” summed individual SNPs at unique sites from the mapping parents for the isogroups.

2.6. Determination of heterozygosity for mapping parents

From the sequence data, individual heterozygosity of the four parents used in the mapping population crosses was determined using a custom Perl script. The script screened all starting SNPs and a subset of SNPs that included at least 60 bp of SNP-free flanking sequence both upstream and downstream of the SNP. If an individual cultivar had less than six reads for a SNP, it was called a "No SNP". The position of a SNP in the transcriptome was determined by using total reads from all four parents and thus, individual parents may show different numbers of total SNPs.

2.7. Determination of synonymous and non-synonomous SNPs

Synonomous and non-synonomous SNPs were characterized from the total SNPs detected. SNPs were first filtered to eliminate those with flanking sequence containing N’s, indels, or trinucleotide SNPs. Then the amino acid translation of the ‘Hass’ transcriptome isotigs was reformatted as a BLAST-searchable database using formatdb (Trick et al., 2012). Using blastall (Trick et al., 2012), the database was queried with the filtered 58,028 SNP 121 mers in a local BLASTx search with an e threshold value of 10^{-8}. The BLASTx report was parsed with a Perl script to identify synonymous or nonsynonomous hits at each SNP position.

2.8. Filtering of variant report to design an Infinium II 6000 SNP chip

SNP data from the variant report was filtered as described in Supplementary Table S1. To remove any 121 mer SNP probe that was interrupted by an intron, the 121 mers were queried in a BLASTn search of preliminary avocado reference genome sequence (*personal communication*, Drs. Luis Herrera-Estrella and Enrique Ibarra-Laclette). Any 121 mer sequence with an alignment length less than 109 bp (90% of total) was removed.

2.9. Genotyping using the Illumina 6000 SNP chip

DNA samples for genotyping on the Illumina SNP chip were isolated using the FastPrep method (MPBio, Santa Ana, CA). The samples included 1339 individuals from the California and Florida mapping populations at SHRS, 453 germplasm accessions from both SHRS and University of California, Riverside (UCR), and 339 individuals of mapping populations from UCR and control DNAs. At least 50 ng/uL DNA and 50 uL volume were sent to Illumina for genotyping. Genotyping data were received as both a text file for analysis using Perl scripts and in a format that could be viewed using GenomeStudio software (Illumina, Inc, San Diego, CA).

2.10. Statistical analysis

All analyses were done with JMP ver. 10.0.2 (SAS, Inc), while some figures were prepared in R ver. 3.1.0 (R Core Team, 2013). Distribution of parameters in the complete dataset and all subsets were calculated and compared to the normal distribution using Analyze Distribution in JMP. Percentages of isotigs in annotation categories were calculated with Fit Y by X and compared with contingency table analyses. Scatter plot matrices were derived from Fit Y by X and Spearman's rho was used to test for significant correlations of x and y.

3. Results

3.1. Transcriptome assembly and analysis

The transcriptome was assembled from 3.15 M clean sequence reads of 359 bp average and N50 of 428 bp, totaling 1.13 Gb (Table 1). The 33,957 isogroups (OD), or estimated unigenes, identified from the transcriptome assembly was similar to the number found in other plant transcriptomes (Angeloni et al., 2011; Ibarra-Laclette et al., 2015; Novaes et al., 2008; Reeksting et al., 2014). The longest isotig in each isogroup was selected from the transcript assembly set and used as the reference for detection of polymorphism in the four cultivars. Average coverage of any base in the transcriptome was 18-fold with a standard

Table 1
Summary of RNA sequencing and transcriptome assembly of Roche 454 reads.

		Assembly	
	High Quality Reads	All Transcripts	Longest Transcripts of Isogroup
Total Number	3,154,979	68,650	33,957
Total Length (bp)	1,132,816,287	91,690,161	37,597,237
N50 length (bp)	428	1629	1336
Mean length (bp)	359 ± 127	1382 ± 1043	1107 ± 675

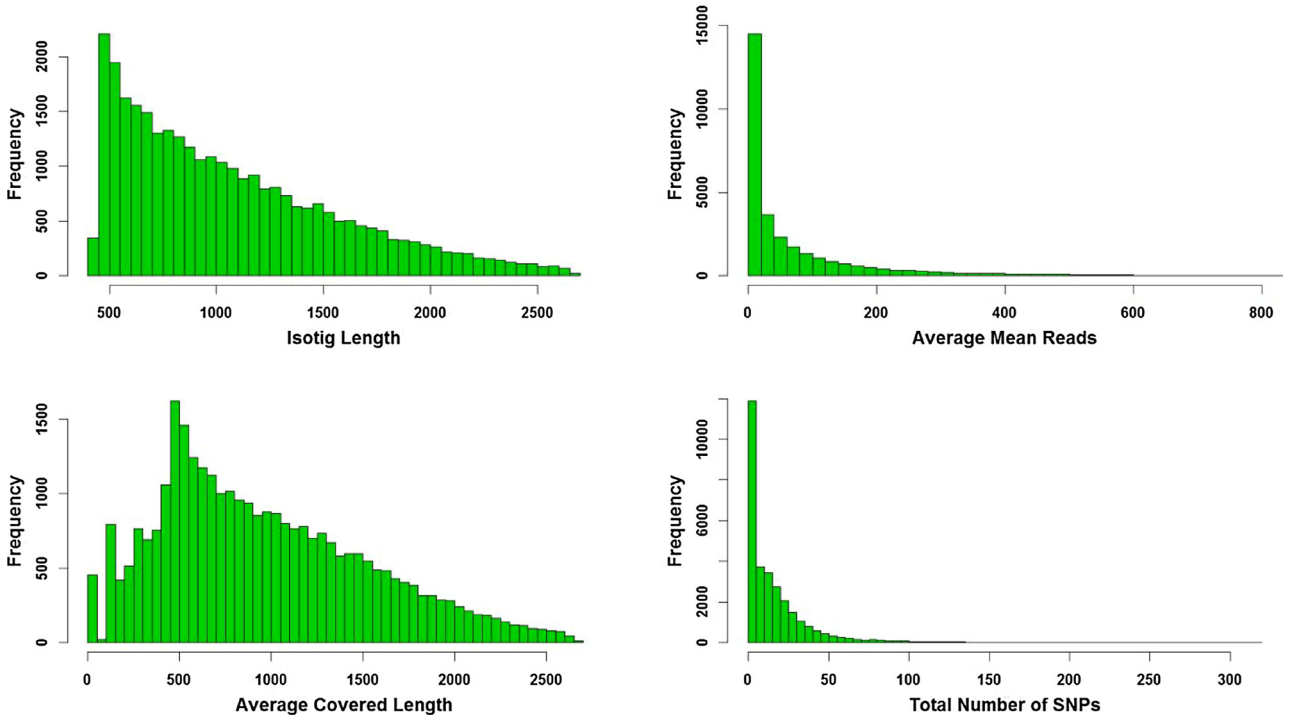

Fig. 1. Distributions of isotig length, average mean reads, average covered length, and total SNPs per isotig for the condensed dataset consisting of 30,543 isotigs.

deviation of 16.7 and a median coverage of 12.2 fold.

Parameters (isotig length, average mean reads, covered length) were not normally distributed for the entire isotig dataset (Fig. 1), and transformation either by log or square root did not improve the goodness of fit to a normal distribution. Outliers were removed from the isotig dataset, as described in Methods, to generate the condensed dataset (Table 2). The number of isogroups in the condensed dataset was 30,543 (89.9%) of the original isogroups. Although removal of outliers decreased the skew of the data for the parameters, this did not result in a normal distribution even with log or square root transformation, so nonparametric methods were used for statistical analyses.

The sequenced genomes of *Arabidopsis, Populus, Vitis* and *Oryzae* (APVO) share orthology in 959 genes identified as single copy across these four species (Duarte et al., 2010). A BLASTx alignment of the above APVO database showed 1370 avocado isotigs to be potential APVO orthologs (Table 2).

We compared the CD (30,543 isotigs), PSCI (19,129 isotigs), and the APVO (1370 isotigs) subsets for representation among the different GO ontologies, P (biological process), F (molecular function), and C (cellular compartment) (Fig. 2). In general, the CD and PSCI subset were similar in annotation percentages. The APVO subset showed some differences in percentage in particular annotation categories and the absence of certain categories of annotation (extracellular for C, receptor binding or activity for F, and electron transport or energy pathways for P). For the CD, we calculated the percentages of genes in two undefined categories (no annotation and unknown component) for the P, F and C annotations. The P annotation total was ~45% undefined (~15% no annotation + ~30% unknown component), the F annotation total was ~30% (~10% no annotation + ~20% unknown component), and the C annotation total was ~40% (~20% no annotation + ~20% unknown component).

SNP discovery identified 660,911 total variants, significantly more than have been discovered by RNAseq in other plants (Blanca et al., 2011; Trick et al., 2009). The variant report was filtered for SNP sequences that contained N at the SNP position (17,950) and indels (26,469), leaving 616,492 SNPs for the 33,957 isotigs (average 18.2 SNPs per isotig). For the condensed dataset there were 555,538 SNPs for the 30,543 isotigs (average 18.2 SNPs per isotig). Of the 30,543 isotigs, 22,881 (74.9%) had at least one SNP and 7662 (25.1%) had no SNPs (Table 3).

For isotigs with reported SNPs, isotig length and mean covered length were highly correlated (Spearman ρ = 0.98, Fig. 3a), while for isotigs lacking SNP calls, this relationship was weaker (Spearman ρ = 0.66, Fig. 3b) and included more transcripts with lower levels of expression and/or sequencing representation in our experiments. Isotig length and average mean reads coverage, as well as isotig length and total SNPs, were not well correlated in isotigs with or without SNP calls (Fig. 3).

Table 2
Avocado isotig length, covered length, and number of reads for different reference data classes. Shown are median (mean) data. Percentage of SNP/no SNP isogroups are in square brackets []. Within each SNP/No SNP pair, numbers with different superscripts (a,b) are significantly different ($p < 0.0001$, Wilcoxon test). Within each subset/not subset pair, numbers with different superscripts (w,x) are significantly different ($p < 0.0001$, Wilcoxon test).

	N	Isotig L.	Covered L.	No. Reads
Condensed dataset	30543	949 (1069)	841 (952)	25 (78)
SNPs	22881 [0.75]	1100^{a} (1188)	1058^{a} (1135)	52^{a} (103)
No SNPs	7662[0.25]	641^{b} (712)	368^{b} (406)	2^{b} (4)
PSCI subset	19129 [0.63]	854^{w} (992)	727^{w} (864)	18^{w} (63)
SNPs	13080 [0.68]	1038^{a} (1127)	994^{a} (1082)	47^{a} (91)
No SNPs	6049 [0.32]	629^{b} (699)	353^{b} (392)	2^{b} (3)
Not PSCI subset	11414 [0.60]	1108^{x} (1198)	1029^{x} (1099)	43^{x} (103)
SNPs	9801 [0.86]	1199^{a} (1270)	1146^{a} (1205)	62^{a} (119)
No SNPs	1613 [0.14]	686^{b} (759)	412^{b} (456)	2^{b} (6)
APVO subset	1370 [0.04]	1091^{w} (1140)	1030^{w} (1050)	50^{w} (85)
SNPs	1129 [0.82]	1188^{a} (1236)	1156^{a} (1191)	66^{a} (102)
No SNPs	241 [0.18]	626^{b} (692)	349^{b} (387)	2^{b} (3)
Not APVO subset	29173 [0.96]	943^{x} (1065)	831^{x} (947)	24^{x} (78)
SNPs	21752 [0.75]	1101^{a} (1186)	1053^{a} (1132)	51^{a} (103)
No SNPs	7421 [0.25]	641^{b} (713)	368^{b} (406)	2^{b} (4)

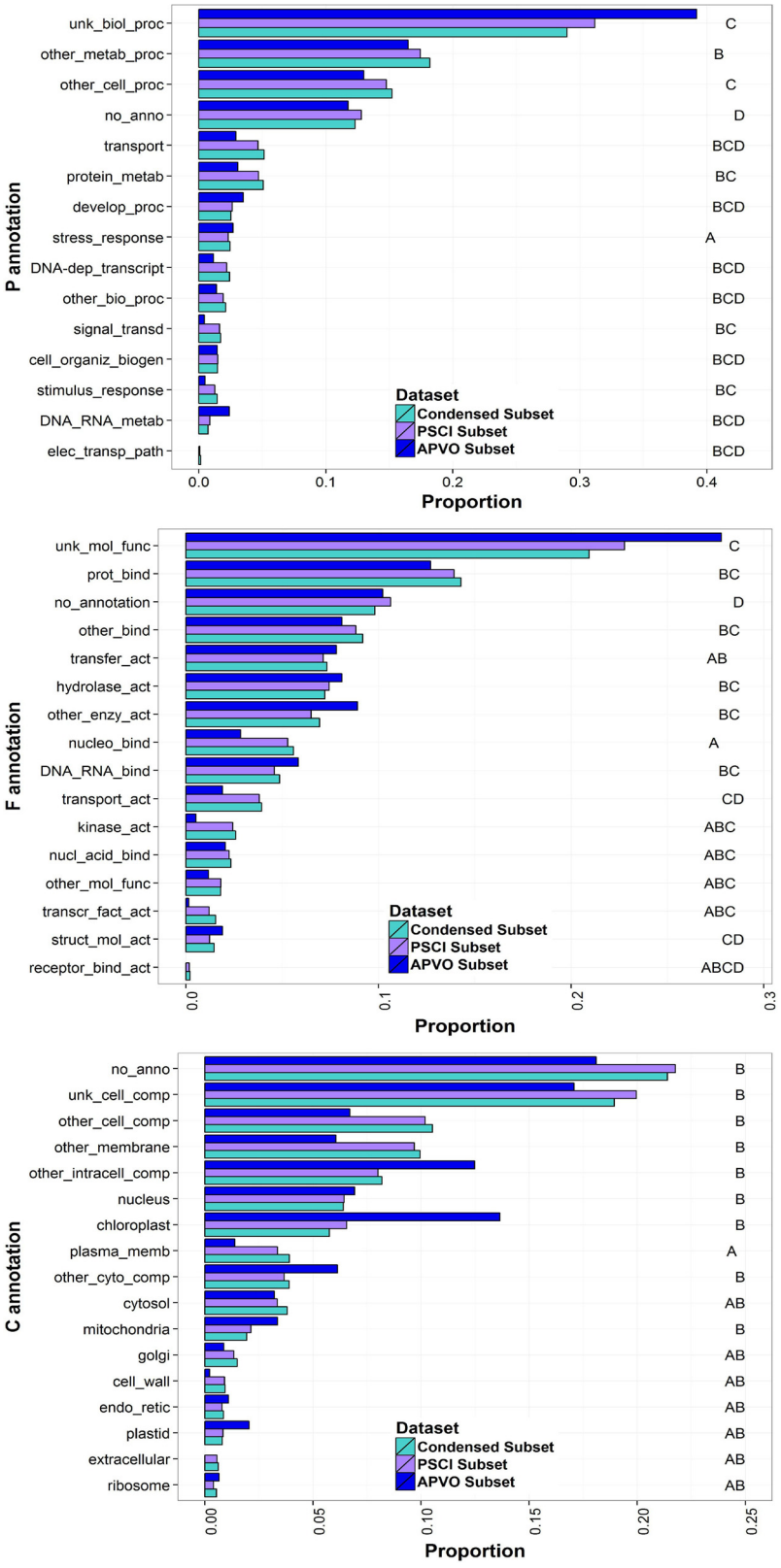

Fig. 2. CD, PSCI and APVO subsets GO annotation. On x-axis: proportion of isotigs in each GO Slim annotation categories. Connecting Letters Report for cellular compartment (C), molecular function (F), and biological process (P) annotation analyzes the total number of SNPs per isotig in the category for CD. Levels not connected by same letter are significantly different ($p < 0.001$) as determined by Wilcoxon test with A being the highest number of SNPs per isotig per category.

Table 3
Statistics for SNP frequency in condensed dataset isogroups to estimate SNP frequency (average number of SNPs per 100 nt). Median values in square brackets [].

	Condensed dataset isotigs	Isotigs with SNPs
Number of isotigs	30,543	22,881
Mean isotig length	1068.8 [949]	1188.3 [1107]
Mean covered length	951.8 [841]	1134.7 [1058]
Mean covered length/Mean isotig length	89.0% [88.6%]	95.5% [95.6%]
Mean total SNPs/isotig	18.2 [11]	24.3 [16]
Mean total SNPs/average isotig length	1.7/100 nt [1.2]	2.0/100 nt [1.4]
Mean total SNPs/average covered length	1.9/100 nt [1.3]	2.1/100 nt [1.5]

3.2. Correlation of SNPs with annotation groups for the CD, PSCI, and APVO subsets

The CD was analyzed for correlations between annotation categories and total SNPs, and results are reported as connecting letters for each annotation group (Fig. 2). Large blocks of annotation groups had similar numbers of SNPs, with some exceptions. In the C annotation, the plasma membrane category had significantly more SNPs than nine other categories. Analysis of isotig length showed that the plasma membrane category was significantly longer than only five other categories. In the F annotation, the nucleotide binding category had significantly more SNPs than nine other categories. When isotig length was analyzed, genes in the nucleotide binding category were also significantly longer than genes in nine other categories. In the P annotation, the response to stress category had significantly higher numbers of SNPs than all 14 other categories, but was not significantly longer or shorter than genes in any other category.

SNPs were initially characterized as heterozygous or homozygous by analyzing the sequence at a particular position for all four parents (Table 4). A custom Perl script was used to screen the 660,911 starting SNPs and a 69,992 subset that represented SNPs with at least 60 bp of SNP-free flanking sequence both upstream and downstream from the SNP. The 69,992 subset of SNP 121 mers was filtered to remove any SNP with N's (10,888) or indels and trinucleotide SNPs (1076), leaving 58,028 SNP 121 mers. If one of the parents was heterozygous, this SNP was termed a het SNP. If all four parents were homozygous at the position but at least one had the alternate allele, this was termed a homo SNP. The number of homo SNPs was greater for each of the four parents than the number of het SNPs. Homo SNPs will not be used for future mapping as all progeny would be heterozygous. True heterozygosity of an individual parent can only be estimated from the number of SNPs that were heterozygous for that parent, calculated as the ratio of het SNPs to total SNPs. From the condensed dataset, the order of cultivars from the most heterozygous to the least heterozygous was 'Bacon' (40%) > 'Tonnage' (38%) > 'Hass' (36%) > 'Simmonds' (30%). From the subset of SNPs with 60 bp flanking regions with no SNPs (ID), the order of cultivars was the same: 'Bacon' (32% heterozygous), 'Tonnage' (28%), 'Hass' (26%) and 'Simmonds' (14%).

Table 4
Heterozygosity of 'Bacon', 'Hass', 'Simmonds', and 'Tonnage' determined from the condensed dataset (CD) that contained 616,492 SNPs (starting number in Supplementary Table 1) and isolated SNP dataset (ID) that contained 69,992 SNPs (after first filtering step in Supplementary Table 1). Heterozygous SNPs (HetSNP) represent a heterozygous locus in the individual parent at that SNP position. Homozygote SNPs (HomSNP) represent a homozygous locus in the individual parent at that SNP position. Total number with percentage total in parentheses. Percentages do not add to 100 for the CD as each cultivar had a small percentage of SNPs for which they had insufficient reads to call a SNP. These SNPs were different for each cultivar and were removed by the first filtering step (Supplementary Table 1).

Cultivar	HetSNP CD	HetSNP ID	HomSNP CD	HomSNP ID
'Bacon'	247379 (40)	22397 (32)	350995 (57)	47595 (68)
'Hass'	223315(36)	18198 (26)	378508 (61)	51794 (74)
'Simmonds'	187814 (30)	9799 (14)	401829 (65)	60193 (86)
'Tonnage'	232440 (38)	19598 (28)	362975 (59)	50394 (72)

3.3. Types of SNPs and their frequency

Frequency of SNP occurrence in the 'Hass' transcriptome is summarized in Table 3. The range of SNP frequencies in coding regions of

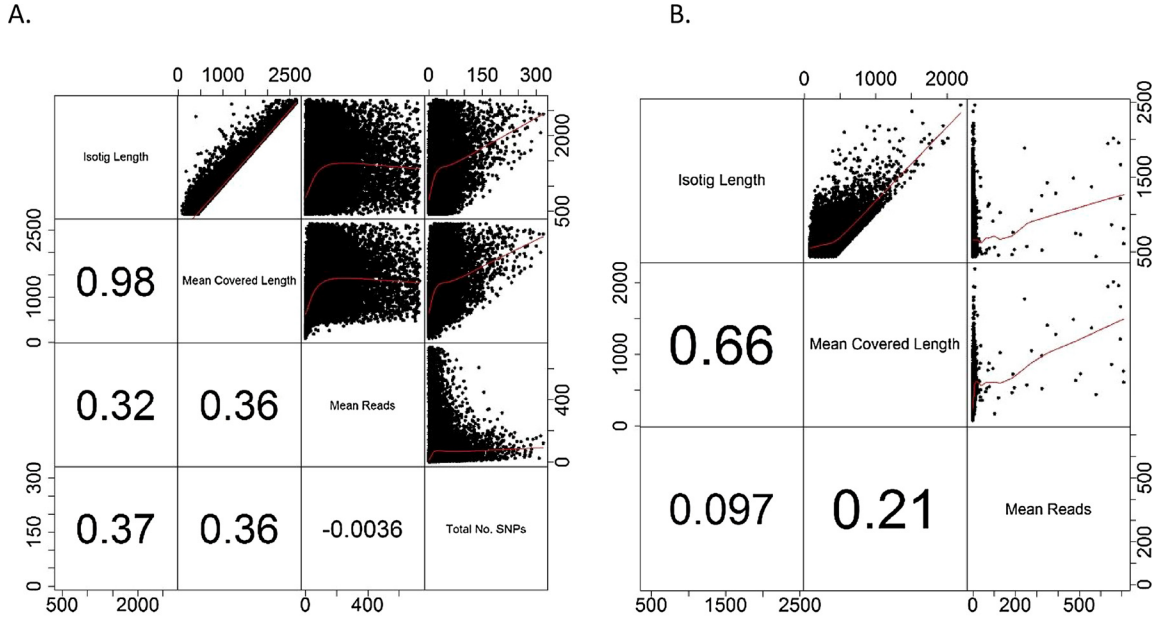

Fig. 3. A. Pairwise comparison graph for isotig length, average mean reads, average covered length and total SNPs for 22,881 isogroups that contain SNPs. B. Pairwise comparison graph for isotig length, average mean reads, average covered length and total SNPs for isogroups without SNPs. Lower half triangle gives the Pearson correlation coefficient.

Table 5
Total transitions and transversions, heterozygous and homozygous SNP transitions and transversions, and the individual categories of transitions and transversions per isotig. In this table, a heterozygous SNP refers to a SNP position where at least one of the parents is heterozygous. A homozygous SNP refers to a SNP position where all four parents are homozygous, but at least one is homozygous for the alternate allele.

Category	Mean per isotig	Std Dev	Median	% of SNPs
Total SNPs	18.2	26.12	11	100
Total Transitions	10.1	15.51	6	59
Heterozygote Transitions	9.0	14.81	4	57 (of total het SNPs)
Homozygote Transitions	1.1	2.43	0	59 (of total hom SNPs)
Total AG	5.01	7.86	3	29
Total CT	5.12	8.09	3	30
Total Transversions	8.1	11.35	4	41
Heterozygote Transversions	6.66	10.76	3	43 (of total het SNPs)
Homozygote transversions	0.80	1.90	0	41 (of total hom SNPs)
Total AT	2.34	3.89	1	13
Total AC	1.86	3.07	1	10
Total GT	1.78	3.03	1	10
Total CG	1.49	2.56	0	8

avocado genes was 1.2–2.1 per 100 bp. Thus, SNPs are ~1–2% of the transcribed gene sequences in avocado, based on the diversity of the parents of our mapping populations.

The types of SNPs and their frequency were determined from the sequence data and the variant report. For all cases the mean number of SNPs per isotig was higher than the median, and the standard deviation was larger than the mean (Table 5). Transitions were more frequent than transversions, whether they were het SNPs or homo SNPs, and no difference in frequency of the type of SNP was noted between het and homo SNPs. For transitions, AG and CT were roughly equal in frequency (~30%). For transversions, the most commonly observed was AT (13%) and the least was CG (8%).

From a total of 58,028 SNP 121 mers as queries, a BLASTx search of the translated 'Hass' transcriptome resulted in 43,010 (74.1%) hits with e score less than 10^{-8} and 15,018 no hits (25.9%). Of the 43,010 hits, 23,685 (55.1%) were synonymous and 19,325 (44.9%) nonsynonymous.

3.4. Filtering of SNPs and design of the Illumina II 6000 SNP chip

SNPs from the CD were filtered as described in Methods and Supplemental Table S1. At step 6, we initially filtered for SNPs that could be mapped in both mapping populations, as they would produce the most resolved joined map. This filter was too stringent since there were less than 6000 SNPs that could be mapped in both populations (Fig. 4). Stringency was reduced for step 6 by accepting SNPs that could be mapped in at least one population.

3.5. Comparison of SNP calls by sequencing and by SNP chip assay

The sequencing results predicting SNP genotypes and the SNP chip results for the four parents were compared. Of the 6000 designed SNP assays, there were 5332 chip genotype loci (88.9%) that were successfully synthesized with 668 loci (11%) without information due to bead failure. Of the 5332 remaining SNPs, 282 (5%) gave no genotype scores against the 2021 individuals. Genotyping of the entire dataset (mapping populations, germplasm, and controls) as described in methods was used for this analysis. Of the 5050 remaining loci, concordance between sequence and chip genotypes are presented in Table 6 for each parent. Reanalyzing the chip data, 307 SNPs were

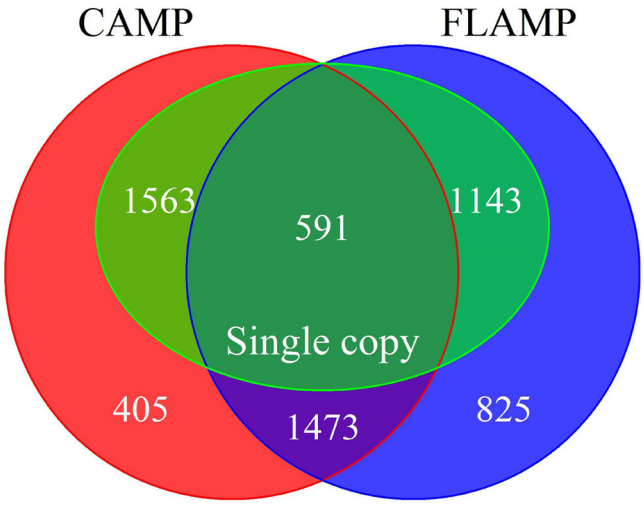

Fig. 4. Venn diagram of SNPs selected for Infinium II 6k chip. Red circle: SNPs mappable in FLAMP, Blue circle: SNPs mappable in CAMP. Green circle: SNPs in Single copy transcripts. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article).

Table 6
Concordance between sequence-based SNP calls and array-based genotyping calls for the 5050 SNPs used on the chip. Values in brackets are with 307 failed chip assays removed (4743 total). Average sequence reads in support of sequence calls are shown in brackets. *Hom Ref,* homozygous reference ('Hass'); *Het*, heterozygous; *Hom Alt*, homozygous alternative allele; *Miss*, data missing.

		Array				
		Hom Ref	Het	Hom Alt	Miss	Congruence
Sequencing	Bacon					
	Hom Ref	1632 [102]	201 [121]	43 [11]	9	90%
	Het	0 [0]	2181 [91]	1 [19]		
	Hom Alt	169 [83]	62 [68]	448 [99]		
	Hass					
	Hom Ref	2957 [86]	224 [84]	0 [0]	5	95%
	Het	1 [18]	1551 [79]	0 [0]		
	Hom Alt	0 [0]	7 [34]	0 [0]		
	Simmonds					
	Hom Ref	2011 [95]	90 [60]	191 [24]	6	88%
	Het	2 [25]	850 [103]	6 [82]		
	Hom Alt	282 [81]	16 [75]	1291 [102]		
	Tonnage					
	Hom Ref	1199 [81]	160 [100]	79 [13]	6	91%
	Het	1 [32]	2648 [90]	2 [82]		
	Hom Alt	157 [63]	40 [50]	454 [90]		

scored as homozygous for all 2021 individuals on the chip and were removed from further analysis, because they were either incorrectly designed or were originally incorrectly identified as SNPs. This left 4743 SNPs (79% of designed SNP assays) that could be called by both sequencing and the chip. Concordance of the sequence and chip genotypes for the 4743 SNPs are shown in Table 6, along with the average sequence read number for each concordance category of the 4743 SNPs as shown in brackets.

The SNP chip had only four types of SNPs (AC, AG, CT, GT) and the distribution was ~10% each for AC and GT and ~30% each for AG and CT (Table 5). The distribution of SNP types of the 307 highly homozygous loci that were removed were essentially the same as the entire chip, suggesting that there was no particular SNP type that had been miscalled by the chip.

There were two types of non-concordance between the sequence and chip calls (Table 6). In the first type, the sequence predicted a

homozygote for a particular parent but the chip called it a heterozygote. This non-concordance will be innocuous for mapping purposes, as at least one of the parents was heterozygous for the SNP locus to have been included on the chip (Supplemental Table S1, Step 7). The correct parental genotype will be determined when the segregation of these SNPs is observed in the mapping populations. In the second type, the sequence predicted a heterozygote for a parent, which could be mapped, and the chip called it a homozygote. For that SNP locus, if only one parent had been predicted by the sequence to be a heterozygote for its inclusion on the chip, then this locus could not be mapped in our populations.

We screened the chip results for loci that had homozygous allele frequencies of greater than 0.9 across all 2021 individuals scored on the chip. For a SNP locus to have been chosen for the chip, it had to be heterozygous in at least one parent and have an alternate allele frequency of greater than 0.2 and less than 0.8. There were 1500 progeny of the four parents genotyped on the chip and 450 unique germplasm accessions, which made such high levels of homozygosity suspect and further suggested a chip genotyping problem. There were 307 loci with a homozygous allele frequency above 0.95. Only 12 showed at least one of the 2021 genotypes to be heterozygous or homozygous for the alternate allele, the other 295 being homozygous for a single allele for all 2021 individuals. Prior to removal of the 307 suspect loci, there were 528 calls that were non-concordant for the sequence heterozygous/chip homozygous categories ('Bacon' 196, 'Hass', 114, 'Simmonds', 114, 'Tonnage', 104, Table 6). After removal, there was a total of 13 non-concordant calls ('Bacon' 1, 'Hass' 1, 'Simmonds' 8, 'Tonnage' 3). Overall concordance of the genotype dataset with the 307 loci removed was in the range of 88–95%.

4. Discussion

SNP marker development for marker-assisted selection and genetic mapping of crop species without a sequenced genome poses the unique problem of generating a reference for aligning sequences and identifying SNPs (Russell et al., 2011). SNP markers are present in much higher frequency than microsatellite markers, especially in transcribed genes. In addition, SNP markers are unambiguous, can be assayed successfully on any platform, and, thus, SNP genotype data can be reliably shared between laboratories. For tree breeding, which can take up to 15 years before cultivar release, it is essential to have sufficient genetic markers to identify offtypes (alleles from a different source than the parents) and progeny without desired alleles at the seedling stage. The majority of the costs in tree breeding are field space, tree maintenance, and evaluation over multiple years once the trees begin bearing fruit. Genotyping seedlings can increase the number of trees with the desired alleles that will be evaluated, saving field space and evaluation costs, as well as improving the efficiency of the breeding process.

We undertook this SNP discovery project to create a SNP array and to generate data that would lead to a highly saturated genetic map with SNPs for the avocado genome that will facilitate QTL mapping in two large F1 populations (FLAMP and CAMP). We had previously completed a medium resolution, SSR map of the FLAMP population (Borrone et al., 2009), but attempting a similar map of the CAMP population with SSR markers would have been too costly, taken too long, and still resulted in a medium resolution map with some linkage groups represented by only a few markers. By developing SNPs and the Infinium II SNP chip, we could complete the genotyping in a shorter period of time, efficiently merge the maps, and estimate genetic diversity in our germplasm collection. This SNP discovery project provided sufficient SNPs to identify a subset that was single copy, could be mapped in one or both populations, and represented 6000 different contigs. This provided maximum information both to construct our genetic recombination maps and to characterize our germplasm collections.

Additionally, we chose to develop SNPs because if the SNPs are mapped, the SNP sequence can be used to align the genome sequence and pull together scaffolds in the genome sequence assembly process. The only platform dependence seen for SNPs is in the differing platforms' requirements for SNP assay design. The SNPs designed for the Infinium II chip had a greater than 99% design success rate for both the Fluidigm and Sequenom platforms. Being able to target particular portions of the avocado genome with SNP markers will be useful for doing association studies in the germplasm once we have identified candidate genes for different traits, such as cold tolerance, *Phytophthora* resistance, fat content, salinity tolerance, etc., through QTL mapping.

4.1. Reference transcriptome

The transcriptome assembly process gave us ~30,000 isotigs of average length ~1000 bp, which compares favorably with other plant transcriptomes (Angeloni et al., 2011) considering we only used leaf and flower tissues. Both the total transcripts and condensed datasets had ~75% isotigs with SNPs. The ~25% of the isotigs that did not contain SNPs were also about one third the covered length of SNP-containing isotigs. We calculated the frequency of SNPs in the dataset as a range of 1.2–2.1 SNPs per 100 bp, so short isotigs could simply be missing SNPs due to chance. The short isotigs that did not have SNPs also had significantly lower numbers of reads (average of two reads per nucleotide) compared to the 25–30 fold higher reads for isotigs with SNPs. In addition, SNPs were only called based on variation in the four parents of our mapping populations. SNP frequency in coding regions and non-coding may be higher if more genetically diverse cultivars are compared to the reference transcriptome.

In contrast, ~300 microsatellites had previously been identified (Borrone et al., 2009) from an avocado EST database consisting of 16,000 unigenes for an average of 1.9 markers per 100 unigenes, as opposed to 18.2 SNPs per isotig. SNPs in coding regions are present ~1000 times more frequently than microsatellite markers. If only polymorphic microsatellite markers are considered, SNPs are present ~1500 times more frequently than microsatellite markers in coding regions.

SNPs in isotigs were not normally distributed, nor was isotig length, isotig coverage or average number of reads per isotig. We attempted to determine whether the number of SNPs was normally distributed in subsets of the dataset, such as the CD, PSCI, or APVO genes. Isotig length, isotig coverage, average number of reads per isotig and total SNPs per isotig were also not normally distributed for these data subsets. Of interest, 1370 of the avocado isotigs had significant identity with 959 APVO genes. As the APVO isotigs were mapped on to complete gene sequences, this probably represents an overestimation of the actual number of genes/transcripts identified in our study.

With the advent of next generation sequencing, SNP discovery projects such as this one can uncover large numbers of SNPs in a short time and allow filtering of the SNPs to get a subset of markers of much higher value than the typical set of markers. However, the sequencing data should be analyzed with regard to how much coverage is necessary to give 95% confidence that a SNP has been identified. We had some concern that the number of reads would play a role in identifying SNPs, but, surprisingly, the number of SNPs did not correlate appreciably with the number of reads. The number of reads was not correlated strongly with covered length, either, suggesting that anything above ~6x coverage per parent for any particular nucleotide position was saturated for SNP information.

4.2. Transcript annotation

We wanted to determine if there would be different distributions of gene annotation types in the CD, the PSCI, and the APVO dataset to assure that we could use a reduced subset of the isotigs for design of the Infinium II chip without decreasing coverage of the transcriptome. We

used the GO Slim terms from the TAIR site because they have a significantly reduced number of categories for each annotation type, which made the analysis possible for such a large number of isotigs. Our analysis showed that the distributions for all datasets were similar with the exception of missing annotation categories in the much smaller APVO subset of isotigs.

Using the GO Slim categories, we also observed the average number of SNPs per gene to identify annotation categories with significantly higher or lower numbers of SNPs (Fig. 2). The most dramatic correlation between SNP number and annotation type was for the process type annotation (P) response to stress category, which had significantly more SNPs than any of the other P categories. Stress response genes may be changing more rapidly than other genes. The higher number of SNPs in the response to stress category was not simply due to longer isotigs, as the analysis of isotig length to GO Slim category showed that the response to stress category was not significantly different from all other categories with regard to isotig length.

Single copy genes had a lower frequency of SNPs but also showed the correlation with a higher number of SNPs for the stress response category, while no significant differences were seen in the APVO subset of genes with regard to SNP frequency, isotig length or a particular correlation with an annotation category. Genes in the No Annotation and No Known Activity categories made up 30–40% of all datasets but generally had the fewest number of SNPs, which may be related to overall shorter sequences for these isotigs.

4.3. Infinium II chip assay design

One goal of this research was to identify SNPs for the design of an Infinium II 6000 SNP chip. The Infinium chip design requirements are that the SNP is in a region of sequence that does not contain another SNP for 60 nucleotides on either side of the SNP. We identified ~60,000 SNPs of the original ~600,000 that fulfilled this requirement. Since creation of a genetic map is our larger goal, we filtered for SNPs that were heterozygous in at least one parent from both the California and Florida mapping populations to allow mapping of the SNP in both populations, which will improve the resolution of the joined maps. We had less than 6000 SNPs that fulfilled this requirement, so we reduced the stringency and accepted SNPs that could be mapped in one population or the other. We used only single copy genes, as defined in Methods, to avoid the problem of false parental heterozygotes due to sequence genotyping of two homozygous paralogs, which would prevent mapping, as all progeny would be heterozygous. We only used a single SNP from each gene to broaden the coverage in our maps. Previous experience with chip design suggested that having more than one SNP per gene was unnecessary (Motamayor et al., 2013). The chip that we designed in this study was used to genotype the mapping populations and all of the germplasm collection. These analyses will be presented in future manuscripts.

4.4. Heterozygosity of the mapping parents

We used the sequence genotype data from the entire dataset, ~616,000 SNPs, and the ~69,000 Illumina design SNPs, to estimate heterozygosity in the parents of the mapping populations. The order of percentage heterozygosity was the same in both datasets, but the actual amounts of heterozygosity were quite different for each parent, with lower amounts of heterozygosity detected in the ~69,000 subset. Perhaps when multiple SNPs are present in a short span of sequence, alignment of the short reads is not as precise and a higher level of heterozygosity is estimated than when SNPs are more isolated. Comparison to the microsatellite data used for the previous map is difficult, as reports were only for loci that were polymorphic in either 'Simmonds' or 'Tonnage', rather than for all four parents. However, microsatellite data for 'Simmonds' and 'Tonnage' also shows the same order with more heterozygotes in 'Tonnage' (93%) than in 'Simmonds' (39%) (Borrone et al., 2009). Exact percentages of heterozygosity estimated by both SNP and microsatellite markers may have been skewed due to subsampling and to constraint on occurrence of markers, as both were identified from expressed genes (Borrone et al., 2007).

4.5. Types of SNPs

We looked at the number of transitions and transversions and the individual types of SNPs for the entire dataset. As seen by others who have used transcriptomes for identifying SNPs (Novaes et al., 2008; Scaglione et al., 2012; Trick et al., 2009), we had more transitions than transversions (59% to 41%), and the majority of SNPs detected were heterozygous in at least one parent (het SNP) rather than being a sequence difference between homozygous parents (homo SNP). Homo SNPs would not have been used for the chip, as they would have resulted in all progeny being heterozygous, and thus the locus could not be mapped. AG and CT transitions were approximately equal, while the transversions were AT > AC~ = GT > CG.

When studying an agriculturally important plant where the genome sequence is not available, SNPs in protein-coding regions are especially interesting. SNPs can potentially lead to amino acid changes that may change the function of a protein or produce nonsense mutations that could truncate the protein. These would be interesting markers to follow in both mapping populations and in germplasm collections. For this project, we did not pursue identifying synonymous and non-synonymous SNPs prior to designing our avocado SNP chip, instead choosing to look at other factors, such as the ability to be mapped in both mapping populations, presence in single copy isotigs, etc. After the chip design, we did analyze synonymous and nonsynonomous SNPs in the ~60,000 well isolated SNPs (60 bp on either side of the SNP without another SNP). By translating the isotigs used for the transcriptome, we obtained a partial set of gene models, as not all isotigs had a continuous reading frame. After analysis of a BLASTx search of the partial gene models with the well isolated SNP 121 mers, we estimated that 45% of the SNPs were nonsynonomous, a surprisingly high percentage for coding regions. Based on these results, we have removed nonsynonomous SNPs as part of our filtering in a similar SNP discovery project for mango (Kuhn et al., 2014), as there is likely less skewing or selection pressure in the inheritance of synonymous SNPs (personal communication, Dapeng Zhang, USDA-ARS BARC, Beltsville, MD).

4.6. Comparison of SNP scoring by sequencing and chip

We had the opportunity to compare the predicted genotypes from our RNA sequencing with the scored genotypes from the SNP chip. The congruence of the two methods of genotyping ranged from 88% for 'Simmonds' to 95% for 'Hass', the cultivar from which the reference transcriptome was made. Initial analysis of congruence was in the range of 86–93%, but removal of 307 highly homozygous loci increased congruence by 2% for each parent. There was no bias to a particular SNP type either for the 307 removed loci or all the remaining non-congruent loci. Fortunately, there were only 13 cases for the four parents where the sequence predicted a heterozygote and the chip a homozygote, which were the most critical types of noncongruence for the success of our genetic map. In addition, the lowest read support for this type of noncongruence was an average of 18 reads, and the average read support for all noncongruent categories was > 50 reads. The correlation between reads and SNP identification suggests that more than six reads for a single parent is sufficient for SNP calling, however, in the absence of further data, it is not possible to assign the miscall to either the sequence or the chip. In general, our chip genotype data strongly supports the accuracy of the RNA sequencing data for this type of SNP discovery project.

4.7. Future work

In the future, we will be able to genotype diverse avocado collections or other avocado F1 populations (e.g. West Indian x Mexican and reciprocal) with an informative subset of the SNPs that are evenly distributed across the linkage groups, and we will be able to use any SNP genotyping platform. An immediate outcome of the SNP discovery process was to estimate homozygosity and heterozygosity in the parents of our mapping populations. A similar preliminary analysis of the germplasm collection has identified trees that are greater than 95% homozygous that will make excellent candidates for resequencing once the avocado genome is available.

In addition, SNP alleles specific to West Indian, Guatemalan and Mexican subspecies in the germplasm collection have been identified; the alleles will simplify parental inference studies for the germplasm collection and commercially grown cultivars. For cultivars that are progeny of recent crosses, specific SNPs may be inherited as haplotype blocks. Determination of the haplotypes of the mapping population parents should increase both the power of the QTL mapping by associating haplotype blocks to the trait, as has been done using the iXora program (Utro et al., 2013), and by inferring the parents of germplasm individuals.

We have already begun SNP discovery projects for *Mangifera indica* (mango) with the goal of creating a genetic map and estimating genetic diversity worldwide germplasm collections. Our success with the avocado SNP discovery process described here has informed and improved the project, particularly with regard to filtering the SNPs for assay design. In addition, the potential to design SNP assays for the APVO orthologs may allow us to generate sufficient molecular markers to estimate germplasm genetic diversity in a variety of lesser crops (carambola, longan, jujube, mamey, sapote) where the expense of an individual SNP discovery project is prohibitive.

5. Conclusions

The 'Hass' reference transcriptome is made up of ~34,000 unique transcript assemblies and a total of ~92 M nucleotides. By mapping RNA sequences from the four parents of the mapping population onto the reference transcriptome, approximately 640,000 SNPs were identified among ~75% of the transcripts. In the transcripts containing SNPs, the average frequency was 1.2–2.1 per 100 nucleotides. Analysis of the SNP frequency in transcripts based on GO Slim annotation identified a significantly higher number of SNPs in stress-related transcripts. Of the ~640,000 SNPs, 6000 SNPs chosen for design of the Infinium II SNP genotyping array (SHRS avochip) fulfilled the following selection criteria: one SNP per transcript, heterozygous in at least one parent, and well-isolated from other SNPs and predicted intron interruptions. The SHRS avochip was used to genotype the parents and ~1500 individuals from the mapping populations. Both SNP chip and RNA sequence data was available for the parents and congruence in parental genotype calls between the RNA sequence data generated for SNP discovery and the SHRS avochip ranged from 88 to 95% depending on the parent. From this congruence, it was concluded that as few as six reads at a nucleotide position allowed SNP determination as accurate as higher coverage.

In summary, a leaf and flower transcriptome was successfully created and used for avocado SNP discovery. Sufficient SNPs that could be used for genetic mapping and germplasm evaluation were generated for design of a 6000 SNP Infinium II chip. Genotyping by RNA sequencing and chip array were essentially identical with respect to accuracy. SNP discovery by RNA sequencing when no genomic DNA sequence is available is a successful strategy for marker generation in specialty crops.

Competing interests

The authors declare that they have no competing financial interests.

Author details

DNK designed the experiment, analyzed the SNP data and authored the manuscript. DSLIII filtered the SNPs and designed the Infinium II chip. JHR provided statistical analysis advice, prepared figures and critically evaluated the manuscript. AC, PM and NVdB participated in experimental design, data analysis and critical evaluation of the manuscript. All authors read and approved the final manuscript.

Acknowledgements

D.N. Kuhn acknowledges support by USDA-ARS CRIS 6631-21000-022. We thank Drs. K. Mockaitis and R. Podicheti (IU), supported by USDA ARS Specific Cooperative Agreement 58-6631-0-101, for transcriptome assembly and SNP calling. We thank James Ford and Zach Smith (IU) for expert sequencing; and Greg Zynda and Aaron Buechlein (IU) for sequence data analysis support. We thank Drs. Luis Herrera-Estrella and Enrique Ibarra-Laclette (Langebio, CINVESTAV) for providing preliminary avocado genomic sequence data for intron identification; Giuliana Mustiga and Dr. J. Conrad Stack (MARS, Inc) for statistical analysis advice; Dr. MaryLu Arpaia for providing germplasm and mapping populations from the UCR collections; and Barbie Freeman (SHRS) for outstanding technical support in RNA isolation.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:https://doi.org/10.1016/j.scienta.2018.10.011.

References

Alcaraz, M.L., Hormaza, J.I., 2007. Molecular characterization and genetic diversity in an avocado collection of cultivars and local Spanish genotypes using SSRs. Hereditas 144, 244–253.

Angeloni, F., Wagemaker, C.A.M., Jetten, M.S.M., den Camp, H., Janssen-Megens, E.M., Francoijs, K.J., Stunnenberg, H.G., Ouborg, N.J., 2011. De novo transcriptome characterization and development of genomic tools for Scabiosa columbaria L. using next-generation sequencing techniques. Mol. Ecol. Resour. 11, 662–674.

Ashworth, V., Clegg, M., 2003. Microsatellite markers in avocado (Persea americana Mill.): genealogical relationships among cultivated avocado genotypes. J. Hered. 94, 407–415.

Ashworth, V.E.T.M., Kobayashi, M.C., De La Cruz, M., Clegg, M.T., 2004. Microsatellite markers in avocado (Persea americana Mill.): development of dinucleotide and trinucleotide markers. Scientia Horticulturae 101, 255–267.

Ayala-Silva, T., Schnell, R., Gordon, G., Winterstein, M.C., 2012. Application of propiconazole in management of laurel wilt disease in avocado (Persea Americana MILL.) trees. Proceedings of the First International Symposium on Wild Realtives of Subtropical and Temperate Fruit and Nut Crops. Acta Horticulturae. pp. 71–78.

Ballerini, E.S., Mockaitis, K., Arnold, M.L., 2013. Transcriptome sequencing and phylogenetic analysis of floral and leaf MIKC(C) MADS-box and R2R3 MYB transcription factors from the monocot Iris fulva. Gene 531, 337–346.

Blanca, J.M., Canizares, J., Ziarsolo, P., Esteras, C., Mir, G., Nuez, F., Garcia-Mas, J., Pico, M.B., 2011. Melon transcriptome characterization: simple sequence repeats and single nucleotide polymorphisms discovery for high throughput genotyping across the species. Plant Genome 4, 118–131.

Borrone, J.W., Schnell, R.J., Violi, H.A., Ploetz, R.C., 2007. Seventy microsatellite markers from Persea americana Miller (avocado) expressed sequence tags. Mol. Ecol. Notes 7, 439–444.

Borrone, J.W., Olano, C.T., Kuhn, D.N., Brown, J.S., Schnell, R.J., Violi, H.A., 2008. Outcrossing in Florida avocados as measured using microsatellite markers. J. Am. Soc. Hortic. Sci. 133, 255–261.

Borrone, J.W., Brown, J.S., Tondo, C.L., Mauro-Herrera, M., Kuhn, D.N., Violi, H.A., Sautter, R.T., Schnell, R.J., 2009. An EST-SSR-based linkage map for Persea americana Mill. (avocado). Tree Genet. Genomes 5, 553–560.

Calderon-Vazquez, C., Durbin, M.L., Ashworth, V.E.T.M., Tommasini, L., Meyer, K.K.T., Clegg, M.T., 2013. Quantitative genetic analysis of Three important nutritive traits in the fruit of avocado. J. Am. Soc. Hortic. Sci. 138, 283–289.

Chen, H., Morrell, P.L., de la Cruz, M., Clegg, M.T., 2008. Nucleotide diversity and linkage disequilibrium in wild avocado (Persea americana Mill.). J. Hered. 99, 382–389.

Duarte, J.M., Wall, P.K., Edger, P.P., Landherr, L.L., Ma, H., Pires, P.K., Leebens-Mack, J., 2010. Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis

and oryza and their phylogenetic utility across various taxonomic levels. BMC Evol. Biol. 10, 61.
FAO, 2016. Global Avocado Production in 2013, by Country.
Ibarra-Laclette, E., Mendez-Bravo, A., Perez-Torres, C.A., Albert, V.A., Mockaitis, K., Kilaru, A., Lopez-Gomez, R., Cervantes-Luevano, J.I., Herrera-Estrella, L., 2015. Deep sequencing of the Mexican avocado transcriptome, an ancient angiosperm with a high content of fatty acids. BMC Genomics 16, 599.
Kendra, P.E., Montgomery, W.S., Niogret, J., Pena, J.E., Capinera, J.L., Brar, G., Epsky, N.D., Heath, R.R., 2011. Attraction of the Redbay Ambrosia Beetle, Xyleborus glabratus, to avocado, lychee, and essential oil lures. J. Chem. Ecol. 37, 932–942.
Kuhn, D.N., Livingstone, D., Main, D., Zheng, P., Saski, C., Feltus, F.A., Mockaitis, K., Farmer, A.D., May, G.D., Schnell, R.J., Motamayor, J.C., 2012. Identification and mapping of conserved ortholog set (COS) II sequences of cacao and their conversion to SNP markers for marker-assisted selection in Theobroma cacao and comparative genomics studies. Tree Genet. Genomes 8, 97–111.
Kuhn, D., Dillon, N., Innes, D., Wu, L.-S., Mockaitis, K., 2014. Development of single nucleotide polymorphism (SNP) markers from the mango (Mangifera indica) transcriptome for mapping and estimation of genetic diversity. XXIX International Horticultural Congress on Horticulture: Sustaining Lives, Livelihoods and Landscapes (IHC2014): IV 1111. pp. 315–322.
Lavi, U., Akkaya, M., Bhagwat, A., Lahav, E., Cregan, P.B., 1994. Methodology of generation and characteristics of simple sequence repeat DNA markers in avocado (Persea americana M.). Euphytica 80, 171–177.
Matvienko, M., Kozik, A., Froenicke, L., Lavelle, D., Martineau, B., Perroud, B., Michelmore, R., 2013. Consequences of normalizing transcriptomic and genomic libraries of plant genomes using a duplex-specific nuclease and tetramethylammonium chloride. Plos One 8.
Mhameed, S., Hillel, J., Lahav, E., Sharon, D., Lavi, U., 1995. Genetic association between DNA fingerprint fragments and loci controlling agriculturally important traits in avocado (Persea-Americana Mill). Euphytica 84, 81–87.
Motamayor, J.C., Mockaitis, K., Schmutz, J., Haiminen, N., Livingstone 3rd, D., Cornejo, O., Findley, S.D., Zheng, P., Utro, F., Royaert, S., Saski, C., Jenkins, J., Podicheti, R., Zhao, M., Scheffler, B.E., Stack, J.C., Feltus, F.A., Mustiga, G.M., Amores, F., Phillips, W., Marelli, J.P., May, G.D., Shapiro, H., Ma, J., Bustamante, C.D., Schnell, R.J., Main, D., Gilbert, D., Parida, L., Kuhn, D.N., 2013. The genome sequence of the most widely cultivated cacao type and its use to identify candidate genes regulating pod color. Genome Biol. 14, R53.
Novaes, E., Drost, D.R., Farmerie, W.G., Pappas, G.J., Grattapaglia, D., Sederoff, R.R., Kirst, M., 2008. High-throughput gene and SNP discovery in Eucalyptus grandis, an uncharacterized genome. Bmc Genomics 9.
Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., Weigel, D., 2008. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res. 18, 2024–2033.
Paritosh, K., Gupta, V., Yadava, S.K., Singh, P., Pradhan, A.K., Pental, D., 2014. RNA-seq based SNPs for mapping in Brassica juncea (AABB): synteny analysis between the two constituent genomes A (from B. rapa) and B (from B. nigra) shows highly divergent gene block arrangement and unique block fragmentation patterns. BMC Genomics 15.
Ploetz, R.C., Pena, J.E., Smith, J.A., Dreaden, T.J., Crane, J.H., Schubert, T., Dixon, W., 2011. Laurel Wilt, caused by Raffaelea lauricola, is confirmed in Miami-Dade County, Center of Florida's Commercial Avocado Production. Plant Disease 95 1589-1589.
R Core Team, 2013. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienne, Austria.
Ramirez-Gil, J.G., Castaneda-Sanchez, D.A., Morales-Osorio, J.G., 2017. Production of avocado trees infected with Phytophthora cinnamomi under different management regimes. Plant Pathol. 66, 623–632.
Reeksting, B.J., Coetzer, N., Mahomed, W., Engelbrecht, J., van den Berg, N., 2014. De novo sequencing, assembly, and analysis of the root transcriptome of Persea americana (Mill.) in response to phytophthora cinnamomi and flooding. PLoS One 9.
Russell, J.R., Bayer, M., Booth, C., Cardle, L., Hackett, C.A., Hedley, P.E., Jorgensen, L., Morris, J.A., Brennan, R.M., 2011. Identification, utilisation and mapping of novel transcriptome-based markers from blackcurrant (Ribes nigrum). BMC Plant Biol. 11.
Scaglione, D., Lanteri, S., Acquadro, A., Lai, Z., Knapp, S.J., Rieseberg, L., Portis, E., 2012. Large-scale transcriptome characterization and mass discovery of SNPs in globe artichoke and its related taxa. Plant Biotechnol. J. 10, 956–969.
Schnell, R.J., Brown, J.S., Olano, C.T., Power, E.J., Krol, C.A., Kuhn, D.N., Motamayor, J.C., 2003. Evaluation of avocado germplasm using microsatellite markers. J. Am. Soc. Hortic. Sci. 128, 881–889.
Schnell, R.J., Tondo, C.L., Brown, J.S., Kuhn, D.N., Ayala-Silva, T., Borrone, J.W., Davenport, T.L., 2009. Outcrossing between' bacon' pollinizers and adjacent' hass' avocado trees and the description of two new lethal mutants. Hortscience 44, 1522–1526.
Sharon, D., Cregan, P.B., Mhameed, S., Kusharska, M., Hillel, J., Lahav, E., Lavi, U., 1997. An integrated genetic linkage map of avocado. Theor. Appl. Genet. 95, 911–921.
Sharon, D., Hillel, J., Mhameed, S., Cregan, P.B., Lahav, E., Lavi, U., 1998. Association between DNA markers and loci controlling avocado traits. J. Am. Soc. Hortic. Sci. 123, 1016–1022.
Trick, M., Long, Y., Meng, J.L., Bancroft, I., 2009. Single nucleotide polymorphism (SNP) discovery in the polyploid Brassica napus using Solexa transcriptome sequencing. Plant Biotechnol. J. 7, 334–346.
Trick, M., Adamski, N.M., Mugford, S.G., Jiang, C.C., Febrer, M., Uauy, C., 2012. Combining SNP discovery from next-generation sequencing data with bulked segregant analysis (BSA) to fine-map genes in polyploid wheat. BMC Plant Biol. 12.
Utro, F., Haiminen, N., Livingstone 3rd, D., Cornejo, O.E., Royaert, S., Schnell, R.J., Motamayor, J.C., Kuhn, D.N., Parida, L., 2013. iXora: exact haplotype inferencing and trait association. BMC Genet. 14, 48.
Wells, R., Trick, M., Fraser, F., Soumpourou, E., Clissold, L., Morgan, C., Pauquet, J., Bancroft, I., 2013. Sequencing-based variant detection in the polyploid crop oilseed rape. BMC Plant Biol. 13.