Linking Candidate Genes to Biochemical Phenotypes in Avocado

Michael T. Clegg
UC Irvine

Mary Durbin, Livia Tommasini, Carlos Calderon, Vanessa Ashworth

Genetic markers can be developed that track the transmission from parent to offspring of high or low levels of nutritionally valuable compounds in the avocado fruit. The goal of this project is to develop SNP (single nucleotide polymorphism) markers from candidate genes, i.e. genes belonging to nutritional pathways and hence most likely to confer variation in nutritional composition. To achieve this goal we are cloning genes in key biochemical pathways, including anthocyanin, carotenoids, sterols, and vitamins B, C, and E, screening DNA sequences for the presence of SNPs in candidate genes, and assaying the nutritional phenotypes of the progeny from a Gwen x Fuerte (G x F) cross. We are also employing biochemical assays to measure the level of nutritional compounds in fruit of the G x F cross (hereafter referred to as phenotypic data). The phenotypic data will be analyzed using the methods of quantitative genetics to determine heritability and ultimately to locate nutritional QTLs (quantitative trait loci). SNPs in candidate genes will be associated with nutritional phenotype for later marker-assisted selection.

Summary of progress

Screening cDNA and genomic libraries for genes involved in nutritional pathways (candidate genes):

SNP development was initiated by a gene discovery phase during which we obtained DNA sequences from the four avocado EST databases (Floral Genome Project, Cornell University; HortResearch (New Zealand), CINVESTAV (Mexico) and our own cDNA library). We performed BLAST-searches against annotated DNA sequences available in the public databases, e.g. TAIR [The Arabidopsis Information Resource], to ascertain whether our sequences matched genes responsible for biochemical phenotypes. Of ca. 65,000 avocado sequence fragments, about a third (22,000) had a homologous sequence in *Arabidopsis* and matched ca. 11,284 distinct *Arabidopsis* genes.

Detecting single nucleotide polymorphisms (SNPs) in candidate genes:

To date, we have sequenced 55 genes at the level of cDNA and 31 at the level of genomic DNA, corresponding to an additional 29 genes (cDNA) and an additional 19 genes (genomic DNA) since April 2009. For each gene we generated sequences of about 500 base pairs in at least 10 different avocado genotypes, evaluated sequence quality using Phred/Phrap software, and examined the sequences for single nucleotide mismatches using the Consed software for SNP detection (Fig. 1). SNPs were confirmed by sequencing both DNA strands. This sequencing effort has yielded 39 SNPs sofar.

Table 1. Progress in the amplification and sequencing of candidate genes. A red “X” denotes a sequence obtained since the mid-year report (April 2009).

Biochemical Pathway or Gene Function Category	Gene Name	Sequenced in cDNA	Sequenced in gDNA
Carotenoids	Beta-carotene hydroxylase (BO2)	X	X
	Lycopene epsilon cyclase (LBC)	X	X

	Phytoene synthase (PSY)	x	x
	carotene beta-ring hydroxylase (LUT5)	x	x
	Beta-carotene hydroxylase 2 (B2)	x	
	Beta-carotene hydroxylase 1 (B1)	X	X
	Zeta-carotene desaturase (ZDS)	x	x
Vitamin B complex			
Vitamin B1 (thiamine)	Cloroplastos alterados 1 (CLA1)	x	x
Vitamin B1 (thiamine)	1-deoxy-D-xylulose-5-phosphate synthase (DXPS1)	x	x
Vitamin B2 (riboflavin)	GTP-Cyclohydolase II; 3,4-dihydroxy-2-butanone-4-phosphate synthase (GCH)	x	
Vitamin B2 (riboflavin)	COI1 suppressor 1 (COS1)	x	
Vitamin B2 (riboflavin)	Riboflavin biosynthesis protein (rib)		
Vitamin B5 (pantothenic acid)	Branched-chain aminotransferase 3 (BCAT3)	x	
Vitamin B5 (pantothenic acid)	Branched-chain aminotransferase 5 (BCAT5)		
Vitamin B6 (pyridoxin)	Pyridoxin biosynthesis 1 (PDX1)	X	X
Vitamin B6 (pyridoxin)	Pyridoxin biosynthesis 2 (PDX2)	X	X
Vitamin B9 (folic acid)	Thymidylate synthase 1 (THY-1)	x	
Vitamin B9 (folic acid)	Thymidylate synthase 2 (THY-2)	x	
Vitamin B9 (folic acid)	aminotransferase class IV family (atrans)	x	x
Vitamin B9 (folic acid)	10-formyltetrahydrofolate synthetase	x	
Vitamin C			
GDP-D-mannose biosynthesis	phosphoglucose isomerase	x	x
GDP-D-mannose biosynthesis	Phosphomannomutase	x	
GDP-D-mannose biosynthesis	GDP-mannose pyrophosphorylase (*VITAMIN C DEFECTIVE 1*)	x	x
GDP-D-mannose biosynthesis	mannose-6-phosphate isomerase	X	
GDP-L-galactose biosynthesis	GDP-mannose-3',5'-epimerase	x	X
Ascorbate biosynthesis	L-galactose-1-phosphate phosphatase	X	
Ascorbate biosynthesis	GDP-L galactose phosphorylase (*VITAMIN C DEFECTIVE 2*)	x	x
Ascorbate biosynthesis	L-galactose dehydrogenase	x	x
Isoprenoid & sitosterol			
Farnesyl diphosphate biosynthesis	farnesyl diphosphate synthase	x	X
Sterol biosynthesis	cycloeucalenol cycloisomerase	x	X
Squalene biosynthesis	squalene synthase (*SQS1*)	x	X
Campestrol and sitosterol biosynthesis	24-dehydrocholesterol reductase	x	x
Vitamin E			
alpha and gamma tocopherol biosynthesis	Tocopherol cyclase (*VITAMIN E DEFECTIVE 1 = VTE1*)	X	
	Homogentisate phytyltransferase (*VTE2*)	X	
	MPBQ/MSBQ methyltransferase (*VTE3*)	x	
	Gamma-tocopherol methyltransferase (*VTE4*)	x	X
	4-hydroxyphenylpyruvate dioxygenase (*PHYTOENE DESATURASE 1*)	x	X
Cell wall hydrolyzing enzymes			
	Polygalacturonase	X	
	Endochitinase	X	
	Cellulase	X	X
Ripening-related genes	Ethylene response sensor (ERS)	X	
	Ripening-related protein gene pAVOe3 ACC oxidase ethylene forming enzyme	X	
Miscellaneous genes	Mitogen activated protein kinase (MAP kinase)	X	
ABA Biosynthesis Pathways	9-cis-epoxycarotenoid dioxygenase NCED1	X	
	9-cis-epoxycarotenoid dioxygenase NCED3	X	X
	Carotenoid cleavage bioxygenase (CCD1)	X	
Amino acid Biosynthesis Pathways	Arginine decarboxylase (ADC)	X	
	Serine/threonine kinase		
Fatty acid pathway	LACS9, long chain acyl-CoA synthetase 9	X	
	LACS9, long chain acyl-CoA synthetase 2	X	
	LACS7, long chain acyl-CoA synthetase 7		
	Acyltransferase, Cuticular 1 (CUT 1)	X	X
	Fatty acid elongase (FAE)		
	ECR, enoyl-CoA reductase	X	
	Lipoxygenase (LOX)	X	
Flavonoid, Anthocyanin & Phenylpropanoid Pathways	Anthocyanidin synthase (ANS)	X	X
	phenylalanine ammonia-lyase 2 (PAL2)	X	X
	flavonol 3'-O-methyltransferase 1 (OTM1)	X	X
	caffeoyl-CoA-O-methyltransferase (caff3)	X	X
	Chalcone synthase (CHS)	X	X

	Flavonol 3-hydroxylase (F3H)	X	X

NB: Enzymes without any entries have been shortlisted for sequencing and SNP development.

Gene expression:

A second phase of the project is to identify candidate transcription factors that are correlated with the expression of genes in nutritional pathways. This phase has just been initiated and the first step is to make RNA from a subset of ten G x F progeny. Fruit are now being collected and dry weight determined; next, fruit samples are vacuum-dried for use in the synthesis of cDNA that will be spotted onto a Nexterion microarray for gene expression studies. A microarray consists of an arrayed series of thousands of microscopic spots of DNA oligonucleotides, each containing tiny amounts of a DNA sequence that uniquely characterizes each of our genes of interest. These "oligos" act as probes that hybridize to a cDNA sample under high-stringency conditions. Hybridization is usually detected and quantified by detection of fluorophore. Microarrays are a means of visualizing and confirming activity of specific genes that are expressed in tissues at the time of sample collection, i.e., ripe fruit versus unripe fruit. We are interested in the candidate genes listed in Table 1 that are active in the biochemical pathways leading to nutritional phenotypes. Transcription factors and other genes with regulatory function are responsible for controlling the expression of our candidate genes. The microarray needs to include both structural and regulatory genes and contrasting tissue stages or types.

Choice of structural genes

To facilitate gene selection for spotting onto the microarray, we annotated the avocado database from CINVESTAV using the Aracyc database that contains the described metabolic pathways and their corresponding genes/proteins in *Arabidopsis*. In total, 2685 avocado sequences match 1859 *Arabidopsis* peptides. There is some redundancy because one avocado sequence may match more than one *Arabidopsis* peptide and vice versa. This annotation is based on a BlastX search against the TAIR9 database using an E-value of 10^{-6} as a cut off in a minimum of 60 nucleotides (20 amino acids).

Choice of regulatory genes (transcription factors)

To target transcription factors (TF) for inclusion on the microarray, we searched the TAIR9 public database using BlastX (stringency E-value of 10^{-6}, cut-off 60 nucleotides). This yielded 536 avocado unigenes matching 240 unique *Arabidopsis* TF loci. The moderate stringency permits discovery of sequences with similarity – not identity - to *Arabidopsis* TF sequences which may be a means of detecting avocado-specific Tfs having functions distinct from those of *Arabidopsis*. We are also investigating the alternative of identifying Tfs via characteristic TF domains plus sequence homology, an approach facilitated by a computer script able to search the Pfam databases (http://pfam.sanger.ac.uk).

Measuring biochemical determinants of nutrition:

All biochemical measurements and molecular studies are being conducted using the avocado trees in our experimental population. The population consists of approximately 200 genotypes that all share the cultivar GWEN (G) as a maternal parent. The paternal parents are ca. 63 Gwen x Fuerte (G x F), 58 Gwen x Zutano (G x Z), 44 Gwen x Bacon (G x B) and the remainder are G progeny having a wide assortment of male parents. Each avocado cultivar is highly heterozygous, so there is tremendous diversity among the progeny of these crosses. Each progeny genotype was asexually propagated four times (via bud grating) with two replicates grown at the South Coast Research and Extension Center in Irvine, CA and two replicates grown at Agricultural Operations in Riverside, CA. Thus each genotype is replicated twice in two environments. The experimental trees are now approximately eight years old and most trees are producing ample fruit.

Sampling:
This fall we harvested immature fruits for the gene expression experiment. We targeted a dry weight (DW) of about 15%. We have collected 7 fruits per tree in 48 G x F progeny trees: first, we took two plugs (cores) from the fruit mesocarp and obtained their DW values. We were able to find 41 G x F genotypes with DW values in the range of 13-17% for at least 4 of the 7 fruits sampled. Since October we have started to check the progression of fruit maturation to sample mature fruits for both the gene expression experiment and to repeat nutrient content evaluation in the second year of the project. Maturation of all G x F progeny is expected to be reached by January or February. For mature fruits we intend to target minimum commercial maturity standards, corresponding to a minimum DW of 24.4 and 19% for Gwen and Fuerte, respectively; for the G x F progeny, the minimum DW is taken to be intermediate between these values.

Biochemical Assays:
The methods for the identification and quantification of anthocyanin, carotenoids, sterols, Vitamin C, and Vitamin E have been elaborated and complete data sets have been gathered for anthocyanins, carotenoids and sterols. The extractions for Vitamin E are complete and samples will be run on TLC plates shortly to carry out quantification. The extraction and quantitation of Vitamin C is 50% complete. We are also progressing well on the methods for the Vitamin B complex.
The carotenoid dataset contains 169 distinct absorbance measurements from 65 G x F genotypes. Absorbance values range from 0.018 to 0.509 [corresponding to a 28-fold difference between the genotypes with the lowest and highest values]. The absorbance values have now been converted to g/g fresh weight using a standard curve calibrated in relation to a commercial standards.

Determining the heritability of nutritional phenotypes following the statistical methods of quantitative genetics:
Statistical calculations using SAS software show that the heritability of total carotenoid content is high (0.763 based on log-transformed absolute values in ng). Both genotype and environment have a highly significant effect on carotenoid content ($p < 0.0001$), while the genotype x environment interaction is not significant ($p = 0.53$). Heritability of proanthocyanidin content is moderately high (0.307 based on log-transformed relative absorbance values). Both genotype and environment have a significant effect on proanthocyanidin content ($p < 0.01$). Genotype effect ($p < 0.0001$) is higher than environmental effect ($p= 0.0087$). Genotype x environment interaction is highly significant as well ($p < 0.001$).
Heritability of fruit maturation rate (based on dry weight measures) is relatively low (0.168 based ranked dry weight values). Maturity is affected both by genotype ($p<0.0001$) and environment ($p=0.0013$) and genotype effect is higher than environmental effect. Genotype*environment interaction is significant as well ($p<0.0001$).
Heritability of fruit weight is moderate (0.315 based on log transformed g values). Both genotype and environment have a significant effect on fruit weight ($p<0.0001$). Genotype*environment interaction is highly significant ($p<0.0001$).
The level of heritability of pulp weight is similar to that of fruit weight (0.359 based on log transformed g values). The effect of genotype, environment and genotype*environment interaction are all highly significant ($p<0.0001$).
Heritability of stone weight is fairly low (0.220 based on log transformed g values). As for fruit and pulp weight, the effect of genotype, environment and genotype*environment interaction on stone weight are all significant ($p<0.0001$).

Searching for statistical associations between candidate gene SNPs and nutritional phenotype in a genetically defined population:
This aspect of the work will be initiated when we have the SNP and phenotypic data in the final year of the project. In an exploratory analysis we searched for correlation between preliminary proanthocyanidin data and the PAL2 gene SNP pattern using the general linear model implemented in TASSEL. We were encouraged to find one SNP to be associated with proanthocyanidin content.

Implementing a program of marker assisted selection based on any SNP-phenotype associations detected:
This is the ultimate objective of our work, but it was not listed as an objective in the proposal because it is planned for implementation after all the data is in hand at the end of the current project.

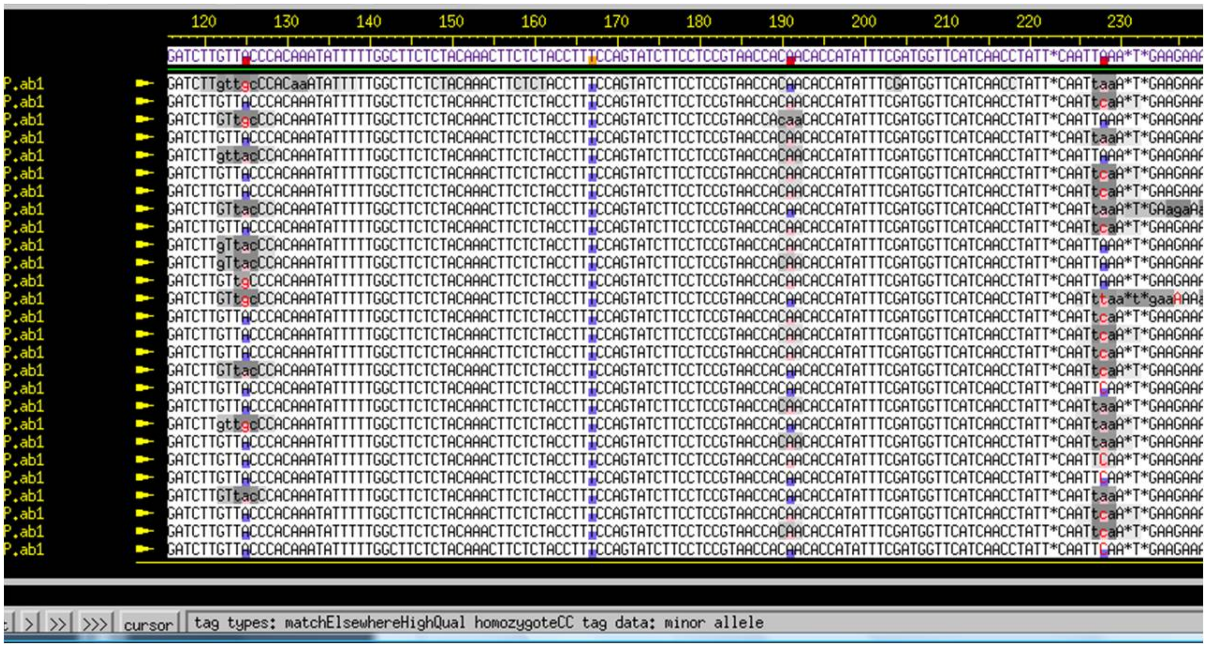

Fig. 1. Sequence alignment for the GDP-D-mannose pyrophosphorylase gene.