

Discovery, mapping, and application of single nucleotide polymorphisms in *Lycopersicon esculentum*

Wencai Yang¹, Xiaodong Bai², Christina Eaton¹, Sophien Kamoun³, Esther van der Knaap¹, and David Francis¹

¹Department of Horticulture and Crop Science, ²Department of Entomology, ³Department of Plant Pathology, The Ohio State University, Ohio Agricultural Research and Development Center, 1680 Madison Ave., Wooster, OH 44691, USA

Summary

Single nucleotide polymorphism (SNP) discovery through *de novo* sequencing is inefficient within cultivated tomato (*Lycopersicon esculentum* Mill.) because the polymorphism rate is lower than the sequencing error rate. The availability of expressed sequence tag (EST) data has made it feasible to discover putative SNPs "in silico" prior to experimental verification. By dividing the tomato ESTs by variety of origin, selecting contigs with a minimum of three sequences to correct for sequence error, and comparing sequences from different varieties, we have identified candidate SNPs for use within cultivated germplasm pools. 1245 contigs having three EST sequences of Rio Grande and three EST sequences of TA496 were used for SNP discovery. We detected 1 SNP for every 8,500 bases analyzed, with 101 candidate SNPs in 44 genes identified. Experimental verification using restriction digestion or Ccl 1 digestion confirmed 83% of the putative polymorphisms tested. SNPs between Rio Grande and TA496 have a high probability (53%) of detecting SNPs between other *L. esculentum* varieties. Twenty-six SNPs in 18 unigenes were mapped to specific chromosomes. SNPs LEOH23 and LEOH37 were linked to quantitative trait loci contributing to fruit color in crosses between elite varieties.

Introduction

A limitation to applying marker-assisted selection to the practice of breeding tomato varieties is that the level of polymorphism between elite varieties is very low. The objective of this research was to assess the potential of using existing public databases for the *in silico* discovery of polymorphisms. The tomato genome project has made available at least 138,100 expressed sequence tags (ESTs). Of these, approximately 15% were derived from the variety Rio Grande or from Rio Grande X Moneymaker crosses. The majority of remaining sequences were derived from TA496, which has a processing tomato pedigree tracing to E6203. By comparing sequence data from Rio Grande and TA496 we identified genetic differences between varieties. Polymorphisms discovered from this data mining were then applied to genetic studies within breeding populations, and markers linked to quantitative trait loci (QTL) contributing to fruit color were identified.

Mapping of SNPs

Two populations were used to map the SNPs. The first was a set of *L. pennellii* LA716 introgression lines (ILs) in the background of *L. esculentum* cultivar M82 (Eshed and Zamir, 1995 Genetics 141:1147-1162). The second population was an F₂ derived from a cross of LA1589 (*L. pimpinellifolium*) and Sun 1642.

Color measurement

Objective descriptions of the red, green, yellow and blue components of tomato color were obtained using the "L*a*b*" CIELAB color space (Commission Internationale de l'Eclairage, 1978) and a Minolta CR-300 colorimeter. The L* coordinate indicates darkness or lightness of color. Chroma (saturation or vividness of color) is calculated from a* and b* as $(a^2 + b^2)^{1/2}$. As chromaticity increases, a color becomes more intense; as it decreases a color becomes more dull. Data were collected for 24 fruit from individual F₂ progeny in two breeding populations, Ohio 8245 x Ohio 2349 consisting of 160 individuals and Ohio 1023 x Ohio 7814 consisting of 80 individuals.

Statistical analysis

Linkage relationships between genotypic classes of SNPs with L and Chroma were determined by ANOVA using single marker-trait analysis. Markers were considered as fixed effects, and replicate measurements and genotype were considered as random effects, and the F-test was performed using the genotype within marker variation as the error term. Total phenotypic variation explained by markers identified for L and Chroma was calculated from variance components estimated using restricted maximum likelihood (REML).

Results

SNPs between TA496 and Rio Grande

1245 contigs with at least three ESTs for each variety were available for identifying potential SNPs. Forty-four unigenes showing 101 potential polymorphisms were identified. Only two polymorphisms were insertion/deletions (indels). Sixty-six candidate SNPs could be recognized by commercially available restriction enzymes. Four varieties: TA496, Rio Grande, E6203, and Money maker, were used for initial confirmation. Forty-three out of 55 SNPs tested (82.7%) were confirmed. The frequency of polymorphisms between TA496 and Rio Grande is 1 SNP in approximately 8,500 bases analyzed. In terms of genes analyzed, one SNP was detected per 15 unique ESTs. The average number of SNPs per EST was 1.79 with most contigs containing only one or two and four contigs containing 5 or more SNPs.

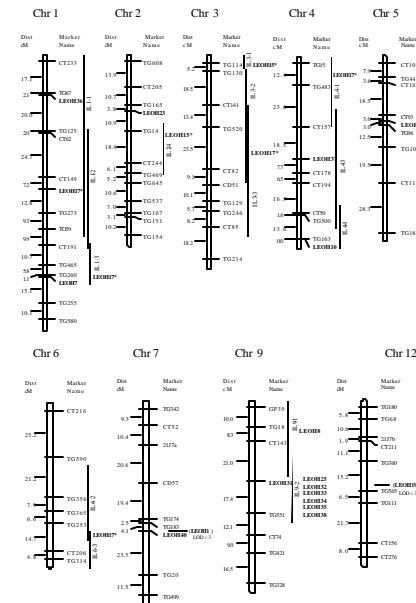


Figure 2. Map position of SNPs. SNPs mapped in the F₂ population derived from a cross of LA1589 (*L. pimpinellifolium*) and Sun1642 (*L. esculentum*) are indicated relative to frame-work markers on chromosomes. Bold lines to the right of each chromosome indicate the positions of *L. pennellii* LA716 introgression lines (ILs). SNP markers mapped to ILs are indicated to the right. SNP markers with an asterisk (*) indicate multiple map positions. Markers mapped with LOD<3.0 are indicated.

Identifying SNPs associated with fruit color

SNP Markers that were polymorphic between the parents of two F₂ breeding populations involving elite x elite crosses detected QTL contributing to fruit color (Table 2). These QTL are located on chromosome 2 and chromosome 4, explain 14% to 22% of the observed variation for color, and do not correspond to previously described genes (for example *og²* on chromosome 6) known to affect color in tomato.



Figure 3. Color differences within elite *L. esculentum* germplasm.

Table 2. SNPs associated with lightness-darkness of color (L) and intensity of color (Chroma) in two elite breeding populations.

Marker	Population	p-value & variance explained Vp			
		L		Chroma	
		p	Vp	p	Vp
LeOH23	OH 1023 x OH 7814	0.022	0.146	ns	ns
LeOH37	OH 8245 x OH 2349	ns	ns	<0.0001	0.216

Conclusions

The frequency of single nucleotide polymorphisms in *L. esculentum* ESTs is lower than in other plant species. The polymorphism rate is approximately ten fold lower than the sequencing error rate.

Public access to EST sequence trace files or Phred quality value data would allow for more efficient SNP discovery by permitting the use of quality information as a substitute for sequence redundancy.

Based on the estimated number of genes in tomato (35,000; Van der Hoeven, et al. 2002 The Plant Cell 14:1414-1456) we estimate that there are 1000 SNPs between TA496 and Rio Grande.

SNPs between Rio Grande and TA496 have a high probability (53%) of detecting SNPs between other *L. esculentum* varieties.

LEOH23 and LEOH37 are linked to quantitative trait loci contributing to fruit color in crosses between elite varieties. These SNPs will be useful in marker-assisted selection.

Table 1. Summary of EST derived SNPs confirmed in *Lycopersicon esculentum*. Forty-three out of 55 SNPs tested (82.7%) were confirmed.

SNP	Chromosome	codon substitution	Forward primer (5'-3')	Reverse primer (3'-5')
LEOH1	7	non-synonymous	TCC ACA TAA AAT AAT GGA CAG AC	TTC TTC GTC AGC ATC GGG TA
LEOH2	9	synonymous	CCA CTG ATC AAT GTG GTG GA	CAA CCA CAA ATG GCT CCT AAA
LEOH3	unknown	non-synonymous	GGC AAT GGC ACT GAC TTA CA	CTC TGT GCT GCT GGT GCT AC
LEOH4	4	synonymous	TGC CAG ATT GAC TGT GAA GG	GSA ACC GTG CATT TGT TGT TGT
LEOH13	unknown	synonymous	TGG CTG GTG ACA TTA TTG GA	GGG GAT CTT GCC ATA TAA TA
LEOH14	unknown	non-synonymous	TGC CAG AGG ACA CCG AAT AA	GTA AGG ACC TTG TCC GAT GCC
LEOH15	2 & 3	synonymous	GGG GTT AAT TGT TGG TCG TC	GTG TGC CAT GGG TAA TCA CCG
LEOH16.1	5	non-synonymous	TGC AGC GTG CAG AAC AAT AC	TTC CTC CTT CTT ACT TCC TTCA
LEOH16.2	5	non-synonymous		
LEOH16.3	5	non-synonymous		
LEOH17.1	multiple	non-synonymous	CAG AGG AGA AGG AAG TTG AGG	CTA CCA CTG GGT GCT TTG AC
LEOH17.2	multiple	non-synonymous		
LEOH17.3	multiple	non-synonymous		
LEOH17.4	multiple	non-synonymous		
LEOH19	12	left indel	AGG GCT CAG AAA GGG TCG AT	TGA GTT CAT CAA CAC ATC ACA CA
LEOH20	unknown	non-synonymous	CAG ACC TAA CAA GAG CAG CA A	ATC AGG CAT GAC CAT GGA AG
LEOH23.1	2	left indel	GAG AGA AAA AGG GCA CAA G	ACC GAC AAA CCG ATA GAT CA
LEOH23.2	2	synonymous	GTA TGC GAT TGT GGG TGG T	CAA GGT AGT TGA AGT TAT GAC CA
LEOH23.3	9	synonymous	GGA GGA AAT AGG GTT TCG AG G	AAT GGC CTG CCT AAT CTG TGT
LEOH26	9	non-synonymous	GAA GAT TGG GAG GTC AAG G	MAC TCC TCA ACT GGC TCA GC
LEOH28	unknown	non-synonymous	CCT CCA TGA CCG ATG TCA CT	TAG TGA TCT CTC GCT GGA CA
LEOH29	9	non-synonymous	TTG CAA TGG CTT CTT CTC TC	ACT TGT CCG TTT CTT GCT GT
LEOH31.2	9	non-synonymous		
LEOH31.3	9	synonymous	TGT TGA TGT GTG CTT GAT T	CCC TGC CCA AAC ATA TAA A
LEOH31.4	9	synonymous		
LEOH31.5	9	non-synonymous		
LEOH31.6	9	non-synonymous		
LEOH32.1	9	synonymous	TGG TGT GGA TCC TGT TGT TA	TGG AAA TCA CAC CAA AAG GA
LEOH32.2	9	synonymous		
LEOH33.1	9	non-synonymous	TGA GGA AGC TTG CTG ACA AA	GCC TTT ATC TTT TAA GGC TGC AAT
LEOH33.2	9	non-synonymous		
LEOH33.3	9	non-synonymous		
LEOH34.1	9	synonymous	CGT TCT ACC AAT TGC ACT CA	CTT ATG TAT CCG GGG CTT CT
LEOH42	9	synonymous		
LEOH43	8	left indel	CAT CAG GCT CCG TCT CTT CT	CAA ACT GCA AGC CAT TTT AA
LEOH5.1	9	non-synonymous		
LEOH5.2	9	non-synonymous		
LEOH5.3	9	non-synonymous		
LEOH5.4	9	synonymous		
LEOH6	1	non-synonymous	TCA CAA AAA TGG CGA TTA GA	CGA CDT GTG GAT CDT TGA CT
LEOH7	4	left indel	TTG ATA TAT TCC ATG TGT GTC TC	ACC TAC AAA TTA ACA AAC TTA AAT GG
LEOH8	9	non-synonymous	CAA GGT TGT GGC TAT GCT CA	ACC TCA GCA GTA TTT ACC AG
LEOH9	7	non-synonymous	TGA GGT GGT GAA CCA TGG AA	CCA ANG TTT GGA CDT TTT GA

SNPs polymorphic in other *L. esculentum* and wild species germplasm

To test if the SNPs identified between TA496 and Rio Grande are also polymorphic among other *L. esculentum* germplasm, 19 varieties were used for further verification. Of the 43 confirmed SNPs, 23 also showed polymorphisms among the sampled germplasm. Thus SNPs discovered in the EST database had a high probability (53.5%) of detecting SNPs between other varieties.

SNPs were also tested for polymorphism in three wild species: LA716 (*L. pennellii*), LA407 (*L. hirsutum*), and LA1589 (*L. pimpinellifolium*). SNPs present between TA496 and Rio Grande had 82.5% polymorphism with LA716, 80% polymorphism with LA407, and 67.5% polymorphism with LA1589.

Map position of SNPs

26 SNPs in 18 unigenes showed polymorphisms in either mapping population 1, population 2 or both. Sixteen unigenes were mapped to a specific chromosome (Figure 2). Map positions of the 10 SNPs in common to both populations were consistent. Two unigenes, LEOH15 and LEOH17, detected multiple gene families and could not be mapped to a specific chromosome. LEOH15 amplifies a *CAB* gene with family members that were mapped to chromosomes 2 and 3, consistent with the location of *CAB1* and *CAB3* respectively. LEOH17 amplifies an *Adh* gene that was mapped to 5 chromosomes in the segmental substitution population and only chromosome 1 using the second population. Although the mapped genes cover 9 of the 12 tomato chromosomes, 8 of 16 were placed on chromosome 9.

Figure 1. Data flow for identification of single nucleotide polymorphisms. Steps to select contigs and combine data were facilitated by routines written in perl. SNPs were verified by either Ccl 1 digestion (A; where H is heteroduplex and C is control) or restriction enzyme digestion (B).

Materials and methods

Identifying single nucleotide polymorphisms (SNPs)

The approach to discovering SNPs in ESTs from *Lycopersicon esculentum* is outlined in Figure 1. Briefly, the NCBI EST data base was parsed into databases consisting of only TA496 and Rio Grande, R11-12 and R11-13. Sequence comparisons between these data sets were used to identify potential SNPs.

Confirmation of Candidate SNPs

Primers were designed flanking the putative polymorphism with the optimal PCR product length set between 150 and 600 bp. Four varieties TA496, E6203, Rio Grande, and Moneymaker were used for verification of SNPs. Another 19 varieties of *L. esculentum* and 3 wild species were used to determine the utility of these SNPs as markers for other crosses. Putative SNPs were verified as cut amplified polymorphisms using PCR followed by restriction enzyme digestion or by Ccl 1 digestion of heteroduplexes formed by mixing amplified products, heating to denature, and re-annealing. Ccl 1 was purified according to published methods and DNA digestion was performed at 45 °C in 20 mM Tris-HCl pH 7.4, 25 mM KCl and 10 mM MgCl₂ for 30 min. Single stranded Ccl 1 digestion products were separated using 10% TBE-urea polyacrylamide gels and visualized by staining with SYBR Gold.